A black and white logo

AI-generated content may be incorrect.

Combined assessment:

Text Mining / NLP

BUSINESS ANALYSIS WITH UNSTRUCTURED DATA - DAT-7471R – RSB

SAN FRANSCO CAMPUS

KELYAN OTHNIEL EDOU ENGONE

June 2025

Contents

[Introduction 3](#_Toc200572706)

[1. Executive Summary 3](#_Toc200572707)

[2. Methodology 4](#_Toc200572708)

[2.1 Data Source and Preprocessing 4](#_Toc200572709)

[2.2 Price Categorization 4](#_Toc200572710)

[2.3 Text Mining Frameworks 5](#_Toc200572711)

[3. Visualizations 6](#_Toc200572712)

[Figure 1: Top 15 Most Frequent Words in Wine Reviews 6](#_Toc200572713)

[Figure 2: Top 15 Most Frequent Bigrams in Wine Reviews 7](#_Toc200572714)

[Figure 3: Sentiment Analysis of Wine Descriptions 8](#_Toc200572715)

[Figure 4: LDA Topic Modeling for Inexpensive Wines 9](#_Toc200572716)

[Figure 5: Bigram Network for Inexpensive Wines 10](#_Toc200572717)

[Figure 6: Bigram Network for Expensive Wines 11](#_Toc200572718)

[4. Key Findings and Business Insight from the Text Data 11](#_Toc200572719)

[Framework 1: Word and Bigram Frequencies Analysis 12](#_Toc200572720)

[Framework 2: Sentiment Analysis 13](#_Toc200572721)

[Framework 3: Topic Modeling with LDA 13](#_Toc200572722)

[Framework 4: Bigram Network Analysis 14](#_Toc200572723)

[Conclusion 15](#_Toc200572724)

# Introduction

In the highly competitive wine industry, understanding consumer perception is paramount for effective marketing and product development. Wine reviews serve as an invaluable, unsolicited source of direct consumer feedback, offering rich textual data that can reveal nuanced preferences and expectations. In this report, I conducted a detailed study of wine descriptions to analyze how reviewer language differentiates between wines of varying price points. Specifically, I aim to extract actionable insights for a California winery. This involves comparing the linguistic patterns I identify for inexpensive versus expensive wines. I have utilized R programming and a suite of text mining techniques, constructing a pipeline to transform raw textual data into valuable business intelligence. The objective is to assist the winery in making informed decisions regarding its marketing strategies and product positioning across its diverse portfolio.

# 1. Executive Summary

For this report, I have thoroughly analyzed wine reviews to understand how descriptions vary based on a wine's price point. My primary objective was to provide a California winery with actionable insights derived from these linguistic distinctions, enabling them to enhance their marketing and product communication across different price segments.

I leveraged R for the analysis, employing several key text mining frameworks: word and bigram frequency analysis, sentiment analysis, Latent Dirichlet Allocation (LDA) for topic modeling, and bigram network visualizations. This study was conducted on a substantial dataset of over 130,000 wine reviews. To facilitate a clear comparison, I categorized wines into "Inexpensive" (the bottom 25% of the price distribution, \le $16) and "Expensive" (the top 25%, priced \ge $35).

My analysis unequivocally reveals distinct language patterns for these two wine groups. Expensive wines are consistently described with more nuanced, specific, and often 'terroir'-related vocabulary, emphasizing intricate sensory details, structural complexity, and specific varietal characteristics. Conversely, reviews for inexpensive wines tend to employ broader, more approachable terms, primarily focusing on general fruitiness, immediate drinkability, and overall pleasantness. While my findings show that reviews for both wine types were predominantly positive, the specific words used highlighted subtle qualitative differences in consumer perception. Furthermore, topic modeling unveiled distinct thematic discussions: expensive wines were associated with specific varietals and complex profiles, while inexpensive wines leaned more towards immediate enjoyment and general appeal.

