ENS 491/492 – Graduation Project

Final Report

Project Title: Wine Reviews

Group Members:

Beyza Burkay

Can Tartan

Hakan Körpe

Supervisor(s): Ezgi Karabulut Türkseven

Date: 04.06.2023

1. EXECUTIVE SUMMARY

This project aimed to analyze a wine reviews dataset to gain insights into wine quality factors and correlations between wine points and various attributes. The primary objectives were to explore regional patterns, examine the relationship between price and quality, and identify popular grape varieties. The project successfully accomplished these goals, providing valuable contributions to the wine industry's knowledge base. The dataset analysis revealed intriguing regional patterns in wine distribution across different countries and provinces. A notable correlation was found between sentiment score and points, suggesting that reviewer comments can serve as an indicator of wine quality. Average prices for each country were examined, along with the relationships between country, color, and points. These findings offer practical insights for professionals in the wine industry and contribute to a deeper understanding of the subject matter. Various methodologies and techniques were employed in the project, including feature selection and engineering, outlier removal using the Interquartile Range (IQR) method, sentiment score extraction through NLP Sentiment Analysis, and the utilization of supervised learning algorithms such as regression models, classification models, ensemble methods, and k-nearest neighbors. Model training and evaluation were performed, considering evaluation metrics like R-squared, mean squared error, and accuracy scores. In conclusion, this project addressed the problem of understanding wine quality factors and correlations with various attributes. The findings contribute meaningfully to the wine industry's existing knowledge, and future work can involve advanced analysis using neural networks and improved sentiment analysis techniques. The project's outcomes provide valuable insights and lay the foundation for further research in this field.

2. PROBLEM STATEMENT

The objective of this project is to investigate the determinants of wine quality and the factors influencing the points assigned to wines by tasters. We aim to explore potential correlations among the ratings provided by different commentators and identify causal factors within the selected datasets sourced from the Kaggle library (Zackthoutt, 2017). We have two separate datasets, one with 130 thousand entries and one with 150 thousand entries, with different numbers of columns. Our first objective is to combine these datasets before starting the project. Our main focus is to conduct an in-depth analysis of the dataset's columns and examine the relationship between these variables and the points awarded to wines. While previous analyses of the same dataset on the web primarily relied on basic visualizations such as point-price or point-variety bar graphs, our intention is to apply more advanced and persistent analytical methods. By utilizing these methods and tools, we aim to gain a deeper understanding of the dataset.

2.1. **Objectives/Tasks**

Objectives:

1. **Data Pre-processing:** Enhancing the dataset's quality and reliability in order to generate meaningful results.
2. **Exploratory Data Analysis (EDA):** Gaining valuable insights and a deeper understanding of the dataset through exploratory data analysis and visualization techniques.
3. **Feature Selection and Engineering:** Choosing and constructing features that can provide valuable insights and focus on the most influential aspects of the data
4. **Modeling Approach:** Selecting the most appropriate models for the dataset to get more accurate predictions.
5. **Model Training and Evaluation:** Training the selected models using the input features to achieve a good accuracy or R-score and trying new approaches after evaluating the predictions.

Tasks:

1. Data Pre-processing:

Merging the datasets: Merging the two datasets, one with 130 thousand entries and one with 150 thousand entries, into one dataset.

Removing Duplicates: Identifying and removing any duplicate entries, which will help minimize the potential for skewed results or biased interpretations.

Adding New Columns: Extracting new columns specifically designed for this purpose. As part of this process, we will create a 'color' and 'year' column to enhance data visualization. Also, eliminating outliers found in the 'price' column using the IQR method before the model training process, in order to maintain data integrity.

Remove Missing Data: Dropping any rows that contain missing values to align with our analysis objectives and maintain consistency. By implementing these measures, we aim to optimize the quality and reliability of our dataset, reducing perplexity and ensuring meaningful results.

2. Exploratory Data Analysis (EDA):

Descriptive Statistics: Calculating and interpreting statistical measures such as mean, median, mode, standard deviation, and range for numerical variables. This will help us unravel overall patterns, spread, and variability in the data.

Scatter Plots: Creating scatter plots to examine relationships between variables. Specifically, we will plot the relationship between price and points to determine if there is a direct correlation.

Quantity-Price and Quantity-Points Plots: Generating plots to visualize the distribution of quantity and price, as well as the distribution of points. These plots will provide insights into patterns and trends within the dataset.

Color and Density-Based Scatter Plots: Creating scatter plots to explore the relationship between average price and points, considering countries and states. This visualization will help us identify popular wine-producing regions and variations within the dataset.

Line Graph: Plotting a line graph to observe trends in average points over the years. This will enable us to identify specific years with exceptional wine quality.

Heat Maps: Generating heat maps to display the counts of wine varieties by geographical regions. This visualization will help us understand the distribution of wines across different regions.

Pivot Tables: Creating pivot tables to visualize the average points of wine colors by countries with high frequencies. This will provide insights into the popularity and scores of different wine colors.

Bar Chart: Generating a bar chart to categorize wine varieties by wine colors. This visualization will allow us to identify the most popular varieties for red and white wines.

Histograms: Creating histograms to visualize the distribution of sentiment scores in the dataset. This will help us understand the predominant sentiment in the comments.

Regression Plots: Generating regression plots to examine the relationship between average sentiment scores and points, as well as the relationship between average points and sentiment scores. These plots will provide insights into the impact of sentiment on wine quality assessment.

3. Feature Selection and Engineering:

New Dataframe for Regression Model: Creating a new binary data frame for regression models, including columns for top countries, continents, wine colors, top varieties, and the last 20 years of data.

Sentiment Analysis: Performing NLP Sentiment Analysis on the 'description' column to extract sentiment scores. By analyzing the text content, we aim to determine the sentiment expressed in each description. As part of this process, we intend to create a new column called 'sentiment_score,' which will range from -1 to 1, indicating negative to positive sentiment. This approach will provide us with valuable insights into the emotional tone of the descriptions and enable us to further analyze and understand the sentiment patterns within the dataset.

4. Modeling Approach:

Identifying the best-performing model among the alternatives, with a particular focus on classification models. By evaluating various models, we can assess their performance and determine the most effective one for our specific task.

5. Model Training and Evaluation:

Evaluating the accuracy and R-Score of the models by training the selected models using the input features that will be selected. By evaluating all possible outcomes, we will define our approach and select the model with highest accuracy.

