Wine Production Quality and Satisfaction: Unsupervised Learning Methods

CSCI 5622: Machine Learning

Author: Samuel Kwon

Department of Computer Science
University of Colorado Boulder
10/02/22

Introduction

The varieties of red and white wine are increasing every year. The increased production and sales of wine make it harder for consumers to purchase good quality wine. As a result, this project aims to explore wine datasets and consumer satisfaction through unsupervised machine learning methods. This paper will apply clustering methods such as K Means, Density, and Hierarchical clustering to the Portugal Wine and Chemical Analysis dataset. In addition, Association Rule Mining will be applied to the NewsAPI data gathered to examine associations to wine from August through September.

Clustering is an unsupervised learning method that utilizes unlabeled data. It seeks to group data points into different clusters depending on the distance. This is important in datasets with no labels as it can highlight the variables and characteristics of the dataset. Furthermore, clustering of the dataset can give insight into any associations between cluster groups. In this paper, K Means, Density, and Hierarchical clustering will be applied to the Portugal Wine and Chemical Analysis dataset to compare the clustering outcomes of these different methods. Since the Chemical Analysis dataset is unlabeled, it is a perfect candidate for clustering. On the other hand, the Portugal Wine dataset contains labels that are represented as discrete variables. Performing clustering on this set will determine if different levels of chemical compounds impact the quality rating.

Association Rule Mining is another unsupervised learning technique that seeks to determine associations in the dataset. Associations, also called rules, can be determined with numerical or text data. In this paper, Association Rule Mining will be performed using text data gained from NewsAPI. Performing text data analysis using ARM will give insight into words associated with wine in news outlets. This will ultimately represent current events on wines and the consumers' attitude towards wine during the time the data was extracted.

Analyses

Analyses: Clustering

Data clustering includes three different methods: K Means Clustering, Density-Based Spatial Clustering, and Hierarchical Clustering. Clustering methods will be performed on the Portugal Wine and Chemical Analysis datasets. The Portugal Wine dataset includes 11 variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, sulfates, alcohol, and quality. The quality variable is discrete, where consumers rate the wine from 1-10. After data cleaning, the mean of the quality variable was 5.631, maximum of 7 and a minimum of 4. In order to determine the clustering pattern of this dataset, the discrete variable was converted to labeled data. Wines with quality greater than or equal to 6 were determined as "good" wine, the quality between ranges of 6 and 4 was determined as "average" wine, and the quality of less than or equal to 4 was determined as "bad" wine. After converting the quality variable into labeled data, it was removed to perform Clustering. The Chemical Analysis dataset has 13 variables: Alcohol, Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280, and Proline. It is important to note that this dataset does not include labeled data; therefore, no additional steps were taken to prepare the dataset for Clustering.

K Means clustering, Density-Based Spatial Cluster, and Hierarchical Clustering are unsupervised methods that require unlabeled data; therefore, preparing the dataset before clustering is important. Clustering is used to discover groups or clusters in the dataset. The user can determine variables and vectors in the unlabeled dataset if clusters are identified. K Means clustering is a partitioning clustering method where clusters are associated with a centroid. In every iteration, clusters are assigned to the closest centroid, making up the number of clusters, K. The goal of this method is to minimize the sum of distances with respect to the centroid. When implementing this in Python, users will plot the dataset and determine the number of clusters. It is important to note that K Means clustering requires the user to know the number of clusters.

Density Based clustering is a clustering method that partitions the dataset into clusters based on the density of regions. The dataset is converted into highly dense and low dense regions. Two variables, eps and minpts, are identified by the user. Eps defines the neighborhood around the individual data point. If two data points have a distance lower than the eps, they are considered to be in the same neighborhood. Next, minpts is the minimum number of data points in the neighborhood. In Density Based clustering, clusters can be divided into core and border points. The core point is where the data points are heavily clustered. The border point is the point that surrounds the core point. Additional noise and outlier points are also shown away from the border point. Density Based clustering does not require the user to know the number of clusters which is helpful when the dataset values are similar.