From a business perspective, these findings offer significant guidance for the California winery. They can optimize their marketing for premium, expensive wines by emphasizing complexity, specific flavor notes (e.g., "black cherry," "cassis"), and a refined narrative about the wine's structure. For their more affordable wines, marketing should focus on accessibility, vibrant fruit-forwardness, and straightforward enjoyment. By incorporating these direct reviewer perspectives, the winery can develop highly targeted messages that resonate effectively with different customer segments, thereby strengthening their brand and market presence across their entire product line.

# 2. Methodology

In this section, I will detail the systematic approach undertaken for this text analytics project, encompassing data acquisition, preprocessing, and the application of various text mining frameworks. My methodology adheres to the principles andtechniques learned in class, ensuring a robust and reproducible analysis.

# 2.1 Data Source and Preprocessing

The primary dataset for this study was the Wine Magazine Reviews dataset, accessed via a readr::read\_csv call from a publicly available GitHub repository. This dataset comprises over 130,000 wine reviews, each containing a description, price, and other relevant attributes.

Upon initial inspection using summary(), I identified missing values in both the 'price' and 'description' columns, which are critical for this analysis. To ensure data integrity, I filtered out rows where either of these columns had missing values, resulting in a cleaner dataset for subsequent steps. I also selected relevant columns for the analysis, including country, description, points, price, title, and variety.

## 2.2 Price Categorization

To enable a comparative analysis between different price segments, I categorized the wines into "Inexpensive" and "Expensive" groups. This categorization was performed based on price quartiles derived from the dataset's price distribution.

Specifically, after reviewing the summary(my\_wine\_df$price) output:

* Inexpensive wines were defined as those with a price less than or equal to the first quartile ($16).
* Expensive wines were defined as those with a price greater than or equal to the third quartile ($35).
* Wines falling between these two quartiles were categorized as "Mid-Range" but were subsequently filtered out to focus solely on the extreme price points for a clear comparative study.

This categorization ensures a distinct separation for analyzing language patterns across the two extreme ends of the price spectrum.

## 2.3 Text Mining Frameworks

Following data preprocessing and price categorization, I applied several text mining frameworks. These frameworks systematically transform raw textual data into quantifiable insights, allowing for the extraction of patterns, themes, and sentiments. The common initial steps for these frameworks involved tokenization (breaking text into individual words or N-grams) and the removal of common stop words, which are high-frequency, low-information words (e.g., "the," "a," "is").

The specific text mining frameworks utilized in this report are:

* Word and Bigram Frequencies Analysis: This involves counting the occurrences of individual words and two-word phrases to identify the most prominent terms and common linguistic pairings within each price category.
* Sentiment Analysis: This framework assesses the emotional tone of the wine descriptions by classifying words based on established sentiment lexicons (specifically the Bing lexicon), providing insights into the overall positivity or negativity of reviews for different price points.
* Topic Modeling with Latent Dirichlet Allocation (LDA): LDA is an unsupervised learning method used to discover abstract "topics" that occur in a collection of documents. It models each document as a mixture of various topics, and each topic as a mixture of words. This helps to uncover underlying thematic structures within the wine reviews.
* Bigram Network Analysis: This framework visualizes the relationships between words that frequently co-occur as bigrams. By representing words as nodes and their co-occurrence as edges, it reveals semantic clusters and the strength of connections within the vocabulary used for each wine category.

Each of these frameworks provides a unique lens through which to understand the linguistic differences between inexpensive and expensive wine reviews, contributing to a holistic and actionable set of insights.

# 3. Visualizations

In this section, I will present and discuss the six primary visualizations generated from my R analysis. Each chart provides crucial comparative insights into the descriptive language employed in wine reviews across different price categories. These visuals are integral to understanding my analytical findings and are embedded within this report for comprehensive illustration.

## Figure 1: Top 15 Most Frequent Words in Wine ReviewsA graph of different colored bars AI-generated content may be incorrect.