2.2. **Realistic Constraints**

There are no realistic constraints regarding the project.

3. **METHODOLOGY**

1. **Data Preprocessing:**

Data preprocessing is a crucial step in any data analysis project as it involves cleaning and transforming the raw data to make it suitable for further analysis and modeling. In this project, several data preprocessing techniques were applied to ensure the quality and usability of the dataset. These steps include:

Merging the Datasets: The project involved merging two datasets with different numbers of columns (11 and 14). Dataset with 150k rows had 3 missing columns which were ‘taster_twitter_handle’, ‘taster_name’ and ‘title’, the rest of the columns were identical. To

combine the datasets, the columns were matched based on their relevance and common identifiers. By doing so, we created a consolidated dataset that only included shared columns, thus minimizing any potential inconsistencies or variations in our analysis.

Handling Missing Values: When conducting specific analyses, such as examining a particular column, we carefully considered the presence of missing values. In certain cases where our primary aim was to gain insights and comprehend the data, we proceeded with the analysis even if certain columns had missing values. For instance, when visualizing or analyzing the data, we prioritized the availability of key columns of interest, such as country, province, variety etc., placing less significance on missing values in other columns. However, for more comprehensive analyses or tasks requiring multiple columns, ensuring complete data became crucial. Consequently, we selectively removed rows that lacked values in the corresponding column(s) essential for the specific analysis. This approach ensured the reliability and consistency of our dataset, even if it meant excluding certain entries due to missing values in specific columns. The decision to retain or discard rows was determined by the nature of the analysis and the specific objectives we aimed to achieve.

Removing Duplicates: Removing duplicate entries was an essential step in guaranteeing the consistency and reliability of our data. By identifying and eliminating duplicate records, we minimized the risk of skewed results or biased interpretations.

Extraction of Year Column: A new 'year' column was created from the 'title' column of the dataset that contained information about the vintage year of the wines. However, it should be noted that this extraction was only possible for the dataset with 130k entries, as the 'title' column was not present in the 150k dataset. One of the issues we encountered was the presence of four-digit numbers in the 'title' column, which turned out to be winery names or designation names instead of years. This caused a disruption in the extracted years, leading to misleading information. To address this problem, we manually fixed it in Excel. This new 'year' column provided temporal information that could be relevant in analyzing trends and patterns over time.

Creation of Region and Province Codes: To further enhance the dataset, 'region_code' and 'province_code' columns were extracted from the existing dataset. These columns provided numerical representations of the geographical regions and provinces associated with each wine. This transformation allowed us to plot the frequency maps in python.

The decision to focus on these particular values and categories was driven by an understanding of their significance and impact within the dataset. Our analysis indicated that these selected categories held a substantial influence on the target variable, whereas other less

frequent categories proved to be less impactful and could potentially introduce noise into our models. Rather than transforming all data into categorical format, which could have increased the complexity and computational demands of our models, we opted to focus on these top-performing categories. This targeted approach allowed the models to capture the influence of these categorical variables more effectively, contributing to their predictive power.

By implementing these data preprocessing techniques, the project ensured that the dataset was cleaned, standardized, and enriched with relevant features. This set a solid foundation for the subsequent stages of exploratory data analysis, feature selection, and modeling.

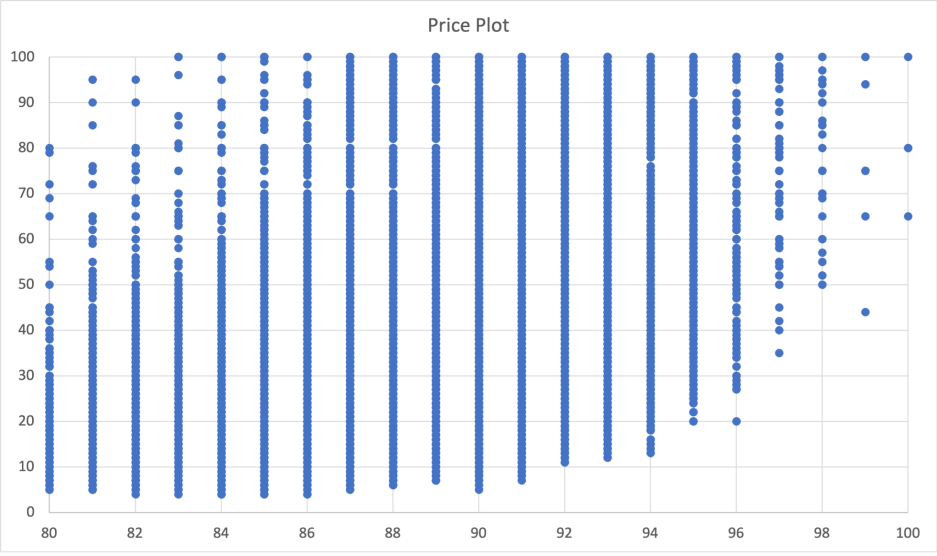
2. Exploratory Data Analysis (EDA):

During the exploratory data analysis phase, we employed various techniques to gain valuable insights and a deeper understanding of the dataset. We explored the data by utilizing descriptive statistics, data visualization, frequency analysis, and summary statistics.

Descriptive statistics played a pivotal role in revealing essential statistical measures such as the mean, median, mode, standard deviation, and range for the numerical variables in the dataset. By examining these measures, we were able to unravel the overall patterns, spread, and variability of the data. This allowed us to comprehend the central tendencies and grasp a sense of how the values were distributed.

