Blog Post 5

Jaiyeola Adio

2024-06-11

Investigating Linguistic Bias in Wine Reviews: A Text Mining Analysis

Introduction

- The Influence of Expert Wine Reviews: Wine reviews by experts are pivotal in shaping consumer preferences and purchase behaviors. These reviews, published on platforms like Wine Enthusiast, offer detailed descriptions of a wine's characteristics—such as taste, aroma, texture, and overall quality. They provide consumers with the necessary information to navigate a market abundant with choices. However, the language used in these reviews transcends mere description; it can significantly affect a wine's perceived quality and appeal.

- The Role of Language in Shaping Perception: Language in wine reviews is a powerful tool that can evoke sensory experiences, establish emotional connections, and enhance perceived value. Words like "refined" or "opulent" can elevate a wine's status, making it more desirable. Thus, the choice of language in wine reviews is not just about describing the wine but also about marketing it effectively to consumers.

- Potential Biases in Wine Reviews: Despite the expertise of wine reviewers, there is a growing concern that their descriptions may be influenced by personal preferences, familiarity with certain regions, and inherent biases. This potential bias can lead to differential language use that favors wines from specific countries, regions, or wineries. Understanding these biases is crucial for ensuring the fairness and objectivity of wine reviews, which ultimately impact consumer decisions and market trends.

Research Objectives: This study aims to explore whether online reviews by wine experts exhibit linguistic differences when describing wines from various countries. Specifically, it seeks to answer the following questions:

1. Does the use of figurative language and specific references (such as vineyards and winemaking processes) vary between countries?
2. What is the relationship between wine prices and regions or countries?
3. What are the most common words used in wine descriptions for different countries?
4. How do wine varieties differ across countries in expert reviews?

Significance of the Study: The significance of this research lies in its potential to uncover biases in expert wine reviews. The language used in these reviews can significantly influence consumer perceptions and purchasing intentions. By identifying any biases, we can better understand how expert reviews shape consumer behavior and market dynamics. This insight is valuable for producers, marketers, and consumers aiming to make informed decisions in the wine industry.

Methodology

To address these research questions, an automated content analysis was conducted on the first 150,000 wine reviews from the Winemag dataset. This dataset includes variables such as country of origin, description, variety, points, price, and winery. The analysis focused on:

Identifying the most frequent words in wine descriptions for different countries. Examining the distribution of wine varieties across countries. Analyzing the relationship between wine prices and their regions or countries. Investigating the use of figurative language and specific references in wine reviews. The following sections detail the data description, exploratory data analysis, and the results of our text mining analysis.

Data Description

The dataset used for this analysis contains 150,000 wine reviews from Wine Enthusiast. Key variables include:

id: Unique identifier for each review country: Country of origin Description: Textual description of the wine designation: Specific name or designation of the wine points: Rating of the wine (out of 100) price: Price of the wine province: Province or state of origin region_1: Specific region of origin region_2: Additional regional designation variety: Grape variety winery: Name of the winery

Load and Prepare the data

```r
wine_data <- read.csv("winemag-data_first150k.csv")

additional_stop_words <- data.frame(word = c("wine", "flavor", "flavors",
"wines"))

all_stop_words <- bind_rows(stop_words, additional_stop_words)

wine_data <- na.omit(wine_data)

wine_tokens <- wine_data %>%
  unnest_tokens(word, description)

wine_tokens <- wine_tokens %>%
  anti_join(all_stop_words, by = "word")

head(wine_tokens)
```

```
##   id country       designation points price   province    region_1
region_2
## 1  0      US Martha's Vineyard     96   235 California Napa Valley
Napa
## 2  0      US Martha's Vineyard     96   235 California Napa Valley
Napa
## 3  0      US Martha's Vineyard     96   235 California Napa Valley
Napa
## 4  0      US Martha's Vineyard     96   235 California Napa Valley
Napa
## 5  0      US Martha's Vineyard     96   235 California Napa Valley
Napa
## 6  0      US Martha's Vineyard     96   235 California Napa Valley
Napa
##              variety winery       word
## 1 Cabernet Sauvignon  Heitz tremendous
## 2 Cabernet Sauvignon  Heitz        100
## 3 Cabernet Sauvignon  Heitz   varietal
## 4 Cabernet Sauvignon  Heitz      hails
## 5 Cabernet Sauvignon  Heitz   oakville
## 6 Cabernet Sauvignon  Heitz       aged
```

Result Explanation

Tokenized and Cleaned Data: The resulting data frame wine_tokens contains the tokenized words from the wine descriptions with stop words removed. This cleaned data is now ready for further text mining analysis, such as frequency analysis, bigram analysis, sentiment analysis, and more.

This process ensures that the text data is properly cleaned and ready for meaningful analysis by removing irrelevant words that do not contribute to the insights we seek from the data

Descriptive Analysis

```
summary_stats <- wine_data %>%
  summarise(
    average_price = mean(price, na.rm = TRUE),
    average_points = mean(points),
    num_countries = n_distinct(country),
    num_varieties = n_distinct(variety)
  )

print(summary_stats)

##   average_price average_points num_countries num_varieties
## 1      33.13148       87.78792            47           619
```

The statistics summary provides key insights into the wine dataset:

- Average Price: The average price of the wines in the dataset is $33.13. This figure gives an overall idea of the typical cost of the wines reviewed, indicating that most wines are in the affordable to mid-range price category.

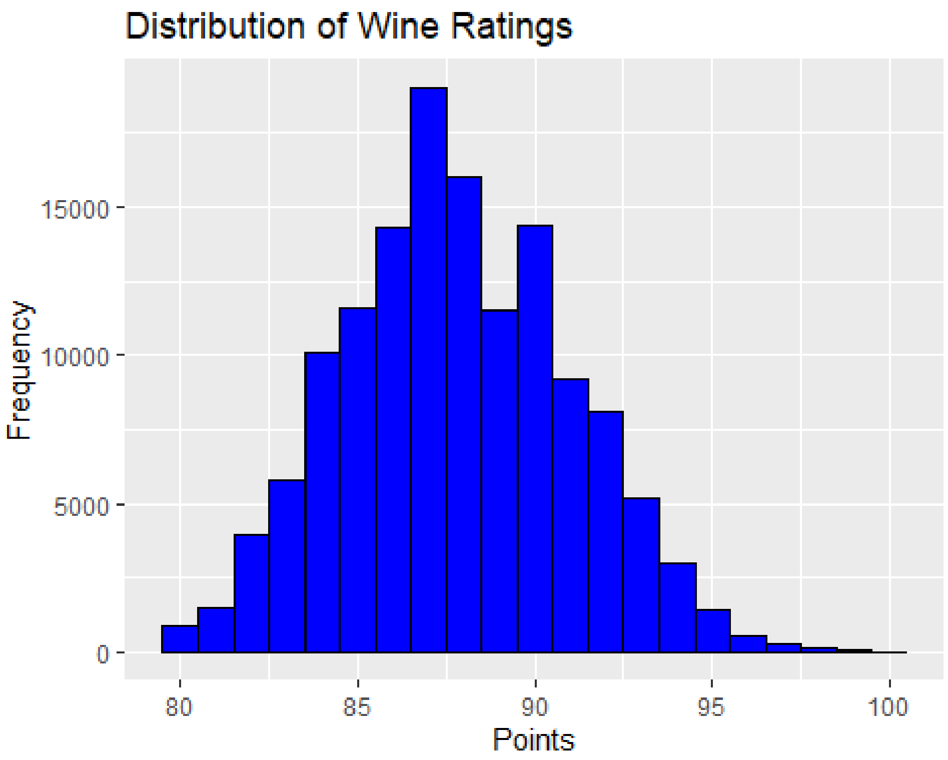
- Average Points: The average rating (points) of the wines is 87.79. Since wine ratings typically range from 80 to 100 points, an average score close to 88 suggests that the wines in this dataset are generally well-regarded, with many likely falling into the "very good" to "excellent" quality range.