Hierarchical clustering can be divided into two types: Agglomerative and Divisive. Agglomerative clustering starts with a data point as an individual cluster, and then after each step, the closest pair of data points merge into the cluster. Divisive clustering is where all data points are considered as one cluster, and then after each step, the cluster is split into individual clusters based on distance. Unlike other clustering methods, Hierarchical

clustering can be represented as a dendrogram where the distance between points is evaluated and represented in each cluster. Based on the dendrogram, users can identify the overall distance between each cluster and the number of clusters present. In Hierarchical clustering, linkages such as single, complete, and average can be employed in addition to a distance metric. Single linkage returns the minimum distance between two data points in two independent clusters. Complete linkage returns the maximum distance between each data point. Average linkage returns the average distance between pairs of data points in individual clusters.

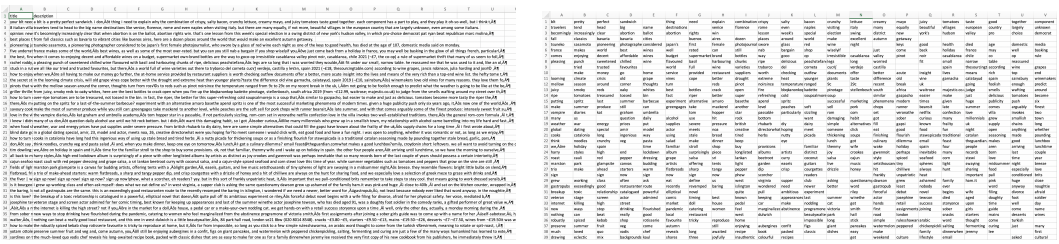
In all three clustering methods mentioned above, a distance metric is utilized to differentiate data points. Distance metrics are important in clustering because they identify clusters' similarity and dissimilarity. Three distance metrics will be employed in this model: Euclidean, Cosine, and Manhattan. The Euclidean distance is also identified as the Pythagorean Theorem's distance metric, which calculates the longest distance of a right triangle. Manhattan distance, also called "city block," calculates the distance between two points in a block-type manner. If we are given a right triangle, the distance is calculated by the sum of traveling along the x and y axes. The Cosine distance measures the angle between two vectors. Generally, it is favorable to use Manhattan distance if the dataset vectors are characterized as a grid. Euclidean distance is the default metric; however, the dimensionality of the data increases, which should be considered when deciding on a distance metric. Lastly, Cosine distance is utilized to better identify similarities between two data points; however, it also leads to higher dimensionality of the dataset.

The Portugal Wine dataset is expected to see three different clusters since the labeled data can be divided into three categories. However, the clustering behavior for Chemical Analysis is hard to determine since chemical concentrations can drastically vary depending on the bottle.

Analyses: Association Rule Mining

Association Rule Mining (ARM) is an unsupervised learning method that utilizes transactional text data. The ARM procedure will observe frequently occurring patterns and associations in the dataset to form rules. Rules can be formatted into an if-then statement where dependencies between items in the dataset are identified. The most famous example is the Market Basket example. A rule for the Market Basket would be that if a consumer purchases milk and bread, then the probability of purchasing coke is high. In forming these rules, the Association Rule Mining algorithm utilizes three parameters: Support, Confidence, and Lift. Support indicates how frequently the rule appears in the database. Confidence is how often the rules are found to be true. Lift is the ratio of confidence in the rule over the expected confidence of the rule. Generally, a lift of greater than one is desired as it indicates positive associations between the elements. A confidence value greater than 0.8 is desired since it indicates that the rule is valid and frequently occurs in the dataset.

This model will utilize the API data gained from News API. It is important to note that the API data gathered includes chunks of descriptions; therefore, it is important to find a way to split the sentences into individual words while filtering out stop words. Please refer to the R code generated to find a way to transfer chunks of code to transactional data. The descriptions were processed so that it was all in lowercase, numerics were removed, punctuations were removed, and stop words were removed. Figure 1 is a comparison of precleaned data with cleaned transactional data. Even after running the code and generating transactional data, some stop phrases were kept in the dataset. Since we are using news data, the phrase "continue reading" was kept in the dataset. Those types of phrases were removed to avoid inaccurate rules. In addition, stop words with unknown symbols were removed from the transactional dataset to improve readability.