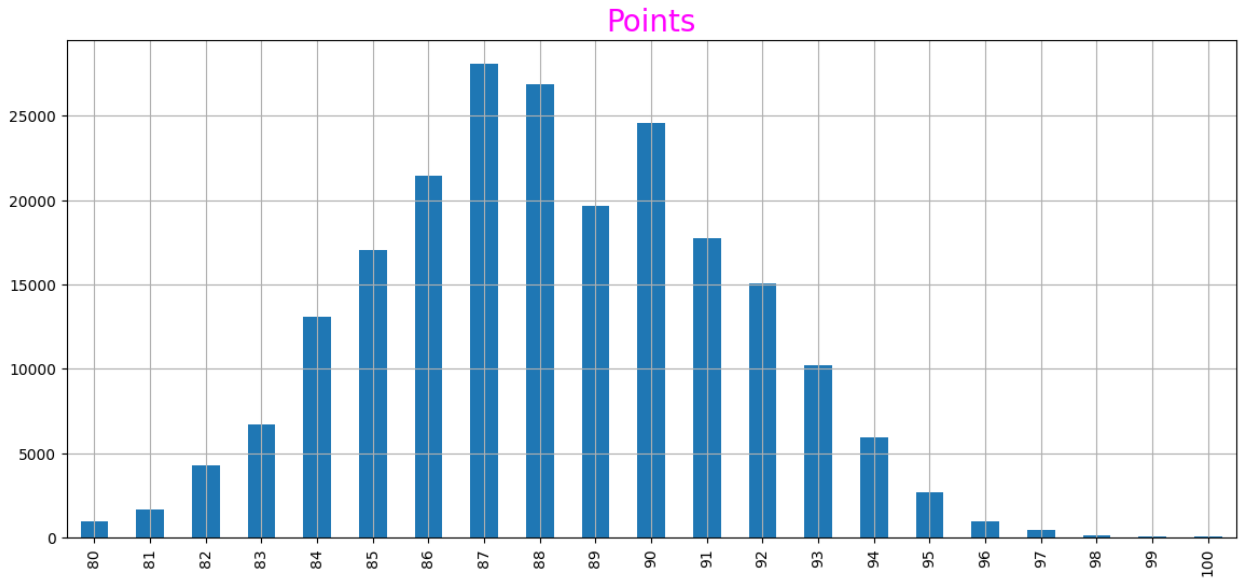
To reveal meaningful patterns, trends, and distributions, we harnessed the power of data visualization through a diverse range of techniques:

Scatter Plot of Prices: To examine the relationship between price and points, our initial step was to create a scatter plot showcasing the distribution of prices. Although the plot did not reveal a direct correlation between price and points, we acknowledge that in the context of this complex problem, the input feature of price still holds valuable information that can contribute to our predictive model.

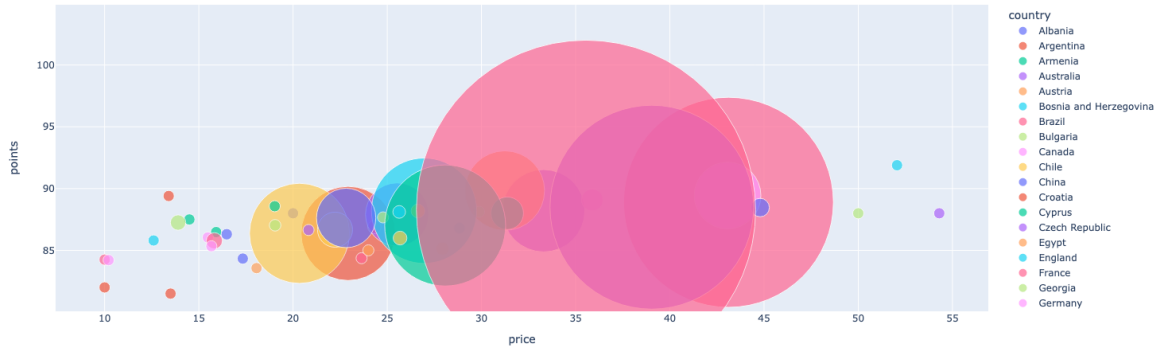


Quantity - Price Plot: This plot provides a visual summary of the quantity and price data, which can be used to identify patterns and trends. Upon observation, it is apparent that the plot exhibits a right-skewed distribution, indicating that the majority of prices fall within the range of 5 to 40 dollars. This information provides valuable insights into the average pricing patterns within the dataset.

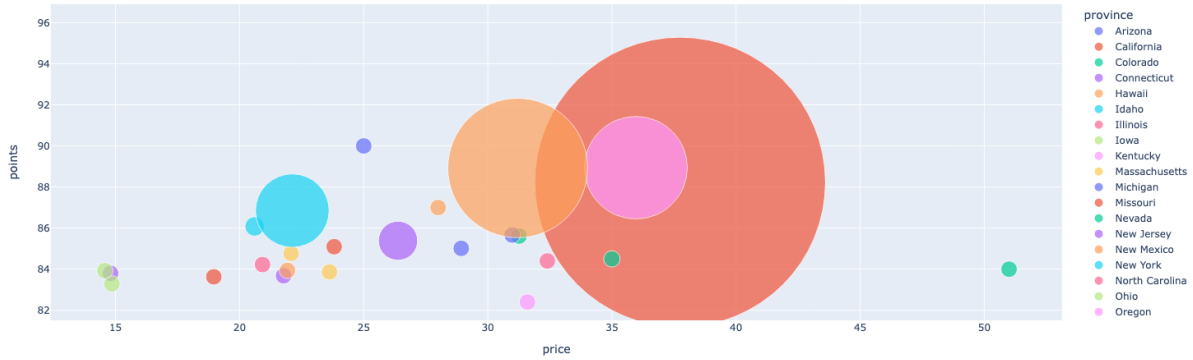
Quantity - Points Plot: This graphical representation provides a visual summary of the distribution of points across the different wines in our dataset. Upon examination, it is evident that the points exhibit a normal distribution, showcasing a bell-shaped curve. The average point value is observed to be 88, with a median of 87. This information allows us to gain insights into the overall distribution and central tendency of the points attributed to the wines, enabling a better understanding of their quality assessment.