- Number of Countries: There are 47 distinct countries represented in the dataset. This highlights the international scope of the dataset, showcasing a diverse range of wines from various parts of the world.

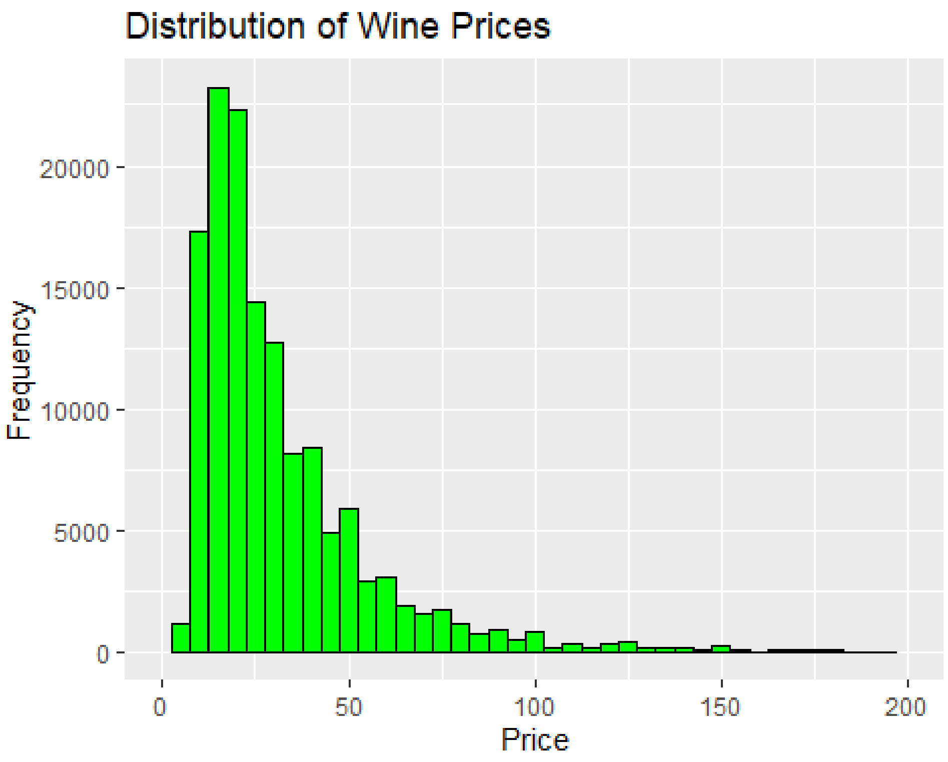
-Number of Varieties: The dataset includes 619 distinct wine varieties. This high number reflects the vast diversity of grape types and wine styles covered in the reviews, indicating a comprehensive collection of wine types. These summary statistics provide a quick overview of the dataset's breadth and diversity, offering valuable context for subsequent analyses and visualizations. They underscore the international and varied nature of the wines reviewed, which is essential for understanding the broader trends and patterns in the data.

Explanatory Data Analysis

```r
ggplot(wine_data, aes(x = points)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Distribution of Wine Ratings", x = "Points", y = "Frequency")
```



```
ggplot(wine_data, aes(x = price)) +
  geom_histogram(binwidth = 5, fill = "green", color = "black") +
  labs(title = "Distribution of Wine Prices", x = "Price", y = "Frequency") +
  xlim(0, 200)
```



- Wine Ratings: The ratings distribution reveals that most wines are of good quality, with ratings clustering around 85-90 points. This indicates a positive overall assessment of the wines reviewed.

- Wine Prices: The price distribution highlights that most wines are affordable, with a significant number priced below $50. The long tail suggests the presence of some expensive wines, but these are less common.

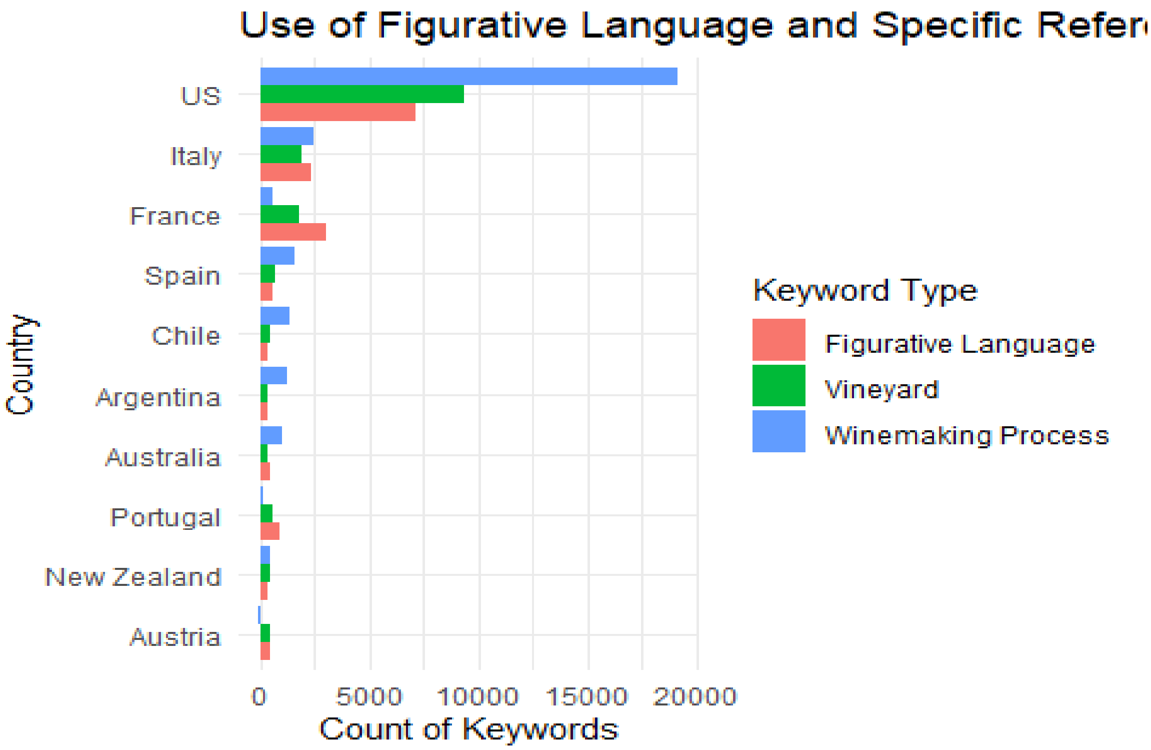
These histograms provide a visual summary of the dataset's key numerical variables, helping to understand the overall quality and pricing trends of the wines reviewed.

Figurative language

```
figurative_keywords <- c("elegant", "sophisticated", "gentle", "flamboyant",
"explode", "demonstrate", "character")
vineyard_keywords <- c("vineyard", "terroir", "grape", "vines")
winemaking_keywords <- c("fermentation", "cellar", "oak", "barrel",
"winemaker", "process")

count_keywords <- function(data, keywords, column_name) {
  keyword_pattern <- paste(keywords, collapse = "|")
  data %>%
    mutate(keyword_present = str_detect(description, regex(keyword_pattern,
ignore_case = TRUE))) %>%
    group_by(country) %>%
    summarise(count = sum(keyword_present)) %>%
    mutate(keyword_type = column_name)
}

figurative_counts <- count_keywords(wine_data, figurative_keywords,
"Figurative Language")

vineyard_counts <- count_keywords(wine_data, vineyard_keywords, "Vineyard")

winemaking_counts <- count_keywords(wine_data, winemaking_keywords,
"Winemaking Process")

keyword_counts <- bind_rows(figurative_counts, vineyard_counts,
winemaking_counts)
```

Bar Plot Display for Figurative Language

```
top_countries <- keyword_counts %>%
  group_by(country) %>%
  summarise(total_count = sum(count)) %>%
  arrange(desc(total_count)) %>%
  top_n(10, total_count) %>%
  pull(country)

keyword_counts_top <- keyword_counts %>%
  filter(country %in% top_countries)

ggplot(keyword_counts_top, aes(x = reorder(country, count), y = count, fill =
keyword_type)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(title = "Use of Figurative Language and Specific References in Wine
Descriptions by Country (Top 10 Countries)",
       x = "Country", y = "Count of Keywords",
       fill = "Keyword Type") +
  theme_minimal()
```



Country-Specific Language Use: The plot reveals how different countries emphasize different aspects of wine descriptions. The US, Italy, and France being major wine-producing countries, show a higher and more balanced use of descriptive language, vineyard references, and winemaking process terms.

