Exploring Patterns and Trends in Wine Quality Through Data Analysis and MapReduce Approach

Justin Bowers
Electrical Engineering and
Computer Science
University of Tennessee
Knoxville, Tennessee
jbower31@vols.utk.edu

Dong Jun Woun
Electrical Engineering and
Computer Science
University of Tennessee
Knoxville, Tennessee
dwoun@vols.utk.edu

Abstract—Customers must evaluate complex parameters for an informed wine purchase. To simplify the process, we present a study that assesses the trends in wine price, quality, and description in wine data. First, our approach cleaned the data by filling in missing values, removing outliers, and condensing duplicates. Afterward, we performed a statistical analysis of quality and price data and a MapReduce analysis on the description of the wine. We identified the country with the highest-rated wine but also concluded that a study on the description of wine brought no statistically significant results.

Keywords—ReduceMap, Data Analysis, Wine, and Data Cleaning

I. Introduction And Motivation

Data analysis is a fastly developing domain of computer science that concentrates on extracting information from various forms of data in mass. Scientists and researchers have solved many problems with more robust data analysis methods. In inspiration, we took the overarching goal to identify the best quality of wine for the best price by analyzing a large-scale wine data set—the research aimed to find common attributes of highly rated wines. The study was motivated by the subjectivity of wine-tasting reviews. Wine-tasting reviews are subjective because many individuals make them with different tastes. Our goal was to create an objective data analysis with the review data.

Kaggle provided the Wine Reviews database[1]. The popular dataset contained an analysis of approximately a hundred-thirty thousand wine reviews from WineEnthusiaist[2]. Each assessment includes the quality points, description, reviewer's name, winery, regional data, price, and other relevant data points. The data was created by a researcher scraping Wine Enthusiast's wine reviews to develop a predictive model to identify wines through blind tasting[1].

II. Methodology

The dataset we used contains a mix of natural language and numeric data, both of which were used to conduct our analysis. Some preprocessing was required, particularly with the numerical values, to properly move forward with our study. Wine prices were an integral part of our study, but some entries were missing price values. For this, we filled missing values with the dataset's price average. In addition to missing price values, some prices were extreme outliers which were removed before performing analyses. The last step in our data cleaning process was to condense duplicate entries. This step only applied to our numerical analyses. To condense the entries, we simply averaged each numerical value into one data entry and removed the extraneous ones.

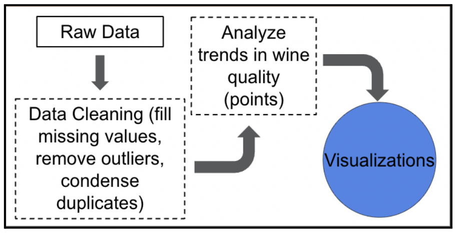

Raw Data
Data Cleaning (fill missing values, remove outliers, condense duplicates)
Analyze trends in wine quality (points)
Visualizations

Fig. 1. Workflow Diagram

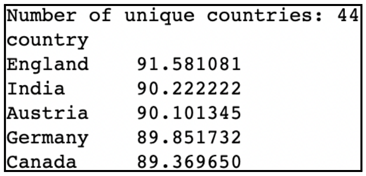
After the dataset was cleaned, we analyzed three variables to find trends in wine quality (points): word frequency in wine descriptions, country of origin, and price values.

A. Description of Wine

The research group compared the frequency of words in wine reviews and the average rating of the wine when a particular word appeared. In addition, we identified the forty most common words as verbs, nouns, and adjectives to identify a common feature among wine descriptions with higher quality ratings. To achieve such, we preprocessed and read the data. Afterward, we mapped each word as a key from the description and reduced the key values with the frequency of each term. Then we classified each word as nouns, verbs, and adjectives utilizing nltk's wordnet library. Finally, we generated the mean and standard deviation of the quality points of wine of each word's usage.

B. Quality and Price Trends

Wine quality may trend with country of origin. When one thinks of excellent wine, the thought is usually associated with France and/or Italy. We put this to the test by averaging the points of every wine for each of the 44 countries represented in the dataset, creating a "score" for each country.



```
Number of unique countries: 44
country
England     91.581081
India       90.222222
Austria     90.101345
Germany     89.851732
Canada      89.369650
```

Fig. 4. Top 5 countries by average wine rating

Another common notion is that higher quality wines are significantly more expensive (even reaching $3300 in this dataset). To test this, we created a quality-to-price ratio. This ratio is essentially the wine's points per dollar value. We compared individual entries on a heatmap to search for trends. In addition to this, we also grouped wines ranging from "acceptable" to "very good" (80-90 points) and "excellent' to "classic" (90-100 points).[2] We then compared general statistics of each group as another way to find trends in quality based on price.