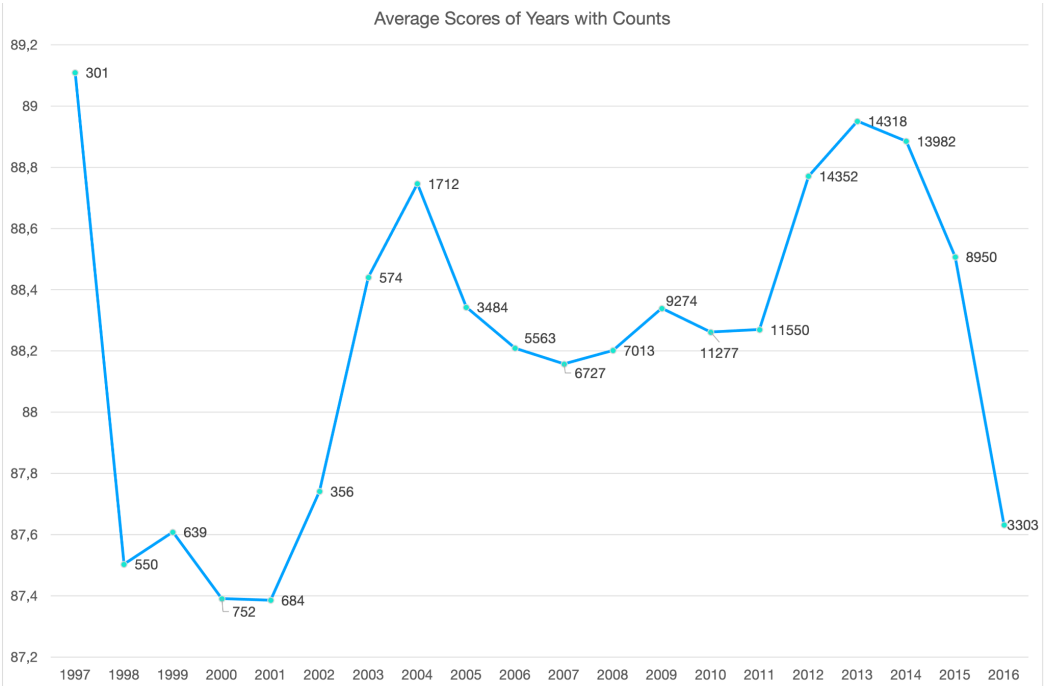
Color and Density-Based Scatter Plot of Average Price by Point for Countries: A scatter plot was generated to explore the relationship between the average price and points (rating) of wines, with each point colored and sized based on the country of origin. This visualization shows the most popular countries in descending order of prices as France, Italy, and America.



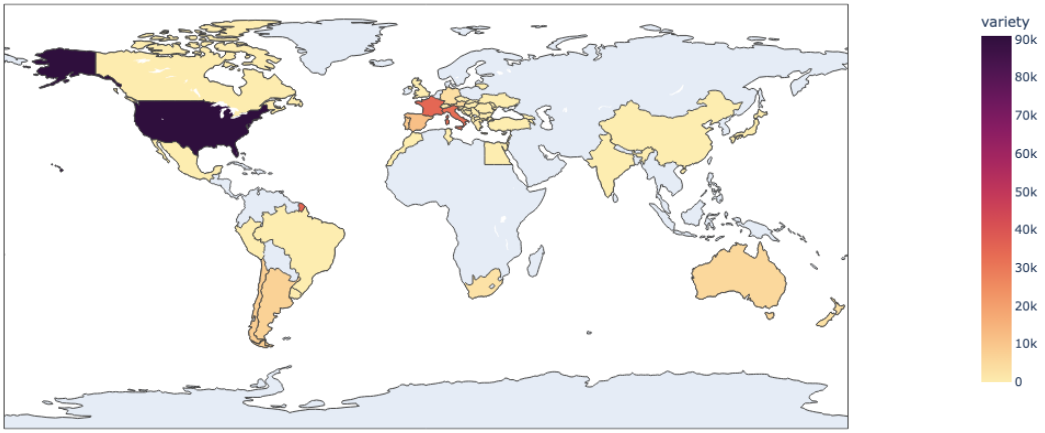
Color and Density-Based Scatter Plot of Average Price by Point for States in the US: Similar to the scatter plot for countries, a scatter plot was created to explore the relationship between the average price and points of wines produced in different states within the United States. This plot showed us most of the wines that are produced in the US come from California, Washington and Oregon respectively. While taking the count frequencies into account, California has the highest price among the states.



Line Graph of Average Points by Year: From the years we extracted in the data-preprocessing phase, we plotted a line graph and visualized the relationship between the years and their corresponding average points. This graph enables us to observe trends and patterns in the data. Upon examination, it is evident that certain years, namely 1997, 2004, 2012, 2013, and 2014, stand out with peak average scores. These years exhibit notably higher average points compared to others, indicating exceptional wine quality during those periods.



Heat-Based World Map of Variety Counts: A heat map was generated on a world map to display the counts of wine varieties by geographical regions. This visualization highlighted the dense distribution of wines in the US region and Western Europe.

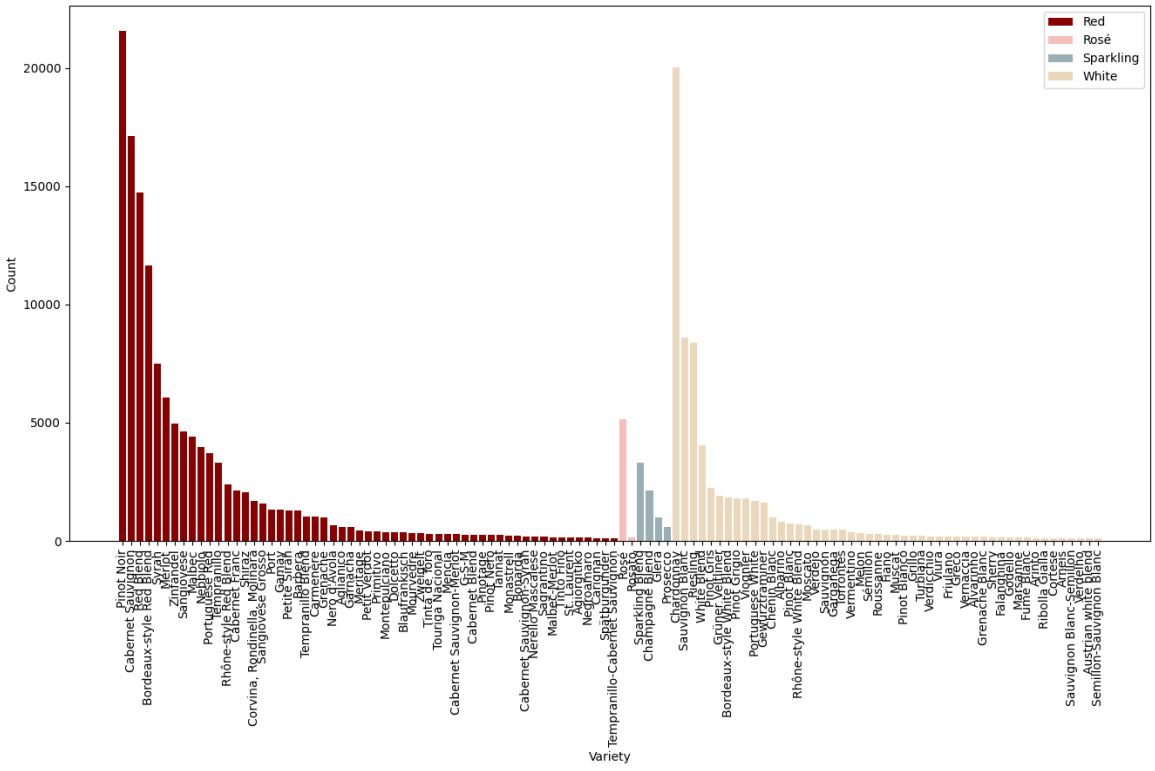


Pivot Tables of Average Points of Wine Colors by Countries : Pivot Tables were created to visualize the average points of wine colors by countries with high frequencies. In the pivot tables we can also see the varieties corresponding to the countries by selecting the plus (+) sign next to the country names. We can see that in red wines, France, Italy, Portugal, Australia and the US come to the fore in terms of score and popularity, respectively.