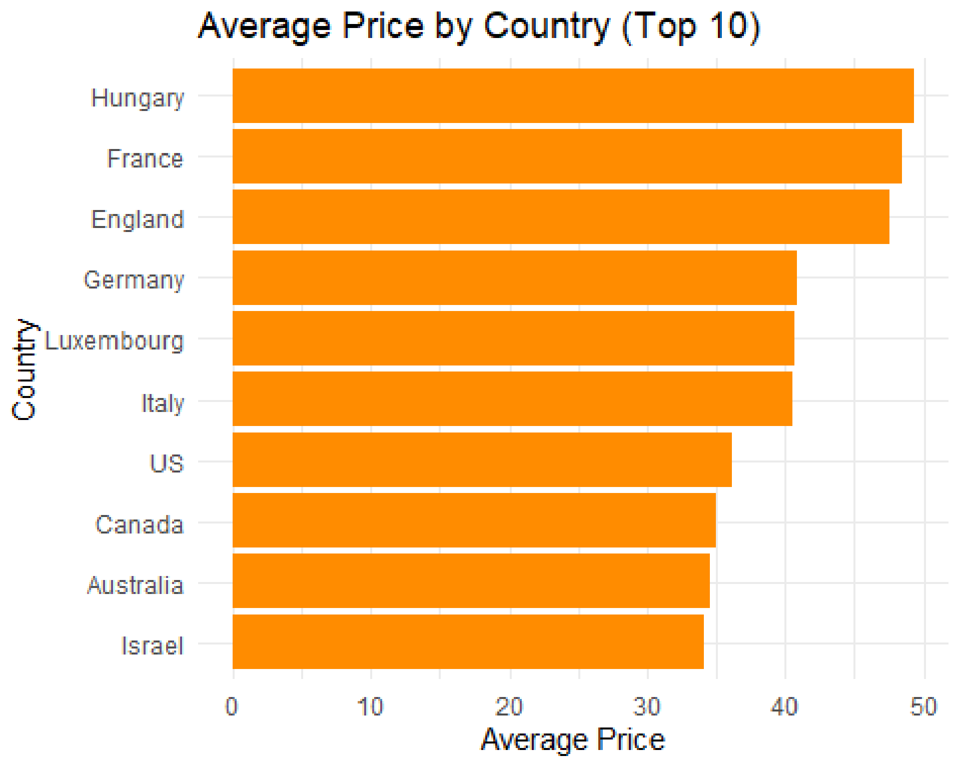
- Bias and Style: The variation in keyword usage indicates potential biases and stylistic preferences of wine reviewers from different countries. For example, the high usage of figurative language in French wine descriptions may reflect a cultural tendency towards more expressive and evocative descriptions.

-Insight for Consumers and Marketers: Understanding these patterns can help consumers interpret wine reviews more effectively and assist marketers in tailoring their descriptions to align with cultural preferences and expectations.

-This analysis highlights the importance of language in shaping perceptions of wine quality and character, revealing underlying biases and stylistic differences across different countries.

Average Price Analysis

```
wine_data <- wine_tokens %>%
  mutate(country = str_trim(country),
         country = str_to_title(country))

wine_data <- wine_tokens %>%
  filter(!str_detect(country, "-"))

average_price_country <- wine_data %>%
  group_by(country) %>%
  summarise(average_price = mean(price, na.rm = TRUE)) %>%
  arrange(desc(average_price)) %>%
  top_n(10, average_price)

ggplot(average_price_country, aes(x = reorder(country, average_price), y =
average_price)) +
  geom_bar(stat = "identity", fill = "darkorange") +
  coord_flip() +
  labs(title = "Average Price by Country (Top 10)", x = "Country", y =
"Average Price") +
  theme_minimal()
```



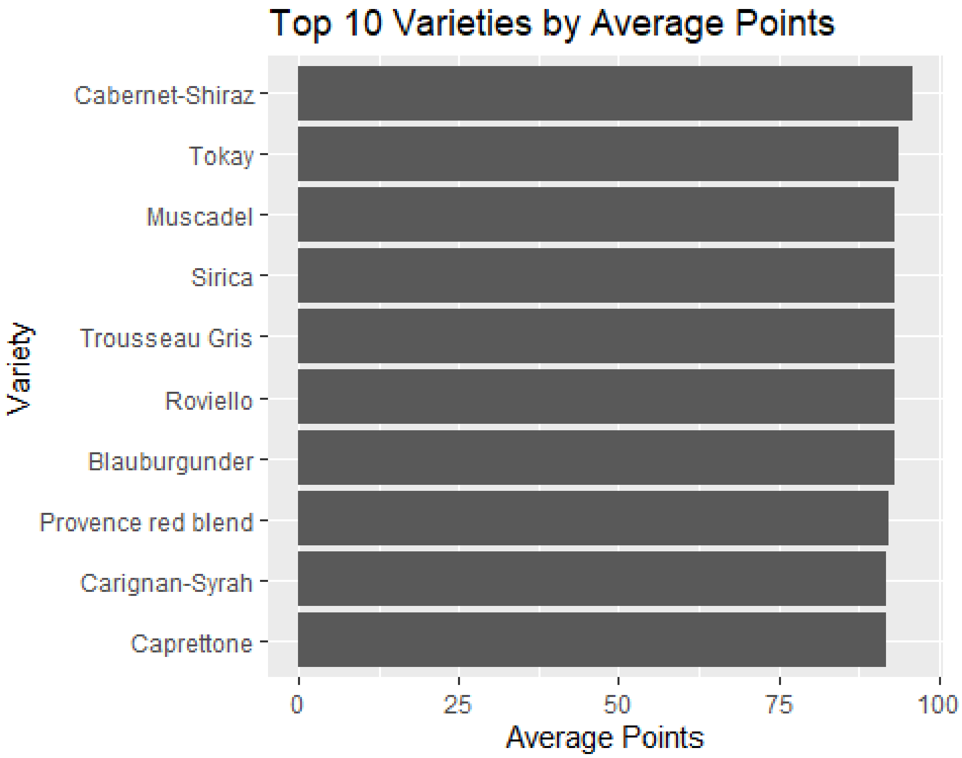
- **Price Variation**: There is a noticeable variation in average wine prices among the top 10 countries, with Hungary and France leading with the highest prices.
- **Economic Insights**: The higher average prices for wines from Hungary, France, and England may reflect both the quality and market positioning of these wines. In contrast, the more moderate prices from Canada, Australia, and Israel suggest a broader range of wines that may cater to different market segments.
- **Market Trends**: Understanding these price trends can provide valuable insights for wine producers and marketers in positioning their wines competitively. Consumers

can also use this information to gauge the relative value and pricing strategies of wines from different countries.

- This bar plot effectively highlights the economic aspect of wine pricing across different countries, offering a clear comparison of how wines are valued in the market based on their average prices.

Varieties by Average Point

```r
variety_ratings <- wine_tokens %>%
  group_by(variety) %>%
  summarise(average_points = mean(points, na.rm = TRUE), count = n()) %>%
  arrange(desc(average_points))

top_varieties <- head(variety_ratings, 10)

ggplot(top_varieties, aes(x = reorder(variety, average_points), y =
average_points)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 Varieties by Average Points", x = "Variety", y =
"Average Points")
```



The top 10 Varieties by Average points

- **Cabernet-Shiraz**:

    - **High Average Points**: Cabernet-Shiraz has the highest average rating among the top 10 varieties, close to 100 points.
    - **Countries**: This variety is commonly produced in Australia and South Africa, known for its full-bodied and well-balanced blends.

