Wine Quality and Price Analysis

IEOR E4523 Data Analytics

Professor Uday Menon

Healthy Pandas
Afnan Khan (ak4854), Devangi Gaikwad (dsg2160), Laurel Hickey (lh3161),
Shruti Agarwal (sa4136), and Yanyi Wang (yw3949)

"Wine can be healthy. Let us help you pick one."

1. *Introduction*

Red wine has been part of social, religious, and cultural events for centuries. In the past, people have theorized that red wine benefits health, particularly alongside a balanced diet.[1] A 2018 study notes that drinking red wine in moderation has positive links to cardiovascular disease, atherosclerosis, hypertension, certain types of cancer, and type 2 diabetes.[2] These links to human health were the primary motivation in looking at a wine dataset as initially the goal of the group for this project was to look at something that could affect health, though the dataset chosen has all types of wine not just red wine and ultimately, the quality of wine has not been linked to more health benefits. However, the objective of this analysis is to gain a better understanding of the quality score of wine and to produce a mechanism that will better predict the quality of a bottle of wine based on a review description, price, and geographic information. In gauging the dataset, it will be easier to identify wines with higher quality and generate a recommendation system.

For the purpose of our analysis, the quality of wine is explored using machine learning and text-mining techniques. The specific dataset used for this analysis was the "winemag-data-130k-v2" dataset from Kaggle[3] that has the rating of wine on a scale of 1 to 100, however, most of the ratings are between 80 and 100 as seen in Figure 1.1. It contains text descriptions of each wine, from different individuals on Twitter. The combination of both textual and numerical data enabled the group to perform text mining, including Vader sentiment analysis and topic analysis, and machine learning using regression, classification and KNN algorithm to get more insights into the dataset as a whole.

2. *Data Processing and Exploratory Data Analysis*

While processing the data, we chose to keep the columns *country, description, points, price, province,* and *variety*. The columns that were dropped included *twitter_handle, vineyard, region_1, region_2,* and *taster_name*. The columns twitter_handle, vineyard, and taster_name had too many unique values in each different column to the point where they were more of a distinct indicator for each different wine and hence the decision to drop these columns. In addition, the columns region_1 and region_2 also had a lot of unique values with many missing columns. Through analysis of these columns, it was found that many of the regions were just a repetition of the province column, therefore, dropping these columns did not cause loss of significant information.

When looking at the price variable and the distribution, it was found that the majority of the data was below $100. In addition, this range makes sense for a normal range of relatively affordable prices and this makes this analysis more accessible to the general public by excluding the wines with extremely high prices (Figure 2.1).

For low-priced wines, the ratings are strongly grouped around points between 85 and 88. As prices increase, points follow accordingly. However, the highest prices are not of the wine with the highest ratings (Figure 2.2).

An interesting observation was made for the relationship between ratings and descriptions' lengths. We noticed that the wines with the highest points also had the longest description length (Figure 2.3).

Lastly, if we specifically focus on the top 5 most common wines (considering variety), Pinot Noir is the variety that has the highest mean points, whereas Red Blend is the lowest (Figure 2.4). Each distribution is exhibited in detail with the stacked plot between points and variety (Figure 2.5).

3. Text Mining:

The text mining section of the project relied on the reviews by customers and wine enthusiasts, proving to be a crucial indicator in the classification model. To further understand more about the wine as the other columns in the dataset were solely focused on the type of wine and geographic location. Beyond this, the only indicator that is not within these two categories is price, therefore, in order to create new columns in the dataset text mining was done in the form of sentiment analysis and topic analysis.

3.1. Sentimental Analysis

When performing sentiment analysis, the vaderSentiment package in python was used to perform this analysis. This package was used on the description column of the data set in order to extract the positive, negative, neutral, and compound scores. We specifically chose to use VADER sentiment analysis as it is not a bag of words sentiment analysis and it takes into account the sentiment of each sentence as a whole. These scores were then added to the dataset in order to be used in further analysis.