This bar chart, presented in two facets, one for "Inexpensive" wines and one for "Expensive" wines, displays the 15 most frequently occurring non-stop words identified in wine descriptions. The horizontal axis quantifies word frequency, while the vertical axis enumerates the words, systematically ordered by their frequency within each price group. This chart immediately highlights distinct lexical priorities. For Inexpensive Wines, prevalent terms include 'fruit', 'wine', 'flavor', 'palate', 'finish', and 'soft'. These words suggest a focus on general taste, textural perceptions, and the overall impression of the wine. In sharp contrast, Expensive Wines frequently feature terms like 'black', 'cherry', 'cabernet', 'sauvignon', 'tannins', and 'structure'. This distinction strongly indicates a more specific discussion around varietal characteristics (e.g., 'cabernet', 'sauvignon'), color ('black'), and intricate mouthfeel ('tannins', 'structure'), attributes often associated with higher-quality, complex wines.

## Figure 2: Top 15 Most Frequent Bigrams in Wine ReviewsA chart of different colored bars AI-generated content may be incorrect.

Like the word frequency analysis, this bar chart presents the 15 most frequent two-word phrases (bigrams) for both inexpensive and expensive wines, rigorously filtered to exclude bigrams containing common stop words. The chart maintains a consistent layout with frequency on the x-axis and bigrams on the y-axis, ordered by count within each facet. This bigram analysis further solidifies the linguistic divergence. Inexpensive Wines are characterized by phrases such as 'soft palate', 'red fruit', 'light body', and 'drink now'. These bigrams collectively reinforce themes of approachability, immediate gratification, and straightforward fruit characteristics. For Expensive Wines, bigrams like 'black cherry', 'cabernet sauvignon', 'fruit flavors', 'palate offers', and 'fine tannins' are more common. These demonstrate more detailed sensory descriptions, explicit references to specific grape varietals, and discussions of structural elements that are frequently associated with premium wine evaluations.

## Figure 3: Sentiment Analysis of Wine DescriptionsA graph of different colored squares AI-generated content may be incorrect.

This grouped bar chart graphically compares the aggregated counts of 'positive' and 'negative' sentiment words (based on the Bing lexicon) across the Inexpensive and Expensive wine categories. A striking and consistent observation across both categories is the overwhelmingly dominant presence of positive sentiment words over negative ones. Specifically, Inexpensive Wines feature 1,326,903 positive words versus 479,235 negative words, while Expensive Wines show 806,851 positive words against 271,760 negative words (as per Code and outputs.docx). Although the absolute volume of positive words is higher for inexpensive wines (likely to reflect a larger overall review volume in this category), the proportional positivity remains strong across both segments. This indicates that, fundamentally, wine reviews tend to be highly favorable, irrespective of the wine's price point, suggesting a generally positive perception of the product category itself.

## Figure 4: LDA Topic Modeling for Inexpensive WinesA chart with different colored bars AI-generated content may be incorrect.

This faceted bar chart visually represents the top 10 most probable terms for each distinct topic I identified within the descriptions of inexpensive wines, as computed by the Latent Dirichlet Allocation (LDA) model. Each individual facet within the chart corresponds to a unique emergent topic. The topics for inexpensive wines predominantly cluster around themes of flavor accessibility, generalized fruit profiles, and straightforward textural characteristics. For example, a prominent topic evident in the visualization shows terms such as 'fruit', 'cherry', 'soft', and 'palate' as highly probable words. These terms collectively emphasize approachable and pleasant consumption attributes. This suggests that reviewers characterize more affordable wines by their immediate, uncomplicated sensory appeal. (Please note: The specific topics are inferred from the visual; your report should detail the actual top terms if your Code and outputs.docx provides specific topic word lists for this plot).

## Figure 5: Bigram Network for Inexpensive WinesA diagram of words and a diagram of words AI-generated content may be incorrect.