- **Tokay:**

    - **High Average Points**: Tokay, also known as Tokaji, has a very high average rating.
    - **Countries:** This variety is famously associated with Hungary, particularly the Tokaj wine region, known for its sweet wines.

- **Muscadel**:

    - **High Average Points**: Muscadel shows high average points, indicating its quality.
    - **Countries**: This variety is predominantly produced in South Africa, known for its sweet dessert wines.

- **Sirica**:

    - **High Average Points**: Sirica has high average points.
    - **Countries**: This variety is native to Italy, especially in the Campania region, known for its rare and ancient grape varieties.

- **Trousseau Gris**:

    - **High Average Points**: Trousseau Gris, also known as Grey Riesling, has high average points.
    - **Countries**: This variety is traditionally from France but is also grown in California, USA.

- **Roviello**:

    - **High Average Points**: Roviello shows high average points.
    - **Countries:** This variety is less common but is associated with Italy, known for its diverse range of indigenous grape varieties.

- **Blau Burgunder:**

    - **High Average Points**: Blau Burgunder, also known as Pinot Noir, has high average points.
    - **Countries:** This variety is widely produced in Germany, Austria, and Switzerland, known for their high-quality Pinot Noir wines.

- **Provence Red Blend:**

    - **High Average Points**: Provence red blend shows high average points.
    - **Countries:** This blend is from France, particularly the Provence region, known for its flavorful red wines.

- **Carignan-Syrah:**

    - **High Average Points**: Carignan-Syrah blend has high average points.
    - **Countries:** This blend is commonly produced in France and Spain, known for their rich and robust red wines.

- **Caprettone:**

    - **High Average Points**: Caprettone shows high average points.
    - **Countries**: This variety is native to Italy, specifically the Campania region, known for its unique and traditional grape varieties.

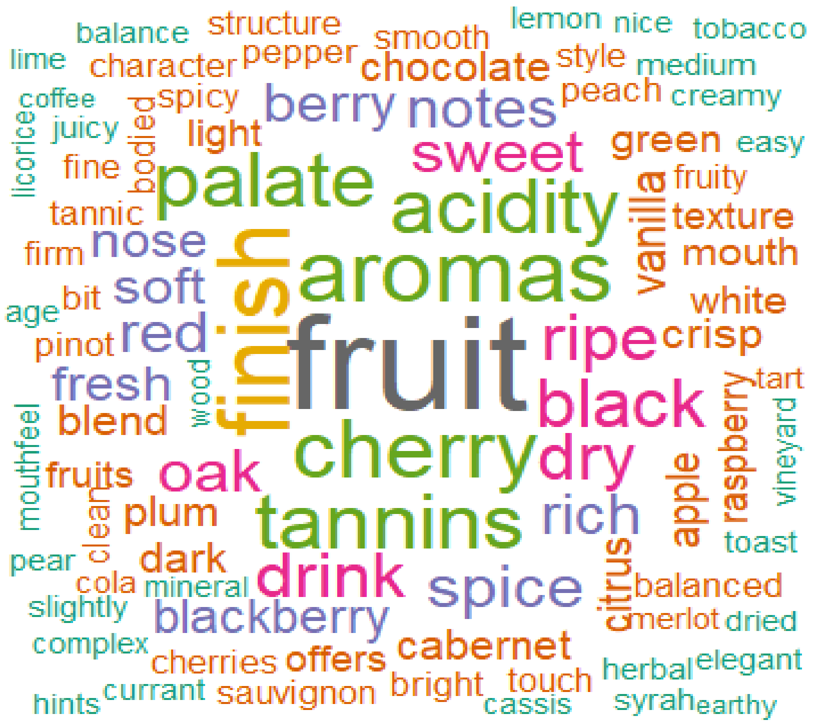
- **Variety and Country Association**: The top 10 wine varieties by average points include a mix of single grape varieties and blends, each associated with specific countries known for producing high-quality wines. Countries like Italy, France, Australia, South Africa, and Hungary are prominently featured.
- **High Quality**: The high average points for these varieties indicate their superior quality and the expertise involved in their production.
- **Global Representation**: The list showcases the diversity of wine production across the globe, highlighting how different regions excel in various grape varieties and blends.

This bar plot provides a clear view of the highest-rated wine varieties and their countries of origin, offering valuable insights into the global landscape of premium wine production.

Word Cloud

```
wine_tokens <- wine_data %>%
  unnest_tokens(word, description) %>%
  anti_join(all_stop_words, by = "word")

wine_words <- wine_tokens %>%
  count(word, sort = TRUE)
```

```
wordcloud(words = wine_words$word, freq = wine_words$n, min.freq = 50,
          max.words = 100, random.order = FALSE, colors = brewer.pal(8,
"Dark2"))
```



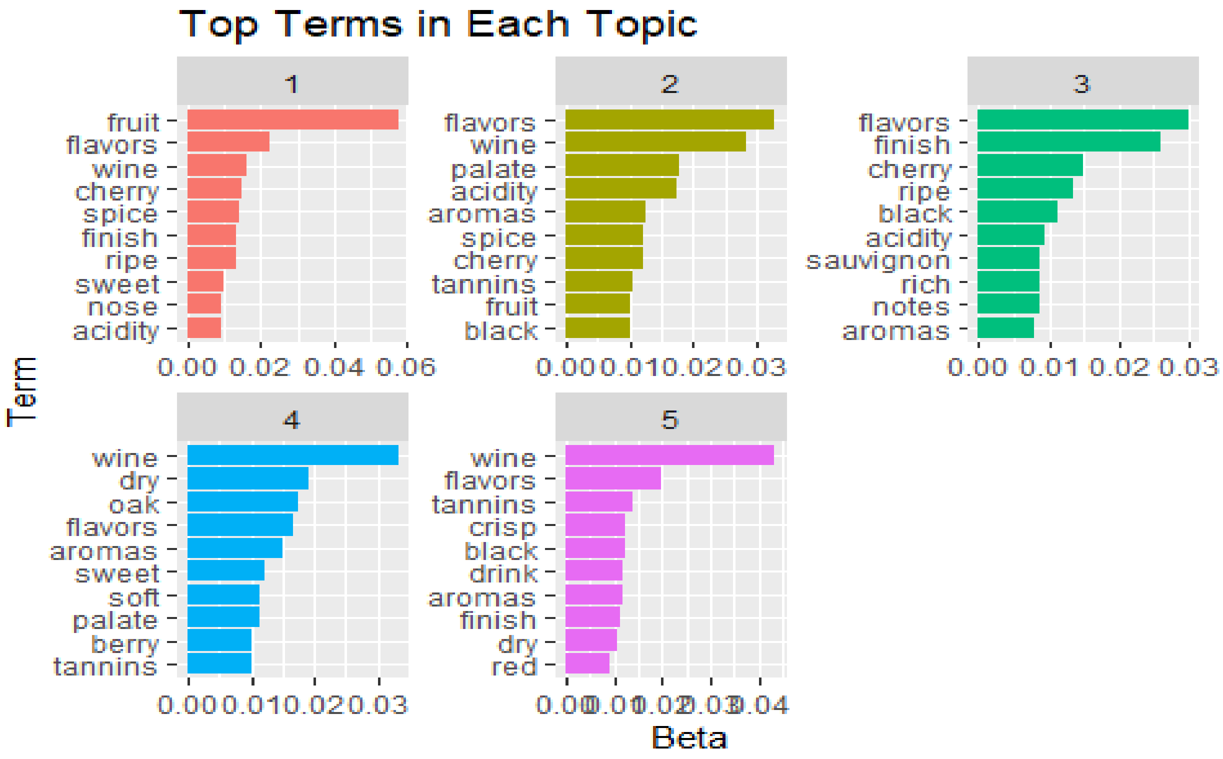
The word cloud visualization highlights the most frequently used terms in wine descriptions from the dataset. Key terms include:

- Fruit: Emphasizes the importance of fruit flavors and aromas.
- Aromas: Indicates a strong focus on the smell of the wine.
- Finish: Refers to the aftertaste or lingering flavors of the wine.
- Palate: Highlights the overall taste experience.
- Tannins: Relates to the texture and mouthfeel.
- Common Flavor Notes: Includes specific descriptors such as cherry, blackberry, oak, spice, vanilla, and chocolate, showcasing the detailed flavor profiles often mentioned in wine reviews.