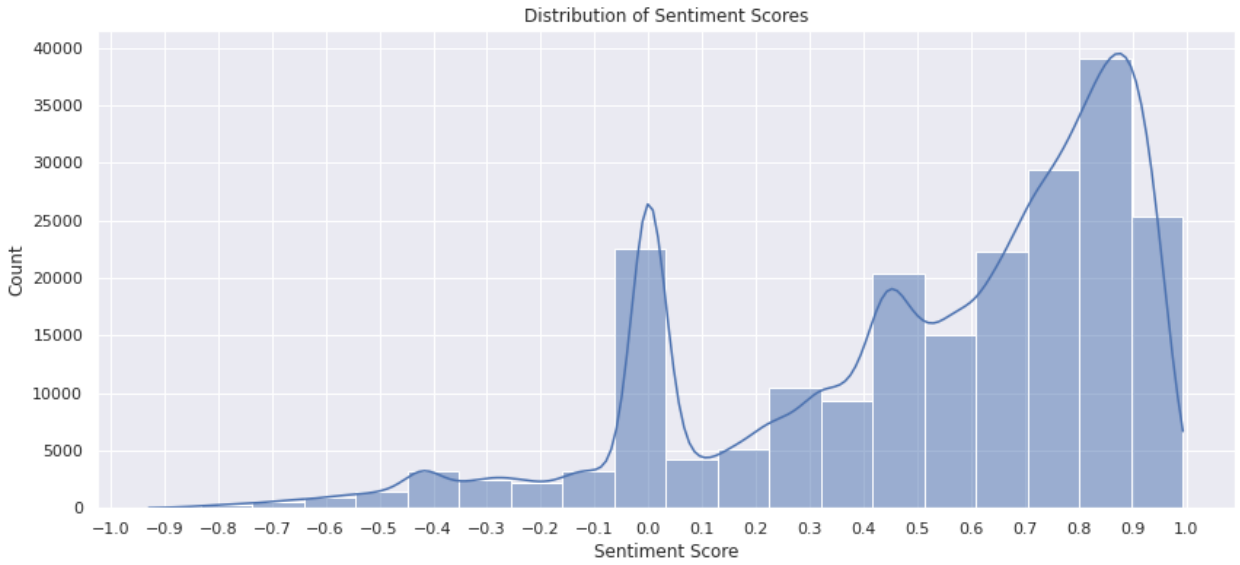
color	(All)	Red wines only!
Row Labels	Count of variety	Average of points
France	6849	89,18980873
Italy	9402	88,94575622
Portugal	2167	88,72404246
Australia	1470	88,6170068
Austria	263	88,57794677
US	27012	88,15004442
New Zealand	537	87,75977654
South Africa	554	87,35740072
Spain	3029	87,12677451
Argentina	2230	86,6206278
Chile	2187	86,56058528
Grand Total	55700	88,25795332

color	White	White wines only!
Row Labels	Count of variety	Average of points
Austria	1037	89,79074253
Germany	1305	89,16628352
France	4884	88,92444717
New Zealand	1022	87,61154599
US	10618	87,43360332
Italy	2337	87,36799315
Portugal	620	87,00483871
Australia	564	86,9822695
South Africa	519	86,8265896
Spain	749	86,435247
Chile	996	85,56526104
Argentina	503	84,56461233
Grand Total	25154	87,71698338

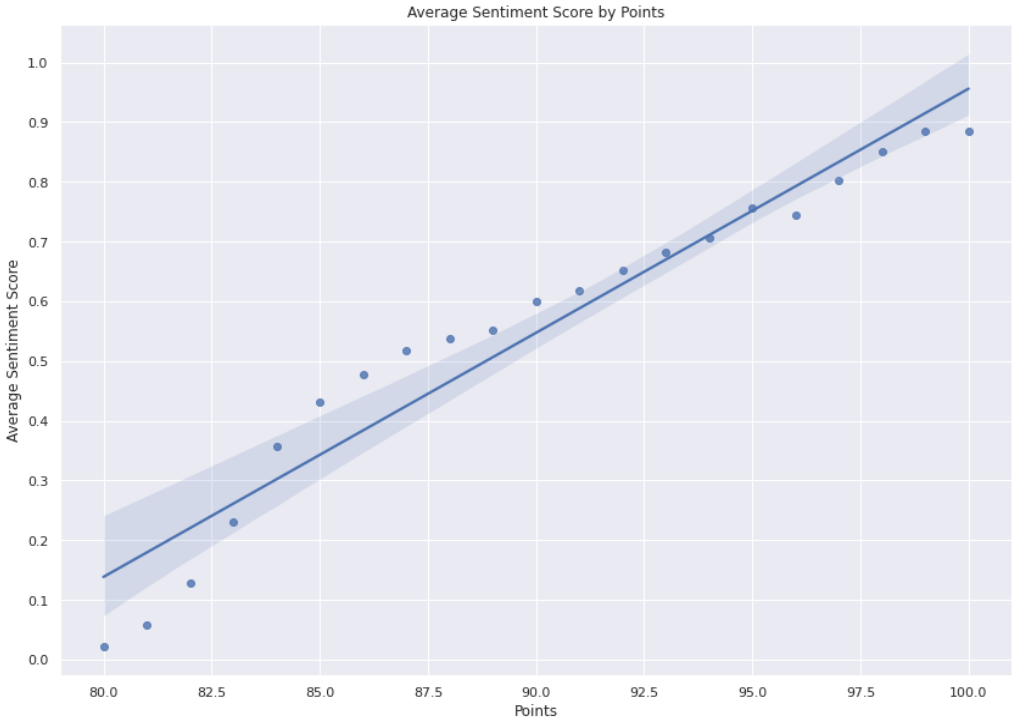
Wine Color-Based Bar Chart of Variety Counts by Varieties: A bar chart was generated to explore the counts of wine varieties categorized by wine colors. This visualization shows that for Red Wines, Pinot Noir is the most popular wine variety. Also for White Wines, Chardonnay is the most popular variety in the dataset.