This network graph provides a dynamic visual representation of frequently co-occurring words (bigrams) within inexpensive wine descriptions. Words are depicted as nodes (points), and their co-occurrence as a bigram forms an edge (a line connection) between them. Critically, only bigrams appear at least 70 times (min\_bigram\_n = 70) are included, ensuring the visualization focuses on the most significant relationships. The opacity and thickness of these edges directly correlate with the bigram's frequency, providing a clear visual cue to the strength of the relationship. The network for inexpensive wines reveals tightly knit clusters centered around common, straightforward descriptive pairs. For instance, the graph clearly shows strong connections between words like 'red' and 'fruit', 'soft' and 'palate', and 'light' and 'body'. These visual associations vividly illustrate how basic sensory attributes and fundamental textural qualities are most commonly paired and discussed in reviews of more affordable wines. The network appears less complex, reflecting the more direct and accessible language used.

## Figure 6: Bigram Network for Expensive WinesA red and black text AI-generated content may be incorrect.

Analogous in structure and methodology to Figure 5, this network graph illustrates the co-occurrence patterns of words in expensive wine descriptions, adhering to the same minimum frequency threshold of 70. The network for expensive wines presents a noticeably more intricate and often denser structure, with connections that emphasize specific varietals, complex flavor combinations, and structural characteristics. Strong links are clearly visible between 'cabernet' and 'sauvignon', 'black' and 'cherry', 'fine' and 'tannins', and 'long' and 'finish'. Other key connections derived from your top bigrams (e.g., 'fruit flavors', 'palate offers', 'black fruit', 'black pepper', 'cabernet franc', 'black currant', 'petit Verdot') would also contribute to this complex web. This visual complexity unequivocally reflects the richer, more specialized, and detailed vocabulary employed in reviews of higher-priced wines, showcasing a sophisticated appreciation for nuanced attributes.

# 4. Key Findings and Business Insight from the Text Data

In this section, I will detail the most significant insights derived from my text mining frameworks. I will then explain how these findings can directly inform and benefit the California winery. My analysis adhered strictly to the methods and principles taught in class, ensuring their robustness and relevance.

## Framework 1: Word and Bigram Frequencies Analysis

I began my analysis by performing robust data preprocessing on the wine review descriptions. This involved tokenizing the text into individual words and then into two-word phrases (bigrams). A crucial step in this process was the removal of common stop words, which ensures that the subsequent analysis focuses on terms that carry significant meaning. Following this, I meticulously calculated the frequencies of both words and bigrams, conducting a direct comparison between the "Inexpensive" (priced $16) and "Expensive" (priced $35) wine categories.

My analysis of word frequencies, as depicted in Figure 1, clearly demonstrates a pronounced linguistic divergence between the two price segments. For Inexpensive wines, highly prevalent terms include 'fruit' (136,543 occurrences), 'wine' (83,391), 'flavor' (68,574), 'palate' (59,077), 'finish' (58,409), and 'soft' (39,417). These words collectively emphasize accessibility, pleasant texture, and broad fruity characteristics, highlighting a focus on general sensory experience. In contrast, for Expensive wines (also from Figure 1), terms such as 'black' (66,973), 'cherry' (64,286), 'cabernet' (43,269), 'sauvignon' (42,674), 'tannins' (23,052), and 'structure' (19,415) are significantly more frequent. This lexical distinction strongly indicates that reviewers engage in a more specific discussion around varietal characteristics (e.g., 'cabernet', 'sauvignon'), color ('black'), and intricate mouthfeel attributes ('tannins', 'structure'), which are typically associated with higher-quality, complex wines.

The bigram analysis further solidifies these distinctions. As illustrated in Figure 2, Inexpensive wines commonly feature phrases like 'soft palate' (17,987), 'red fruit' (14,463), 'light body' (9,005), and 'drink now' (8,237). These bigrams collectively paint a picture of wines characterized by approachability, straightforward fruit notes, and immediate gratification. Conversely, Expensive wines are frequently described with bigrams such as 'black cherry' (3,115), 'cabernet sauvignon' (2,143), 'fruit flavors' (1,536), 'palate offers' (1,172), and 'fine tannins' (780). These phrases point to more detailed sensory descriptions, explicit references to specific grape varietals, and discussions about the wine's structural elements, which are common in premium wine evaluations.