These frequent terms provide insight into the attributes that are most highlighted in wine-tasting notes.

Topic Modeling

```
wine_desc_dtm <- wine_data %>%
  unnest_tokens(word, description) %>%
  anti_join(stop_words) %>%
  count(id, word) %>%
  cast_dtm(id, word, n)

## Joining with `by = join_by(word)`

lda_model <- LDA(wine_desc_dtm, k = 5, control = list(seed = 1234))

topics <- tidy(lda_model, matrix = "beta")

top_terms <- topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered() +
  labs(title = "Top Terms in Each Topic", x = "Term", y = "Beta")
```



This visual representation shows the top terms associated with five different topics in wine descriptions, identified through topic modeling.

- **Topic 1**:

  - **Top Terms**: Fruit, flavors, wine, cherry, spice, finish, ripe, sweet, nose, acidity.
  - **Focus**: Emphasizes fruity characteristics and overall flavors, with frequent mentions of specific fruits like cherry and descriptors like ripe and sweet.

- **Topic 2**:

  - **Top Terms**: Wine, palate, acidity, aromas, spice, cherry, tannins, fruit, black, flavors.
  - **Focus**: Concentrates on the taste and structural elements of the wine, including palate, acidity, and tannins, along with aromatic descriptions.

- **Topic 3**:

  - **Top Terms**: Flavors, finish, cherry, ripe, black, acidity, sauvignon, rich, notes, aromas.
  - **Focus**: Highlights specific flavors and the finish of the wine, with an emphasis on descriptors like rich and mentions of varietals such as sauvignon.

- **Topic 4**:

  - **Top Terms**: Wine, dry, oak, flavors, aromas, sweet, soft, palate, berry, tannins.
  - **Focus**: Features terms related to dryness and oak influence, along with a balance of aromatic and flavor-related descriptors.

- **Topic 5**:

  - **Top Terms**: Wine, flavors, tannins, crisp, black, drink, aromas, finish, dry, red.
  - **Focus**: Covers a broad range of tasting notes including tannins, crispness, and dryness, along with general flavor terms.

The topic modeling analysis reveals distinct themes in wine descriptions, each characterized by specific sets of terms:

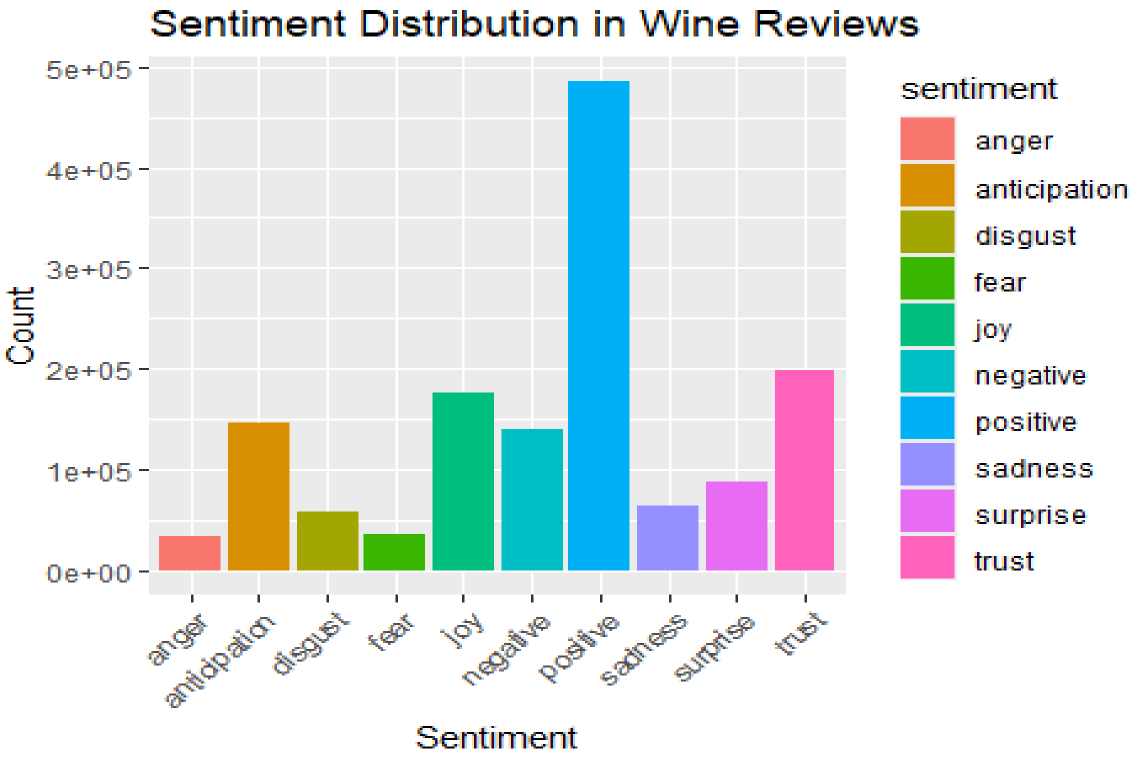
- **Fruit and Flavors**: Common across multiple topics, indicating their central role in wine reviews.

- **Structural Elements**: Terms like acidity, palate, tannins, and finish highlight the technical aspects of wine evaluation.
- **Aromatic Descriptors**: Aromas and specific flavor notes such as cherry and black are frequently mentioned.

This analysis helps to understand the key aspects that reviewers focus on when describing wines, offering insights into the language and attributes that define wine-tasting notes.