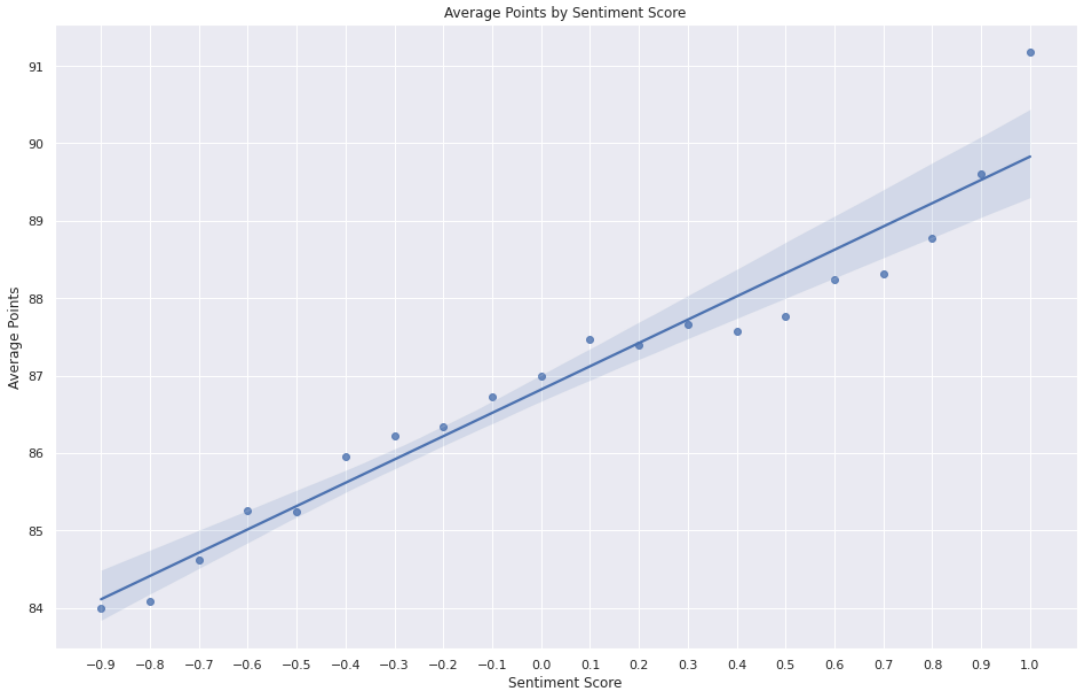
Histograms of Sentiment Score Distributions: Histograms were created to visualize the distribution of sentiment scores in the dataset. The resulting histograms depict a left-skewed distribution, indicating that the majority of comments possess positive sentiment. This conclusion can be drawn as the dataset's point range falls between 80 and 100 .



Regression Plot of Average Sentiment Score by Points: A regression plot was generated to explore the relationship between the average sentiment scores of wines and their corresponding points (ratings). Notably, we observed that at the extremes of the points scale, the variance in sentiment scores tends to increase. This indicates that wines with exceptionally high or low ratings can elicit a wider range of sentiment responses.



Regression Plot of Average Points by Sentiment Score: Another regression plot was created to examine the relationship between the average points of wines and their corresponding sentiment scores. We observed that as the sentiment scores increase, the variance in points also tends to rise. This indicates that wines with higher sentiment scores exhibit a wider range of point values. Also, this plot confirms that the sentiment scores impact positively on the predictive performance.



In conclusion, our exploratory data analysis and visualization techniques have provided valuable insights into various aspects of the wine dataset. The scatter plot of prices revealed no direct correlation with points, emphasizing the complexity of the problem at hand. The quantity-price plot demonstrated a right-skewed distribution, indicating average prices predominantly ranging from 5 to 40 dollars. The quantity-points plot showcased a normal distribution of points, with an average of 88 and a median of 87. Further analyses using color and density-based scatter plots explored the relationships between average price and points, considering countries and states. These visualizations highlighted France, Italy, and the US as prominent wine-producing regions, while also revealing California's dominance within the US. The line graph depicting average points by year uncovered exceptional wine quality in specific years, such as 1997, 2004, 2012, 2013, and 2014. A heat-based world map indicated the dense distribution of wines in the US and Western Europe. Pivot tables provided insights into the average points of wine colors by countries, emphasizing France, Italy, Portugal, Australia, and the US in terms of score and popularity for red wines. A wine color-based bar chart showcased

the popularity of Pinot Noir for red wines and Chardonnay for white wines. Histograms displayed a left-skewed sentiment score distribution, indicating a predominance of positive sentiment. Regression plots demonstrated that sentiment scores positively impacted predictive performance, with higher scores leading to greater point variance.

Overall, our analyses and visualizations have enhanced our understanding of the dataset, enabling us to uncover patterns, trends, and influential factors related to wine quality and sentiment. These insights will contribute to building an effective predictive model for this complex problem.

3. Feature Selection and Engineering:

To refine the dataset and enhance the relevance of our analysis, we undertook a very careful process of feature selection and engineering. Our goal was to carefully choose and construct features that would provide valuable insights and focus on the most influential aspects of the data. Through a thorough evaluation of the existing features, we identified the need for additional attributes to capture essential characteristics these approaches:

Outlier Removal: Through this process, we made informed decisions on how to treat outliers, ensuring that our subsequent analysis accurately reflected the underlying trends and patterns in the data. Prior to the model training process, outliers in the 'price' column were eliminated using the Interquartile Range (IQR) method. This step aimed to minimize the influence of extreme values on the models' performance.