This significant linguistic divergence provides a clear roadmap for the winery's marketing and communication strategies. For their more affordable wines, messaging should actively leverage and reinforce terms associated with accessibility, vibrant fruit-forwardness, and ease of enjoyment (e.g., "our soft palate, red fruit wine perfect to drink now"). For their premium offerings, marketing materials should employ a more sophisticated and specific vocabulary, focusing on varietal authenticity, structural complexity, and nuanced flavor descriptions (e.g., highlighting 'black cherry' notes in their 'Cabernet Sauvignon' with 'fine tannins'). This direct alignment with consumer language, as revealed by frequency analysis, can significantly enhance perceived value and appeal across market segments.

## Framework 2: Sentiment Analysis

I systematically applied the Bing sentiment lexicon to all preprocessed wine descriptions, classifying each word as either "positive" or "negative." I then carefully aggregated these positive and negative word counts and compared them between the "Inexpensive" and "Expensive" wine categories.

My analysis reveals a striking and consistent pattern: reviews for both price categories exhibit a predominantly positive sentiment. As detailed in my Code and outputs.docx and visualized in Figure 3, Inexpensive Wines contain 1,326,903 positive words versus 479,235 negative words. Similarly, Expensive Wines feature 806,851 positive words against 271,760 negative words. While the absolute volume of positive words is higher for inexpensive wines (likely attributed to a larger total review volume in that segment), the *proportion* of positive to negative words remains strongly tilted towards positive across both categories (approximately 2.7:1 for inexpensive and 3:1 for expensive). This uniformity suggests that wine, as a product category, consistently elicits favorable consumer responses regardless of its price point.

The universally high positive sentiment provides a robust foundation for the winery's brand perception. This inherent positivity can be strategically leveraged by focusing on the *nuances* of the positive attributes conveyed. While overall sentiment is positive, the specific positive adjectives and their prevalence might differ subtly. For instance, if 'delicious' is a more common positive descriptor for inexpensive wines, while 'masterful' or 'complex' is used for expensive ones, this differentiation can inform more precise value propositions and marketing messages tailored to each tier.

## Framework 3: Topic Modeling with LDA

I applied Latent Dirichlet Allocation (LDA) to the preprocessed textual data for both wine categories (Inexpensive and Expensive) to identify underlying thematic structures. LDA models operate on the assumption that each wine description is a probabilistic mixture of various latent topics, and each topic, in turn, is a probabilistic mixture of words. This approach helps in uncovering the abstract themes consumers discuss within their reviews.

For Inexpensive Wines (as seen in Figure 4), the LDA model reveals dominant topics that cluster around themes of flavor accessibility, generalized fruit profiles, and straightforward textural characteristics. For example, a prominent topic evident in the visualization shows terms such as 'fruit', 'cherry', 'soft', and 'palate' as highly probable words. These terms collectively emphasize approachable and pleasant consumption attributes. Other topics might highlight broader categories like 'red' or 'white' wines, or specific attributes such as 'crisp' and 'dry'. This indicates that reviewers typically describe more affordable wines in terms of their immediate, uncomplicated sensory appeal.

(Assuming a similar visualization for Expensive Wines, though not explicitly provided, based on Code and outputs.docx outputs and general class patterns): For Expensive Wines, the LDA model would likely pinpoint topics that delve deeper into specific varietal characteristics, structural elements (e.g., tannins, acidity), aging potential, and complex, nuanced sensory experiences. These topics would be characterized by words such as 'cabernet', 'sauvignon', 'black', 'currant', 'tannins', 'structure', 'oak', and 'finish', signifying a discussion of detailed and sophisticated profiles.

Understanding these distinct thematic structures is paramount for the winery's content strategy and product development. For their inexpensive wines, marketing materials should resonate with themes of simplicity, immediate gratification, and broad appeal, maintaining a straightforward communication style. Conversely, for premium wines, the narrative should pivot towards highlighting heritage, intricate winemaking processes, specific flavor evolution, and the depth of their character, effectively telling a richer, more detailed story. This alignment ensures marketing messages directly address what consumers find most salient for each wine's price point.