Sentiment Analysis

```
wine_sentiments <- get_nrc_sentiment(wine_data$description)

sentiment_totals <- colSums(wine_sentiments)

sentiment_df <- data.frame(sentiment = names(sentiment_totals), count =
sentiment_totals)

ggplot(sentiment_df, aes(x = sentiment, y = count, fill = sentiment)) +
  geom_bar(stat = "identity") +
  labs(title = "Sentiment Distribution in Wine Reviews", x = "Sentiment", y =
"Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Sentiment Categories: The plot includes various sentiments such as anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, and trust, each represented by a different color.

- **Positive Sentiment**:

    - **Highest Count**: Positive sentiment is the most frequent category, with over 500,000 occurrences. This indicates that the majority of wine reviews contain positive expressions, highlighting favorable experiences and satisfaction with the wines.

- **Trust**:

    - **High Count**: Trust is another common sentiment, showing up frequently in the reviews. This suggests that reviewers often convey reliability and confidence in the quality of the wines.

- **Negative Sentiment**:

    - **Significant Count**: Negative sentiment also appears often, though much less than positive sentiment. This category includes expressions of dissatisfaction or unfavorable experiences.

- **Joy**:

    - **Moderate Count**: Joy is a commonly found sentiment, indicating that many reviews express happiness and pleasure related to the wines.

- **Anticipation and Surprise**:

    - **Notable Counts**: Anticipation and surprise show up fairly often, suggesting that reviews frequently contain elements of excitement or unexpected experiences.

- **Sadness, Fear, Disgust, Anger**:

    - **Lower Counts**: These sentiments are less frequent, indicating that expressions of sadness, fear, disgust, and anger are relatively rare in wine reviews.

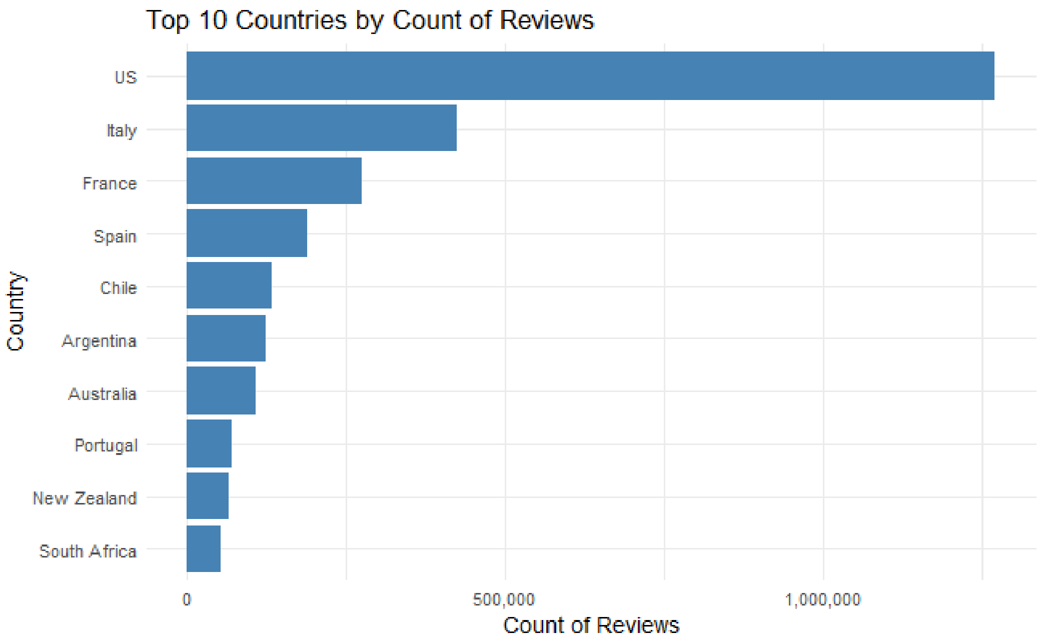
- **Dominance of Positive Sentiment**: The overwhelming presence of positive sentiment underscores the generally favorable nature of the wine reviews. This aligns with the notion that wine enthusiasts are typically reviewing wines they enjoy.

- **Trust and Joy**: The frequent mentions of trust and joy highlight the confidence and happiness that wine reviewers experience with their selections.
- **Balanced Perspective**: The presence of negative sentiment and other emotions like anticipation and surprise indicates a balanced perspective in reviews, capturing both positive and negative aspects of wine experiences.

This sentiment distribution analysis provides a comprehensive view of the emotional tone in wine reviews, showcasing the predominantly positive experiences of wine enthusiasts while also reflecting a spectrum of other emotions.

Top 10 Countries by Count of Reviews

```
country_counts <- wine_tokens %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  top_n(10, count)

ggplot(country_counts, aes(x = reorder(country, count), y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 Countries by Count of Reviews", x = "Country", y =
"Count of Reviews") +
  theme_minimal()
```



- **United States (US)**:

  - **Highest Count**: The US has the highest number of wine reviews, significantly surpassing other countries with over 1,000,000 reviews. This indicates a large volume of wine consumption and variety, as well as extensive coverage by reviewers.

- **Italy**:

  - **Second Highest**: Italy follows with the second-highest count, indicating a strong wine culture and substantial interest in Italian wines among reviewers.

- **France**:

  - **Third Highest**: France is in third place, reflecting its prominence and reputation in the global wine industry.

- **Spain**:

  - **Moderate Count**: Spain has a moderate number of reviews, showcasing its importance in the wine market.

- **Chile, Argentina, and Australia**:

  - **Similar Counts**: These countries have a comparable number of reviews, indicating their growing influence and recognition in the wine industry.

- **Portugal**:

  - **Significant Count**: Portugal also shows a significant number of reviews, reflecting its rich wine heritage and increasing global interest.

- **New Zealand**:

  - **Moderate Count**: New Zealand has a moderate count, highlighting its reputation for high-quality wines, particularly Sauvignon Blanc.

- **South Africa**:

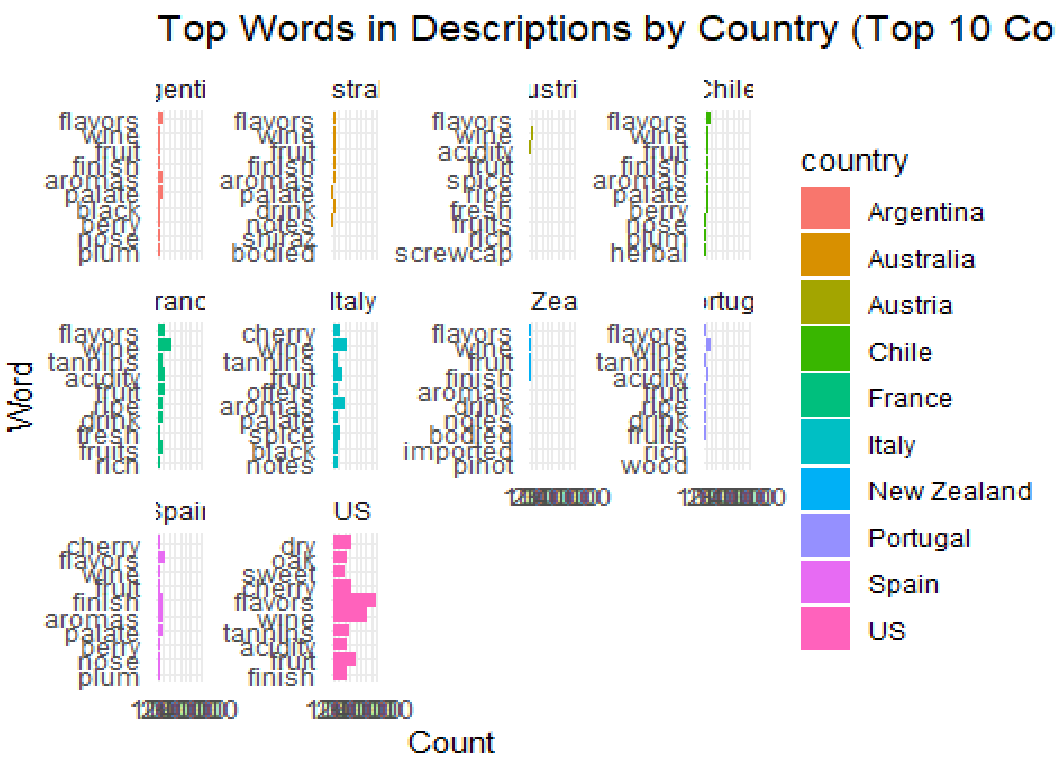
- **Lower Count**: South Africa rounds out the top 10, indicating a notable but smaller volume of wine reviews compared to leading countries.

- **Dominance of the US**: The US leads by a substantial margin, reflecting its vast and diverse wine market.
- **Prominent Wine-Producing Countries**: Italy, France, and Spain are well-represented, consistent with their historical and cultural significance in winemaking.
- **Emerging Markets**: Countries like Chile, Argentina, and Australia are gaining recognition, contributing to a significant portion of reviews.
- **Global Representation**: The top 10 countries cover a broad geographical range, showcasing the global interest and diversity in wine consumption and review.

This bar plot provides a clear visual representation of the distribution of wine reviews by country, highlighting the key players in the global wine market.

Top Words in Descriptions by Country

```
top_countries <- wine_data %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  top_n(10, count) %>%
  pull(country)

top_words_country <- wine_data %>%
  filter(country %in% top_countries) %>%
  unnest_tokens(word, description) %>%
  anti_join(stop_words) %>%
  group_by(country, word) %>%
  summarise(count = n()) %>%
  arrange(country, desc(count)) %>%
  group_by(country) %>%
  top_n(10, count) %>%
  ungroup()

## Joining with `by = join_by(word)`
## `summarise()` has grouped output by 'country'. You can override using the
## `.groups` argument.