Creation of Binary Data Frame: To enhance our dataset for regression modeling, we employed a strategic approach by creating a binary data frame that captured the most influential categories within the data. During the exploratory data analysis (EDA) phase, we identified the top 10 wine-producing countries. Consequently, we generated corresponding columns in the binary data frame, such as 'is Italy', 'is US', 'is France', and so on. Furthermore, we included columns like 'is Australian', 'is North American', and 'is European' to indicate the continents to which the countries belonged. Additionally, we leveraged the insights gained from data pre-processing, where we extracted information about wine colors from their varieties. Hence, we added columns such as 'is Red', 'is White', 'is Rosé', and 'is Sparkling' to capture these color categories. Lastly, taking into account the findings from the EDA phase regarding the top 10 varieties, we incorporated corresponding columns like 'is Pinot Noir', 'is Chardonnay', 'is Sangiovese', and so on.

Sentiment Score Extraction: Sentiment scores were extracted from the reviewers' comments using NLP Sentiment Analysis techniques. This process involved analyzing the sentiment of the comments and assigning a sentiment score ranging from -1 (negative) to 1

(positive). The sentiment score was added as a new column ('sentiment_score') in the dataset. This new feature provided a quantitative representation of the sentiment expressed in the reviews, which could be valuable in predicting wine scores.

During the evaluation of different sentiment score extraction techniques, we initially experimented with TextBlob, which uses a simple bag-of-words approach to calculate sentiment scores (Loria et al., 2014). However, the scores were heavily centralized around the neutral point (0), leading to a less diverse score distribution.

Alternatively, we utilized the NLTK VADER lexicon, which is more attuned to recognizing subtleties in sentiment expression by considering context, intensifiers, and contrastive conjunctions (Hutto & Gilbert, 2014). This method provided a more diverse spread of scores, giving a nuanced representation of the sentiments within the reviews. Given its superior performance, we ultimately decided to employ NLTK's VADER lexicon for the sentiment analysis in this project.

Selection of Key Input Features: In accordance with the project's objectives and analysis requirements, we carefully selected key input features to be used for model training. These features were chosen based on their significance in capturing crucial information related to wine scores while maintaining the integrity of the dataset.

The selected input features included 'country', 'province', ‘region1', 'variety', 'winery', ‘color’, 'price', and 'sentiment_score'. These features were considered essential as they involved various aspects such as geographical information, categorical attributes, and sentiment analysis related to the wines. By incorporating these features into the model, we aimed to ensure a comprehensive representation of the data, enabling us to effectively analyze and interpret the wine scores.

4. Modeling Approach:

The modeling approach involved applying various supervised learning algorithms to predict and analyze wine score based on the selected and engineered features. The following algorithms were utilized in this project:

Linear Regression: Linear regression models were used to establish a linear relationship between the input features and the target variables. This approach aimed to predict continuous variables, such as wine scores, based on the selected features (Neter et al., 1996).

Multiple Linear Regression: Multiple linear regression models extended the linear regression approach by incorporating multiple input features to predict the target variables. This technique allowed for capturing more complex relationships between the features and the target variables (Draper & Smith, 2014).

Ridge Regression: Ridge regression models employed L2 regularization to prevent overfitting and improve model performance. This technique was particularly useful when dealing with multicollinearity among the input features (Hoerl & Kennard, 1970).

Decision Tree Regression: Decision tree regression models utilizes a tree-like structure to make predictions based on hierarchical decision rules. This approach aims to predict continuous variables by dividing the feature space into regions based on the selected features (Breiman et al., 1984).

Decision Tree Classifier: Decision tree classifier models are employed to predict categorical variables, such as wine quality intervals. The decision tree algorithm makes predictions by partitioning the feature space and assigning classes based on the selected features (Quinlan, 1986).

Random Forest Classifier: Random forest classifier models combine multiple decision trees to improve prediction accuracy and reduce overfitting. This ensemble learning approach involves creating an ensemble of decision trees and making predictions based on the aggregated results (Breiman, 2001).

XGBoost: XGBoost, an optimized implementation of gradient boosting, is used to build gradient boosting models. This technique iteratively adds decision trees to the model and optimizes the loss function to improve prediction performance (Chen & Guestrin, 2016).

AdaBoost: AdaBoost, an ensemble learning method, is employed to combine multiple weak classifiers and create a strong classifier. This technique assigns weights to each classifier based on their performance and aggregates their predictions to make final predictions (Freund & Schapire, 1997).

k-Nearest Neighbors (kNN): kNN models utilize the concept of similarity to classify data points. This algorithm assigns labels to new data points based on the class labels of the k-nearest neighbors in the training dataset (Cover & Hart, 1967).

These modeling approaches covered a wide range of supervised learning techniques, including linear regression, decision trees, ensemble methods, and nearest neighbors. The selection of these algorithms allowed for comprehensive analysis and prediction of wine scores based on the selected and engineered features.

5. **Model Training and Evaluation:**

The model evaluation and training phase involved assessing the performance of the selected models using appropriate evaluation metrics. Various evaluation approaches were utilized to gauge the models' accuracy, robustness, and generalization capabilities. The following steps were undertaken:

Data Split: The dataset was divided into training and testing sets to evaluate the models' performance. A common split ratio of 80:20 (80% for training, 20% for testing) was used for the initial model training and evaluation.

Model Training: The dataset was used to train supervised learning models, including linear regression, multiple linear regression, ridge regression, decision tree regression, decision tree classifier, random forest classifier, XGBoost, AdaBoost, and k-nearest neighbors. Each model was trained on the training dataset using the selected and engineered features.

Evaluation Metrics: To assess the models' performance, several evaluation metrics were employed based on the specific task at hand. For regression models, metrics such as R-squared (R^2) and mean squared error (MSE) were used to measure the goodness of fit and the predictive accuracy of the models. For classification models accuracy scores were employed to evaluate the models' classification performance.

Decision Tree Model Outputs: For the decision tree classifier, the points intervals were categorized into different quality levels (normal, good, high, very high, excellent). This categorization allowed for a more interpretable analysis of the model's outputs and the classification of wines into quality categories.

Cross-Validation: To assess the models' performance more robustly, k-fold cross-validation (specifically 10-fold cross-validation) was performed. This technique involved splitting the dataset into k subsets and training and evaluating the models k times, with each subset serving as both the training and testing set. Cross-validation provided a more reliable estimate of the models' performance and their ability to generalize to unseen data.

Iterative Model Training: Since the sentiment score column ('sentiment_score') was added to the dataset, the supervised learning models were trained and evaluated again. This step ensured that the models were trained on the updated dataset to capture the influence of sentiment on wine ratings.