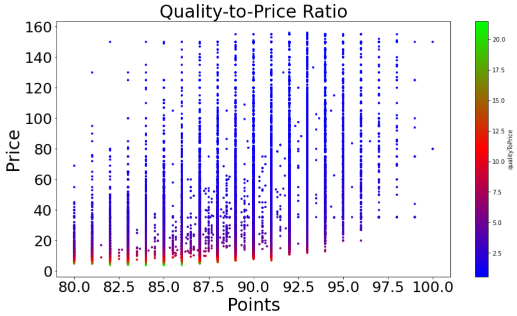


Fig. 3. Heatmap of each wine's quality to its price

III. Results

As a result, we found that the mean of all three groups was 88.9, and the standard deviation was 1.07. In addition, the mean of adjectives, verbs, and nouns were 89.0, 89.2, and 88.5, and the standard deviations of the three were 1.17, 0.78, and 1.14. Some common adjectives were ripe, fresh, and soft. In comparison, some common verbs were bodied, long, and balanced. Of the 44 countries represented, 15 had point values greater than the dataset average (88.4 points). Among those 15, only one country, England, had a point value greater than one standard deviation above the mean. Higher rated wines (90-100 points) are slightly more expensive in general than their lower rated (80-90 points) counterparts.

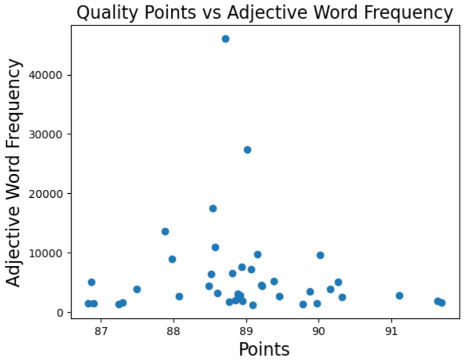


Fig. 4. Quality Points vs Adjective Frequency

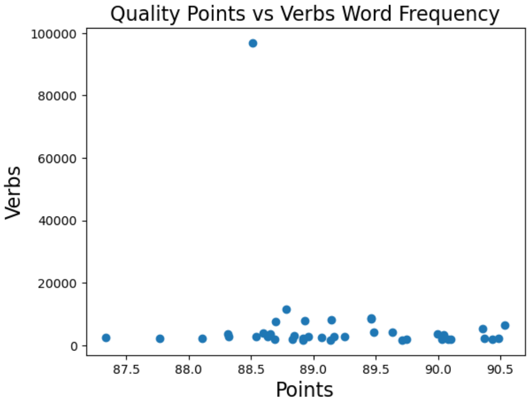


Fig. 5. Quality Points vs Verbs Word Frequency

IV. Conclusion And Future Work

Verbs had a slightly higher mean and a smaller standard deviation. However, the statistical analysis is insignificant as the values meet within one standard deviation. Therefore, it is possible to conclude that verbs in wine descriptions have a higher trend for wine quality but not a statistically significant conclusion.

During our analysis, we realized that the wine reviewers wrote many reviews in French and Italian. So, we could not correctly classify Italian and French words' frequency and speech form despite their high frequency. In the future, we would like to utilize a natural language model that accounts for multiple languages to increase the accuracy of our analysis.

Consumers will have the best chances at purchasing a high quality wine from the 15 above average countries, particularly England, India, and Austria as their average point score is considered "Excellent" on WineEnthusiat's rating scale[3]. Wines in the 90-100 point range are not significantly more expensive than the wines in the 80-90 point range. In fact, consumers wishing to get the best "bang for their buck" are better off buying some of the wines in the lower rating range as they have far more points per dollar.

V. References

[1] "Wine Reviews", kaggle.com, https://www.kaggle.com/datasets/zynicide/wine-reviews

[2] "Wine Enthusiast Magazine", https://www.winemag.com/

[3] "Wine Enthusiast", winesearcher.com, https://www.wine-searcher.com/critics-17-wine+enthusiast#:~:text=Wine%20Enthusiast's%20100%2Dpoint%20wine%2Dscoring%20scale%3A&text=90%E2%80%9393%20%E2%80%93%20Excellent,80%E2%80%9382%20%E2%80%93%20Acceptable