3.2. Topic Analysis

3.2.a. WordClouds

WordClouds were created for the top 10 countries (Figure 3.2.1) with the highest number of reviews using the WordCloud library and eliminating common words (wine, review, drink, etc.). The most frequent 400 words were masked on an image with each country's map (Figure 3.2.3-Figure 3.2.12). Base map images were obtained from reference code found on Kaggle.[4] The predominant words found in reviews from a specific country provided insights into the choice of flavors, palette descriptions and posed to be initial recommendations for the region. The visualization of WordClouds using maps was an intriguing catch for people not familiar with machine learning and its interpretation. People's taste and description of the wine were deemed to be influenced by social and geographical factors.

3.2.b. Topic Modeling

As we noticed earlier, many wines can convey similar taste closely related to its variety or country of origin. As such, we wanted to approach all the descriptions with topic modeling. The results not only helped us attain better apprehension of how wines are usually characterized, but also matched new wines with a specific topic.

After tokenizing all descriptions, removing stop words (wine, drink), and lemmatizing them, we applied a Latent Dirichlet Allocation model (LDA). We tried numbers of topics ranging from 3 to 15 and concluded that 5 topics were enough to avoid overlapping while giving a good representation of how wines could be characterized (Figure 3.2.13).

1	flavors, apple, citrus, white, acidity, lemon, palate, fresh
2	flavors, aromas, finish, berry, plum, palate, cherry, fruit
3	ripe, acidity, fruit, tannins, rich, fruits, red, ready
4	aromas, palate, black, offers, dried, tannins, red, opens
5	flavors, fruit, black, cherry, shows, oak, pinot, rich

4. Machine Learning Models and Comparison

Since our data is mostly categorical data, we used one-hot coding to convert categorical data to new binary features with 1 and 0 for each unique category in categorical data. Initially when creating the models, a regression analysis was performed in order to attempt to predict the quality of wine. The models that were created were Logistic Regression with accuracy of 16%, Decision Tree regressor with a R^2 score of 48%, Random Forest Classifier with accuracy of 16%, Bagging Classifier with accuracy of 16%, and KNN Regressor with R^2 score of 34% (Figure 4.1). Due to all of these models having very low accuracy, in order to make a better prediction, the quality column was changed to a categorical data type with values 1-9. According to the EDA we have done previously, points of wine are normally distributed. Therefore we set the range for each quality point 1-9, from 10th percentile to 90th percentile respectively. This change of data type to categorical made the model more intelligible and did not result in any significant loss of information. For an individual using the model, it is more comprehensive to observe the difference between a category for the higher quality versus a lower quality than understanding the difference in points on a continuous scale.

After we mapped continuous point data to categorical data, we employed classification models for prediction. Since classification models were designed to predict categorical data, and there were 9 classes to cluster, the accuracy of prediction was expected to increase. After we performed a Random Forest Classifier, we obtained a 95% accuracy and a 95% f_1 score, 86% accuracy score and 81% f_1 score for Decision Tree Classifier, 89% accuracy and 89% f_1 score for KNN Classifier.(Figure 4.2)

5. Recommendation system

Recommendation systems are becoming increasingly important in today's hectic world. People are always in the lookout for products/services that are best suited for them. Therefore, the recommendation systems are important as they help them make the right choices given their usual preferences.

Recommendation systems can be broadly divided into two categories : content based and collaborative filtering based. Content based recommender systems focus on the properties of the content to recommend items to users, and this is what we are going to focus on. More precisely, given that a user likes a certain wine, we are interested in suggesting similar items they may like in the future.

For that purpose, the Nearest Neighbors algorithm was used, an unsupervised learner for implementing neighbor searches. We chose *province, variety* and *points* features and *cosine* as the metric for distance computation. Similarity was found to be the cosine of the angle between the two item vectors for A and B. Closer the vectors, smaller was the angle and larger was the cosine.

To choose the number of neighbors we looked at how many varieties had more than 1500 appearances in the dataset. 15 varieties fell under this group. With that, we were able to produce a recommendation system where, given a preferred wine, the model would return a list of similar wines the user may like to try. For instance, for a Raboso wine, recommendations are Friulano, Manzoni, Corvina, Tocai and Marzemino.

6. Conclusion and Further Steps

By using text mining and turning the quality variables into a categorical data type, we obtained a machine learning model with a high accuracy rate. Modifying quality into a categorical data type made it more coherent for the everyday wine drinker. In addition, to make the model more convenient for an everyday individual to use, the geographic information could be omitted as most casual wine drinkers are not very concerned with where a wine came from geographically but more with the type of wine, which is red or white.