In summary, model training and evaluation phase involved splitting the data, training various models, evaluating their performance using appropriate metrics, and comparing their accuracy. The inclusion of outlier removal, handling missing values, cross-validation, and iterative training allowed for robust and reliable assessment of the models' performance.

4. RESULTS & DISCUSSION

In this section, we present the results and discuss the performance of our regression and classification models. Our main output feature is "points." We initially trained the models without considering sentiment scores, and then we extracted sentiment scores from reviewer comments and trained the models again, which resulted in improved performance.

Regression Models:

We evaluated the performance of our regression models using different combinations of input features. However, the outcomes did not meet the project's expectations. The highest R-Score we obtained was 41% using the features Country, Variety, Price, and Year.

Input Features	Multiple Linear Regression	Regression With Sentiment Score
Country, Variety, Price, Year	0.353	0.41
Region, Variety, Price, Year	0.345	0.40
Country, Price, Year	0.344	0.40
Variety, Price, Year	0.338	0.40
Country, Variety, Price	0.323	0.40
Country, Color, Price	0.317	0.39
Country, Price	0.315	0.39
Variety, Price	0.307	0.39

Classification Models:

We also evaluated the performance of our classification models, both with and without sentiment scores. The highest accuracy achieved was 70% using the Decision Tree model.

Input Features	Decision Tree	Decision Tree with Sentiment Score	XGBoost	XGBoost with Sentiment Score
Country, Province, Variety, Winery, Price	0.656	0.700	0.576	0.582
Country, Province, Variety, Winery, Price, Color	0.655	0.702	0.573	0.582
Country, Color, Variety, Winery, Price	0.658	0.699	0.567	0.573
Country, Province, Variety, Winery, Price, Region1	0.600	0.695	0.571	0.578

Other classification models were also tested, but their accuracies were relatively lower compared to the Decision Tree and XGBoost models. Due to limited computational resources and time constraints, we could not train the Random Forest model extensively.

Input Features	Random Forest	AdaBoost	Adaboost with Sentiment Score	kNN	kNN with Sentiment Score
Country, Province, Variety, Winery, Price	0.57	0.54	0.54	0.52	0.49

Effect of Sentiment Scores:
We examined the impact of sentiment scores on model accuracy. We split the sentiment scores into two categories: low sentiment scores (scores below 0.02) and high sentiment scores (scores above 0.02). We trained separate models for each category and evaluated their accuracies. Surprisingly, the sentiment scores with 0 values actually improved the model's accuracy since the output feature, points, ranged from 80-100.

Input Features	Decision Tree with Sentiment Score
Train/Test Split Accuracy for low sentiment scores	0.652
10-fold Accuracy with only low sentiment scores	0.686
Train/Test Split Accuracy for high sentiment scores	0.668
10-fold Accuracy with only high sentiment scores	0.689

Final Evaluation:
Finally, we present the accuracy of the Decision Tree model using 10-fold cross-validation. This approach splits the data into 10 parts, trains each combination of 1:9 train/test splits, and calculates the mean accuracy across the 10 combinations, providing a more accurate and unbiased prediction result.

Input Features	Decision Tree with Sentiment Score
Country, Province, Variety, Winery, Price, Color	0.732
Country, Province, Variety, Winery, Price	0.730

In conclusion, our models showed varying levels of performance with different input features. Despite the inclusion of sentiment scores, the overall accuracy did not meet the project's desired outcome of at least 80%. The highest accuracy achieved by the classification models using these features was 73% with the Decision Tree model. While this accuracy is relatively higher compared to other models and feature combinations, it does not guarantee that these features alone can accurately predict the quality of a wine. However, the findings provide insights into the effectiveness of different input features and the impact of sentiment scores on model accuracy.

5. **IMPACT**

This project has an impact on both consumers and producers, as the data analysis conducted helps to uncover the factors that contribute to favorable wine reviews, ultimately leading to a better understanding of what makes a wine exceptional. By thoroughly examining this dataset, we have not only provided valuable insights for current researchers but have also paved the way for future studies seeking to comprehend the qualities of wine based on its characteristics using robust data analysis methods. This analysis enables wine buyers to make informed decisions by selecting highly rated yet affordable wines. Furthermore, the application of data visualization methods demonstrated in this report extends beyond wine reviews and can be advantageous for the wine producers. This will enable the wine producers to harness consumer preferences and produce wines that align with market demand.

6. **ETHICAL ISSUES**

When it comes to the analysis of the wine reviews, there are no ethical issues.

7. **PROJECT MANAGEMENT**

The implementation part had difficulties arising from the sheer amount of data the database had and the data pre-processing actually took an unexpected amount of time and effort. We were also unable to achieve any type of success with the neural network because of the amount of time it required to run the code as well as the sheer size and complexity of the code being unable to run it on our personal computers.

8. **CONCLUSION AND FUTURE WORK**

In conclusion, this project has been successful in achieving its primary objectives of understanding the factors that contribute to wine quality and the correlations between wine points and various attributes. Our findings regarding regional patterns, the relationship between price and quality, and the impact of variables such as country and color on wine scores provide meaningful contributions to the existing knowledge base. The project had certain limitations due to limited computation power and time constraints. Other constraints were the lack of the datasets integration to one and another because one of the datasets didn't have all of the columns that the other had such as titles and had many missing values in the price column. Another important setback with the data was the reviews only had points above 80 from 100 thus making the dataset predominantly positive. These limitations impacted the scope and depth of the analysis, as well as the ability to implement certain models and algorithms. The overall findings provide useful insights into the effectiveness of different input features of wines. Future work can include the further analysis of the data with even stronger tools like

neural networks. The project can benefit from this further analysis with the deepening of the understanding of the dataset by creating an even higher standard of sentiment analysis and improved artificial intelligence training methods. The future work may include re-doing the sentiment analysis with a more advanced “bot” specifically designed to analyze human written texts.

9. APPENDIX

https://github.com/hakankorpe/ENS492-Wine_Reviews

10. REFERENCES

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. CRC press.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

Draper, N. R., & Smith, H. (2014). Applied regression analysis. John Wiley & Sons.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

Loria, S., Keen, P., Honnibal, M., & Yankovsky, R. (2014). TextBlob: simplified text processing. Secondary TextBlob: Simplified Text Processing.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models (Vol. 4). Chicago: Irwin.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

Zackthoutt. (2017, November 27). Wine Reviews. Kaggle. Retrieved April 2023, from https://www.kaggle.com/datasets/zynicide/wine-reviews