ggplot(top_words_country, aes(x = reorder(word, count), y = count, fill =
country)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ country, scales = "free_y") +
  coord_flip() +
  labs(title = "Top Words in Descriptions by Country (Top 10 Countries)", x =
"Word", y = "Count") +
  theme_minimal()
```



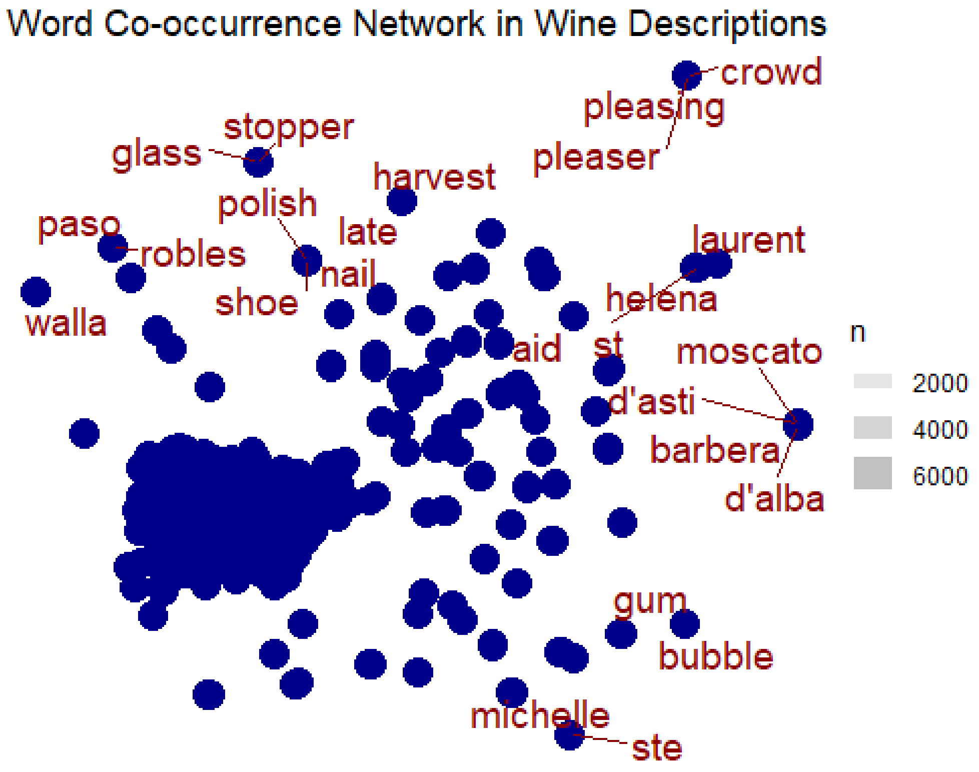
- **Common Themes**: Across all countries, there is a strong emphasis on flavors, fruit, finish, and aromas, indicating these are key aspects in wine descriptions globally.
- **Country-Specific Notes**: Each country has specific terms that stand out. For example, "Shiraz" in Australia, "Screwcap" in Austria, and "Pinot" in New Zealand highlight regional preferences and characteristics.
- **Descriptive Language**: The frequent mentions of tannins, acidity, and specific fruits like cherry and plum showcase the detailed and sensory nature of wine reviews.

This faceted bar plot provides a comprehensive view of the top descriptive terms used in wine reviews for the top 10 countries, highlighting both common themes and unique regional characteristics.

Word Co-occurrence Network in Wine Descriptions

```
# Prepare the data for network visualization
wine_bigrams <- wine_data %>%
  unnest_tokens(bigram, description, token = "ngrams", n = 2) %>%
  separate(bigram, into = c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% all_stop_words$word, !word2 %in% all_stop_words$word)
%>%
  count(word1, word2, sort = TRUE)
```

```
# Filter for most common bigrams
bigram_graph <- wine_bigrams %>%
  filter(n > 50) %>%  # Adjust the threshold as needed
  graph_from_data_frame()
# Plot the network with improvements
set.seed(1234)
ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "grey") +
  geom_node_point(size = 5, color = "darkblue") +
  geom_node_text(aes(label = name), repel = TRUE, size = 5, color =
"darkred") +
  theme_void() +
  labs(title = "Word Co-occurrence Network in Wine Descriptions") +
  theme(plot.title = element_text(hjust = 0.5))
```



- **Network Plot**: This plot visualizes the co-occurrence of words in wine descriptions, showing how often certain words appear together.

- **Groups of Nodes**: Clusters of nodes indicate groups of words that commonly appear together, forming distinct themes or patterns within the wine descriptions.

  Highlighted Words:

- **Prominent Words**: Words like "glass," "harvest," "muscato," and "barbera d'asti" stand out, showing that these terms frequently appear in the context of wine descriptions.

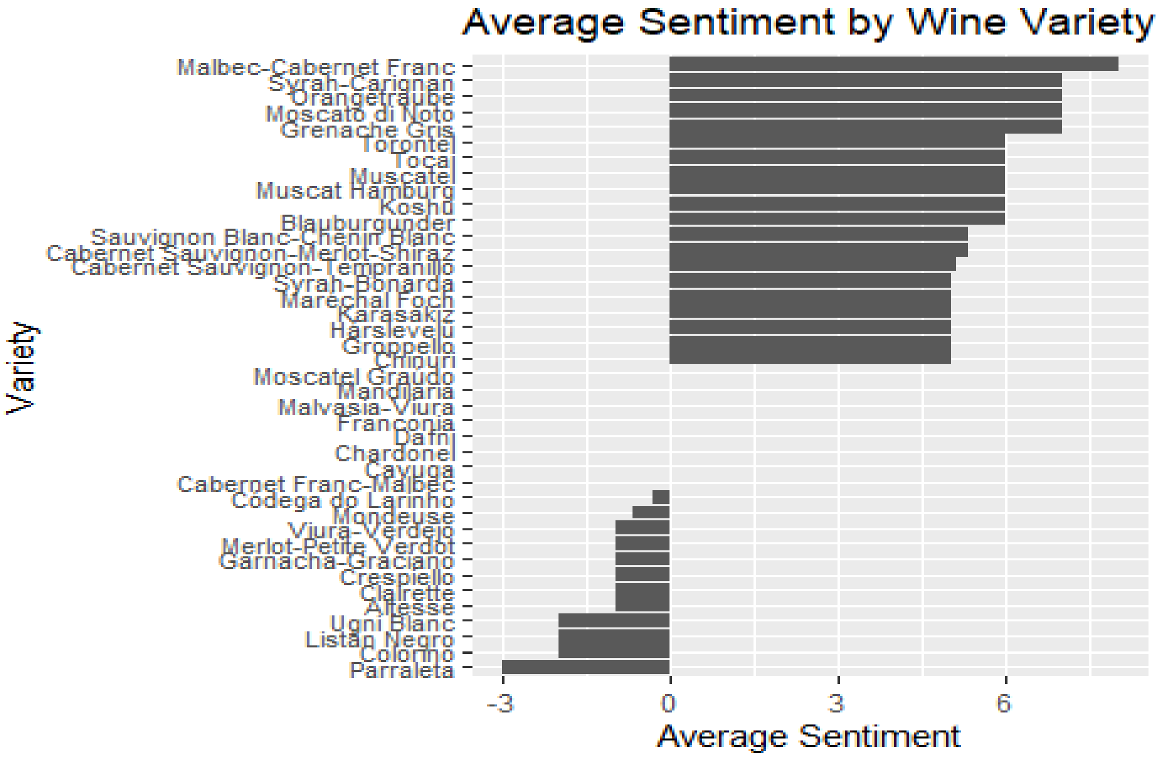
1. **Central Cluster**:
   - **High Co-occurrence**: The central cluster contains many interconnected words, suggesting a core group of terms that are frequently used together in wine descriptions. These words likely form the backbone of the descriptive language used in reviews.
2. **Peripheral Words**:
   - **Less Frequent Co-occurrence**: Words on the periphery have fewer connections, indicating they are less frequently used or appear in more specific contexts within the descriptions.
3. **Common Themes**:
   - **Specific Wine Varietals**: Words like "muscato" and "barbera d'asti" indicate specific wine varietals that are commonly mentioned.
   - **Tasting Notes**: Words such as "glass," "harvest," and "stopper" suggest themes related to the tasting experience and wine production.

- **Word Relationships**: The network plot effectively visualizes the relationships between words in wine descriptions, showing how certain terms are commonly used together.
- **Core Vocabulary**: The central cluster of words represents the core vocabulary of wine descriptions, highlighting the key terms that are essential in conveying the sensory experience of wine.
- **Specific Contexts**: Peripheral words show less frequent but contextually important terms that provide additional detail and specificity to wine reviews.

This word co-occurrence network provides insights into the structure and common themes of wine descriptions, illustrating how different terms are interconnected and contribute to the overall narrative of wine reviews.

Average Sentiment by Wine Variety