A further step in order to make the analysis more accessible would be to try to find additional data similar to test our model to see how accurate it is to predict when the source of the data changes. In the analysis, topic analysis was done to add an additional feature to the data, however, ultimately this was not included in any of the models that we made. Another step in this analysis could be to create a more user interactive model and provide more options to users based on flavor or fruit categories. Similarly, one could consider using descriptions and their main topics for improving the recommendation system.

7. References:

[1] *10 health benefits of drinking red wine that will keep you healthy*. Whitehall Lane Winery. (2017, June 3). Retrieved December 17, 2022, from https://whitehalllane.com/10-health-benefits-of-drinking-red-wine-that-will-keep-you-healthy/

[2] Snopek, L., Mlcek, J., Sochorova, L., Baron, M., Hlavacova, I., Jurikova, T., Kizek, R., Sedlackova, E., & Sochor, J. (2018). Contribution of Red Wine Consumption to Human Health Protection. *Molecules (Basel, Switzerland)*, *23*(7), 1684. https://doi.org/10.3390/molecules23071684

[3]Zackthoutt (2017) *Wine reviews*, *Kaggle*. Available at: https://www.kaggle.com/datasets/zynicide/wine-reviews (Accessed: November 7, 2022).

[4] Skrzym, M. (2017) *Wine review word clouds*, *Kaggle*. Kaggle. Available at: https://www.kaggle.com/code/skrzym/wine-review-word-clouds (Accessed: November 19, 2022)

8. Appendix:

Figure 1.1:

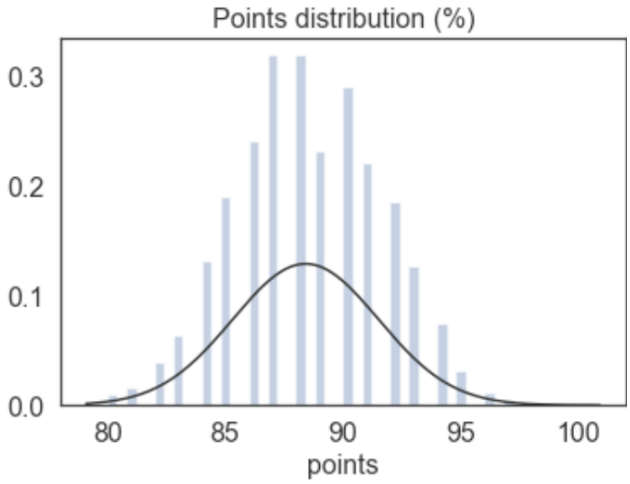


Figure 2.1:

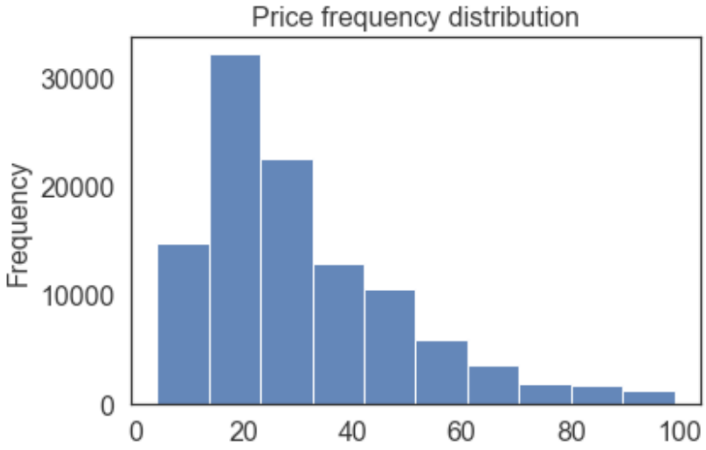


Figure 2.2:

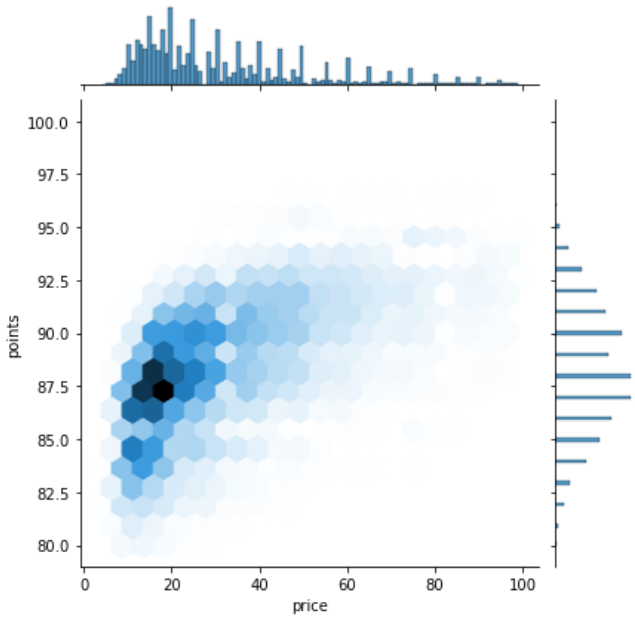


Figure 2.3:

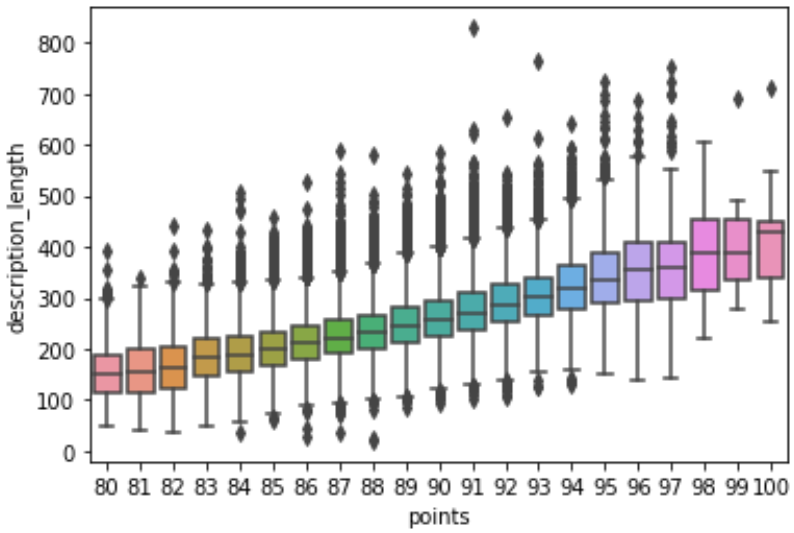


Figure 2.4:

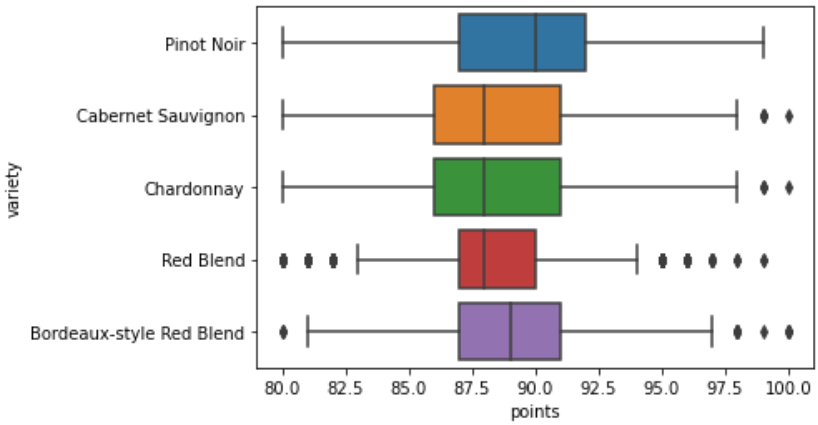
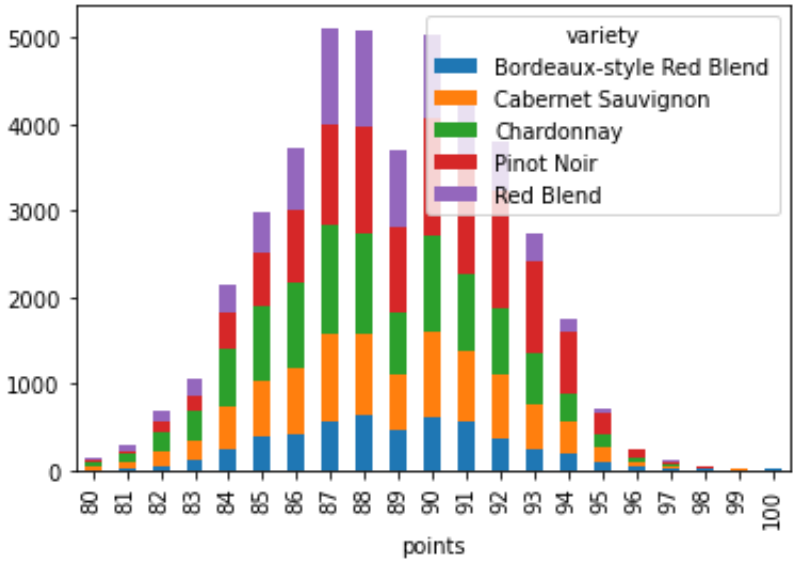