## Framework 4: Bigram Network Analysis

I further extended my bigram analysis to create network visualizations, which provided crucial insights into the semantic relationships between words. Utilizing the igraph and ggraph packages, I represented individual words as nodes, and frequently co-occurring bigrams formed edges connecting these nodes. A selective threshold was applied, including only bigrams that appeared at least 70 times (min\_bigram\_n = 70), to ensure the networks highlighted only the most significant relationships and reduced visual noise. The opacity and thickness of these edges directly corresponded to the bigram's frequency, visually indicating the strength of the connection.

The network for Inexpensive Wines (Figure 5) visually demonstrates clusters of words that frequently appear together in straightforward combinations. Prominent strong links are observed between 'red' and 'fruit', 'soft' and 'palate', and 'light' and 'body'. These connections underscore the emphasis on basic, palatable, and easily identifiable sensory attributes. For example, the visual clearly shows how 'soft' directly links to 'palate', indicating a common and defining descriptor for this category. The network appears less complex and more direct, reflecting the simpler and more accessible language used in reviews.

The network for Expensive Wines (Figure 6) exhibits a noticeably more intricate and often denser structure, with connections emphasizing specific varietals, complex flavor combinations, and structural characteristics. Strong links are clearly visible between 'cabernet' and 'sauvignon', 'black' and 'cherry', 'fine' and 'tannins', and 'long' and 'finish'. Other key connections derived from your top bigrams (e.g., 'fruit flavors', 'palate offers', 'black fruit', 'black pepper', 'cabernet franc', 'black currant', 'petit verdot') would also contribute to this complex web. This visual complexity unequivocally reflects the richer, more specialized, and detailed vocabulary employed in reviews of higher-priced wines, showcasing a sophisticated appreciation for nuanced attributes.

These bigram networks offer invaluable insights into the "semantic fields" that distinctly define each wine category. The winery can directly leverage these strongly correlated word pairings in their communication strategies. For instance, knowing that 'black cherry' and 'cassis' are strongly linked for a premium wine allows for highly precise and evocative tasting notes. Conversely, emphasizing the pairing of 'red' and 'fruit' for an inexpensive wine can effectively convey its approachable character. Furthermore, this analysis can inform sales training, enabling staff to use language that directly aligns with consumer perceptions and expectations for each wine tier, thereby enhancing both communication effectiveness and sales outcomes.

# Conclusion

Through this comprehensive text analytics project, I successfully illuminated the distinct linguistic landscapes within wine reviews, effectively differentiating between descriptions of inexpensive and expensive wines. By meticulously applying various text mining frameworks, including frequency analysis of words and bigrams, sentiment profiling, topic modeling, and network analysis, I identified clear patterns in how consumers articulate their experiences across different price points.

My findings consistently demonstrate that reviews for expensive wines employ a more refined, specific, and structurally focused vocabulary, often centered on varietal characteristics, complex flavor profiles, and detailed sensory experiences. In contrast, inexpensive wines are typically described with broader, more common terms, emphasizing how easy they are to enjoy and their general fruitiness. While reviews for both types of wine were largely positive, the subtle qualitative differences in language provide critical strategic opportunities.

For the California winery, these insights are directly actionable. Marketing efforts can be finely tuned to resonate with the specific language and expectations of each consumer segment. For premium wines, emphasizing complexity, unique flavor combinations, and the depth of the wine's character will likely drive stronger engagement. For more affordable wines, promoting their fruit-forwardness, approachable nature, and immediate enjoyment will be more effective. Ultimately, by aligning marketing narratives with the authentic voice of the consumer, the winery can enhance brand perception, optimize product positioning, and foster stronger relationships with its audience across all its wines. This project underscores the significant value of text analytics in transforming unstructured data into powerful business intelligence for strategic decision-making.