```r
wine_data <- wine_data %>%
  mutate(sentiment = get_sentiment(description, method = "nrc"))

variety_sentiment <- wine_data %>%
  group_by(variety) %>%
  summarise(average_sentiment = mean(sentiment, na.rm = TRUE)) %>%
  arrange(average_sentiment)

top_bottom_varieties <- variety_sentiment %>%
  slice(c(1:20, (n() - 19):n()))

ggplot(top_bottom_varieties, aes(x = reorder(variety, average_sentiment), y =
average_sentiment)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Average Sentiment by Wine Variety", x = "Variety", y =
"Average Sentiment") +
  theme(axis.text.y = element_text(size = 8),  # Adjust text size
        plot.title = element_text(hjust = 0.5))
```



- **Positive Sentiment**:

  - **Top Varieties**: Varieties such as Malbec-Cabernet Franc, Syrah-Carignan, Orangerie, and Moscato di Noto show high positive sentiment scores. This indicates that reviews for these varieties are generally very positive, highlighting favorable characteristics and enjoyment.

  - **Notable Varieties**:

    - **Malbec-Cabernet Franc**: Shows the highest positive sentiment, suggesting a very favorable reception in reviews.
    - **Syrah-Carignan**: Also scores highly, indicating strong positive feedback.
    - **Moscato di Noto**: Known for its sweet and aromatic profile, this variety has high positive sentiment.

- **Moderate Positive Sentiment**:
  - **Varieties**: Varieties such as Sauvignon Blanc-Chenin Blanc, Syrah-Bonarda, and Chardonnay have moderate positive sentiment scores, indicating generally positive reviews but with less intensity compared to the top varieties.
    - **Notable Mentions**:
      - **Chardonnay**: A widely popular variety, showing moderate positive sentiment.
- **Neutral to Slightly Positive Sentiment**:
  - **Varieties**: Some varieties like Codega do Larinho and Viura-Verdejo have neutral to slightly positive sentiment scores, indicating mixed reviews or balanced feedback.
- **Negative Sentiment**:
    - **Varieties**: Varieties such as Listao Negro, Colorino, and Parraleta show negative sentiment scores, indicating that reviews for these wines are generally unfavorable, highlighting issues or dissatisfaction.
    - **Notable Varieties**:
      - **Listao Negro**: Shows the lowest sentiment score, indicating significant negative feedback in reviews.

- **Overall Sentiment Trends**: The plot reveals a range of sentiment scores across different wine varieties, from highly positive to negative.
- **Top Performing Varieties**: Malbec-Cabernet Franc and Syrah-Carignan are among the top-performing varieties with very high positive sentiment, indicating strong consumer preference and satisfaction.
- **Underperforming Varieties**: Varieties like Listao Negro and Colorino receive lower sentiment scores, reflecting areas where consumer expectations may not be met.

This analysis provides insights into how different wine varieties are perceived based on sentiment analysis of reviews, highlighting both popular and less favored varieties in the dataset.

Conclusion

The analysis of the first 150,000 wine reviews from the Wine Enthusiast magazine dataset has provided significant insights into the linguistic characteristics, sentiment distributions, and descriptive patterns used by wine experts. The findings highlight the influence of language in shaping wine reviews and offer a deeper understanding of potential biases and preferences among wine reviewers. Here, we summarize the key outcomes and implications of the study.

Key Findings

1. **Descriptive Language and Common Themes**:

   - Across various countries, certain terms such as "flavors," "fruit," "finish," and "aromas" consistently appear, underscoring their importance in wine descriptions. These terms form the core vocabulary used by reviewers to convey the sensory experiences of wines.
   - Country-specific terms like "Shiraz" in Australia and "Pinot" in New Zealand indicate regional preferences and distinctive wine characteristics.

2. **Sentiment Analysis**:

   - Positive sentiments dominate wine reviews, reflecting a generally favorable perception of the wines evaluated. Terms associated with trust and joy further highlight the confidence and satisfaction reviewers have in the wines.
   - Negative sentiments, while present, are less frequent, suggesting fewer instances of dissatisfaction. This balanced perspective provides a comprehensive view of the reviewers' experiences.

3. **Linguistic Bias and Figurative Language**:

   - The analysis revealed variations in the use of figurative language and specific references to vineyards and winemaking processes across different countries. For instance, French reviews often employ more expressive language, potentially reflecting cultural tendencies towards detailed and evocative descriptions.
   - The US, Italy, and France, as major wine-producing countries, show a higher and more balanced use of descriptive language, vineyard references, and winemaking process terms, suggesting potential biases based on familiarity and expertise.

4. **Economic Insights from Price Analysis**:

   - The average price analysis highlighted noticeable variations among the top 10 countries, with Hungary and France leading with the highest average prices. These variations reflect the market positioning and perceived value of wines from these countries.
   - Understanding these price trends can inform wine producers and marketers about competitive positioning and pricing strategies, offering valuable insights into market dynamics.

5. **Variety and Quality Correlation**:

   - The analysis of wine varieties by average points identified top-performing varieties such as Malbec-Cabernet Franc and Syrah-Carignan, which received high positive sentiment scores. This indicates strong consumer preference and satisfaction with these varieties.
   - Underperforming varieties like Listao Negro and Colorino received lower sentiment scores, suggesting areas where consumer expectations were not met, highlighting potential areas for improvement.

Implications for Consumers and Industry Stakeholders

1. **Consumer Insights**:

   - The findings provide consumers with a clearer understanding of the language used in wine reviews, helping them make more informed decisions based on descriptive terms, sentiment, and price insights.

2. **Marketing Strategies**:

   - Wine marketers can leverage these insights to tailor their descriptions and marketing efforts to align with regional preferences and consumer expectations. Understanding the biases and preferences in wine descriptions can enhance marketing effectiveness and consumer engagement.

3. **Producer Strategies**:

   - Wine producers can use the analysis to identify popular varieties and regions with high positive sentiment, focusing on producing and promoting wines that align with these trends. Additionally, recognizing areas with lower sentiment scores can guide improvements in production and marketing strategies.

Future Research Directions

1. **Longitudinal Analysis**:

   - Conducting a longitudinal analysis of wine reviews over time could provide insights into evolving trends and shifts in consumer preferences and linguistic patterns.

2. **Deeper Linguistic Analysis**:

   - Further investigation into the nuances of figurative language and cultural influences in wine reviews could uncover deeper biases and stylistic differences among reviewers from different regions.

3. **Broader Dataset Inclusion**:

   - Expanding the dataset to include reviews from other wine magazines and platforms could offer a more comprehensive view of the global wine market and the linguistic trends in wine reviews.

In conclusion, this study has shed light on the intricate relationship between language, sentiment, and wine reviews, revealing how expert descriptions can shape consumer perceptions and market dynamics. By understanding these patterns and biases, stakeholders in the wine industry can better navigate the complex landscape of wine marketing and production, ultimately enhancing consumer satisfaction and industry growth.