Figure 2.5:



Text Mining & WordClouds

Figure 3.2.1:

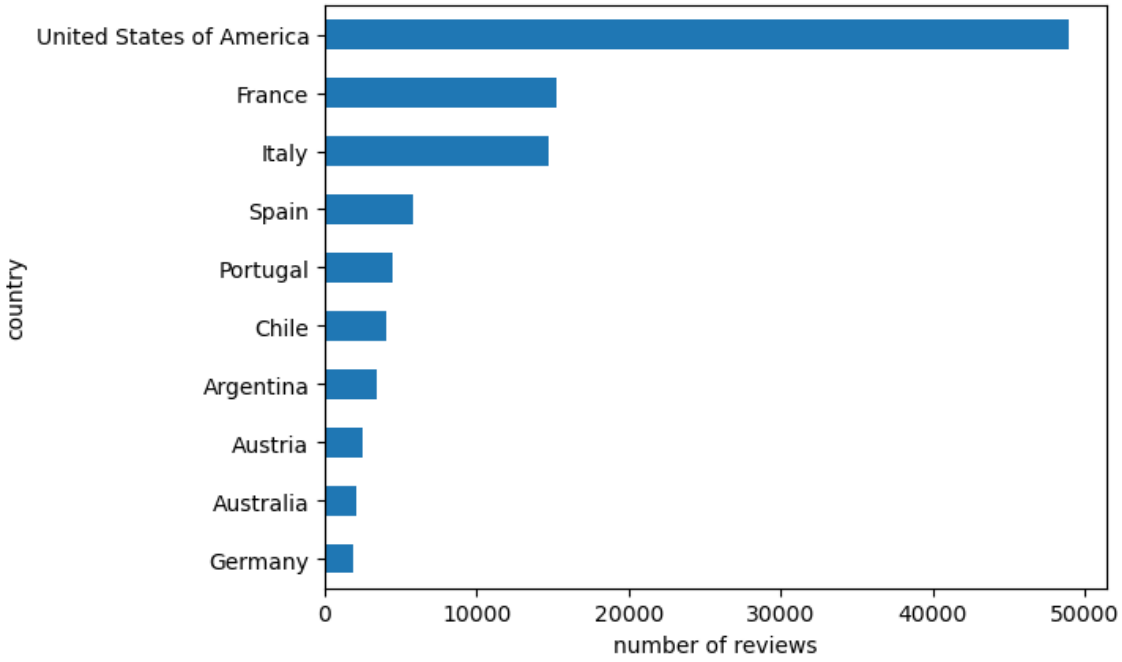


Figure 3.2.2:

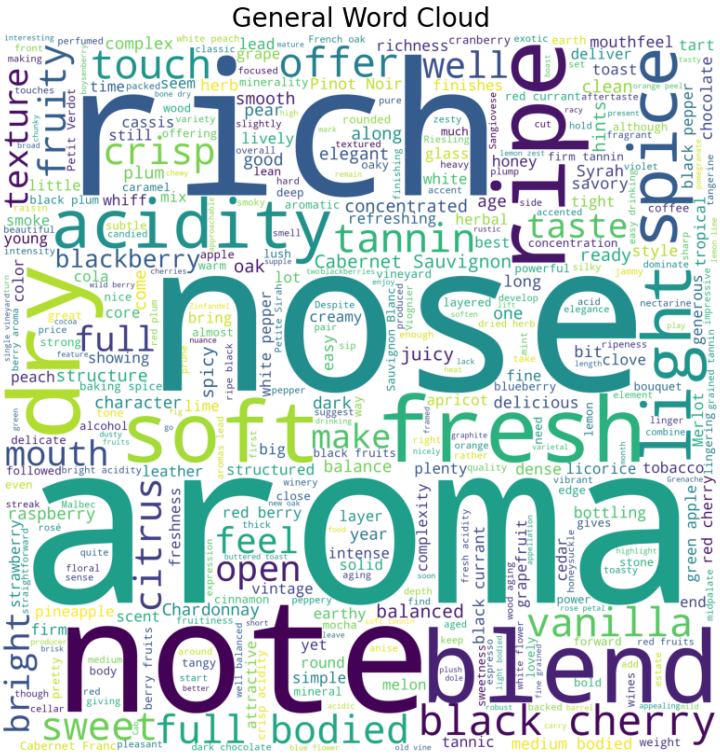


Figure 3.2.3: United States of America

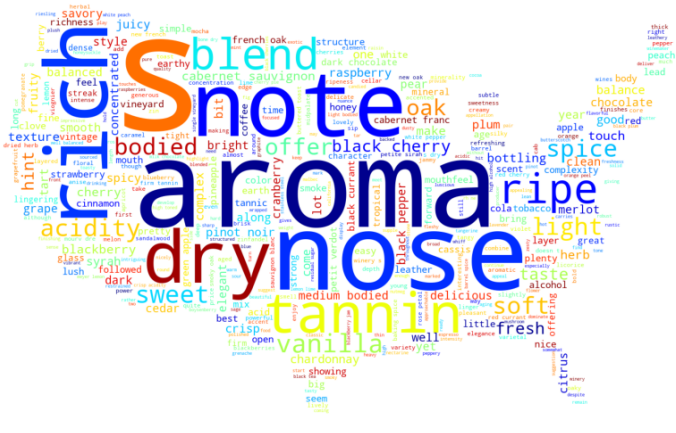


Figure 3.2.4: France

Figure 3.2.5: Italy

Figure 3.2.6: Portugal

Figure 3.2.7: Spain

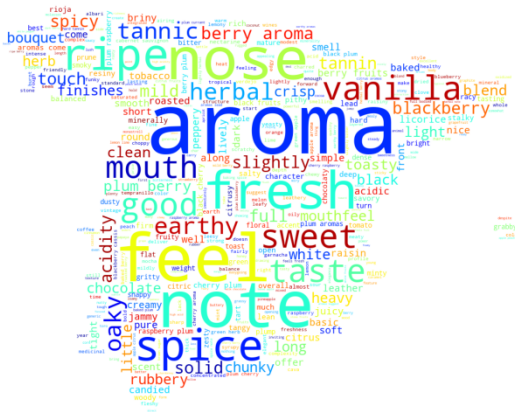


Figure 3.2.8: Austria

Figure 3.2.9: Chile

Figure 3.2.10: Argentina

Figure 3.2.11: Australia

Figure 3.2.12: Germany

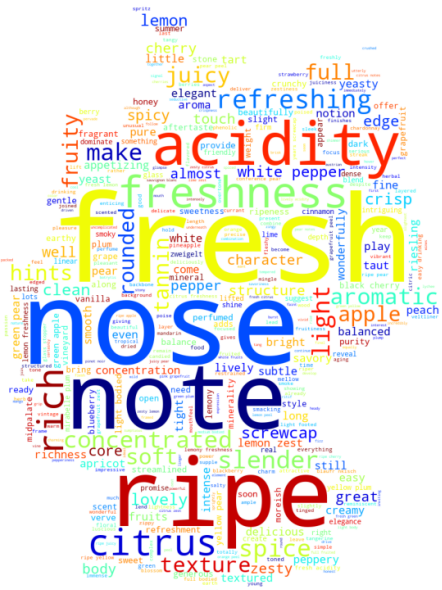


Figure 3.2.13:

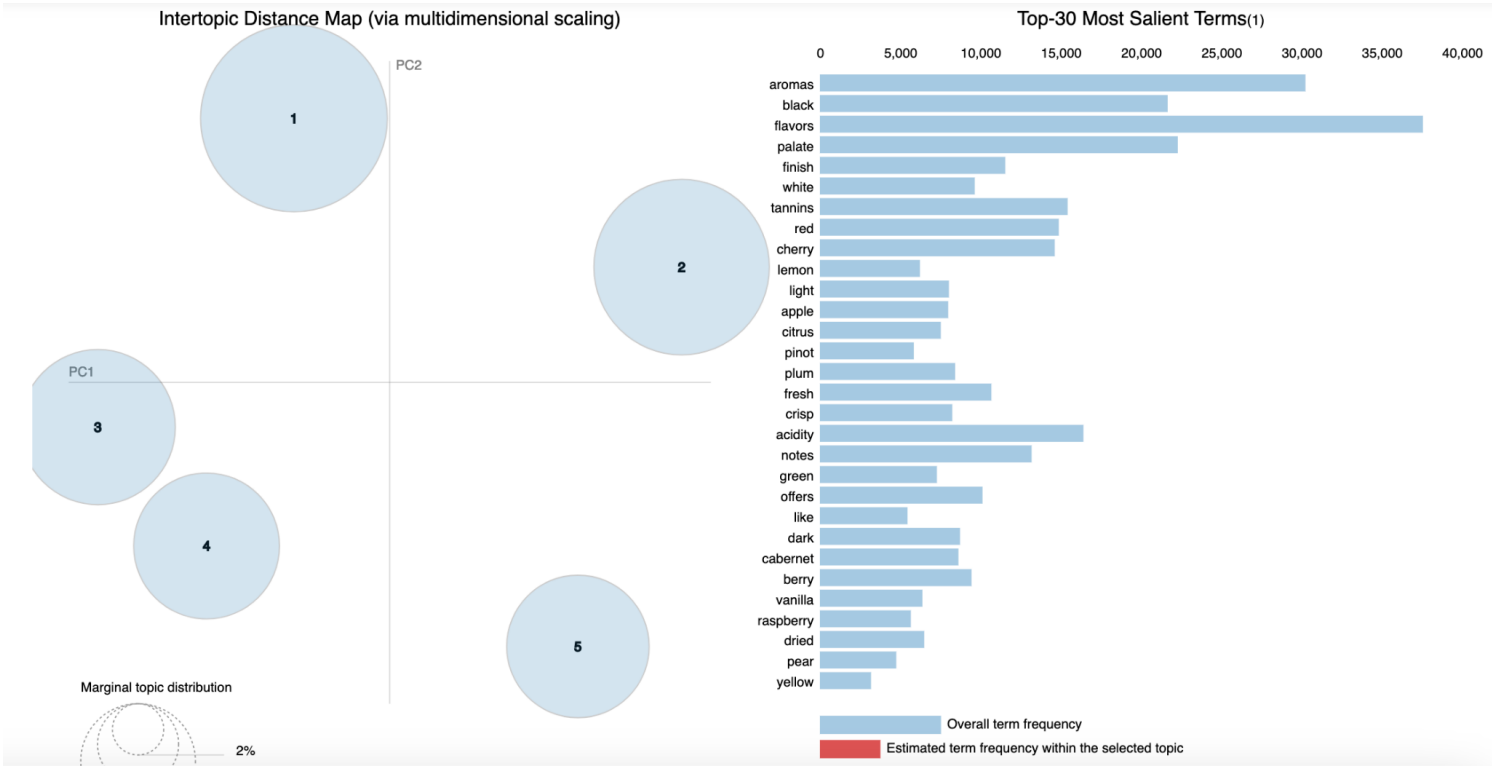


Figure 4.1

Model	Accuracy	RMSE	R^2
Logistic Regression	16.019%	2.5933	27.5%
Random Forest Classifier	15.822%	2.7649	20.2%
Bagging Classifier	16.044%	2.7182	22.9%
KNN Regressor	-	2.5139	34.1%
Decision Tree Regressor	-	2.217	48.2%

Figure 4.2

Model	Accuracy	Precision	Recall	f_1 score
Logistic Regression	27%	19%	28%	20%
Random Forest Classifier	95%	96%	96%	95%
Decision Tree Classifier	86%	78%	87%	81%
KNN Classifier	89%	90%	89%	89%