(a) (b)

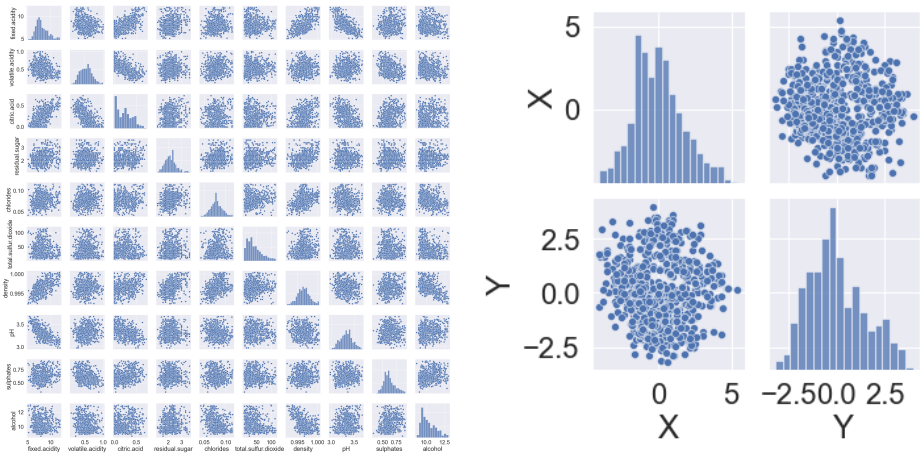
Figure 1: News API Data to Transactional Data

Results

Results: Clustering

Portugal Wine Dataset

The Portugal Wine dataset is a high-dimension dataset with 11 variables. The dataset was prepared for clustering by scaling. Principle Component Analysis with a dimension of two was employed to reduce the dimensions, making it easier to visualize and perform clustering. The original dataset was explored using a pairwise plot. Figure 2a represents the pairwise plot of the original dataset. Most pairwise relationships show heavy clustering in the center and some randomness in the dataset. Figure 2b shows a normal distribution of datapoint in the center after performing PCA and scaling.



(a) (b)

Figure 2: Portugal Wine Dataset Initial Exploration

K Means clustering will be performed using K=2, K=3, K=4, and K=5. In order to determine the optimized K value, the Elbow method will be employed to compare the inertia at varying K values. The K Means clustering models will also be compared using the Silhouette Score. Comparisons of the K Means cluster with varying K values are shown in figure 3.It is important to note that the Euclidean distance was utilized. In the figure, the red dot indicates the centroid of each cluster. The Silhouette score indicates how well the dataset clustered with respect to distance. This score ranges between -1 and 1, where -1 through 0 represent poor clustering. Any Silhouette value close to 1 represents good clustering. In this comparison model, the Silhouette score using K=2 is 0.380, K=3 is 0.397, K=4 is 0.382, and K=5 is 0.371. Comparing these four K Means clustering models, the model with K=3 is the better model since the Silhouette score is closer to one. The Elbow plot backs up this fact. Figure 4 shows the inertia at different numbers of clusters. The optimized number of clusters is determined where non-linearity starts to occur. K values from one to three show decreasing linearity. However, after three clusters, the graph starts to display nonlinear behavior. As a result, the elbow method indicates that the optimized number of clusters is three.

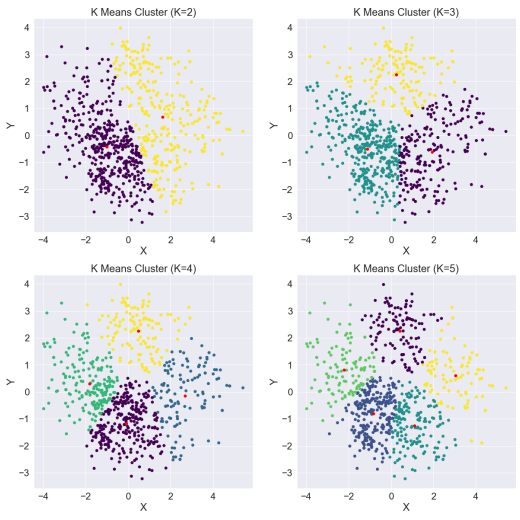


Figure 3: Portugal Wine K Means Comparison Plot

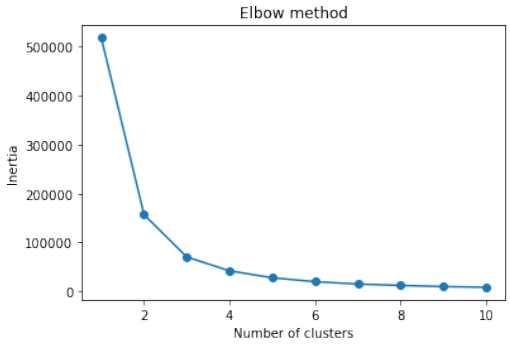
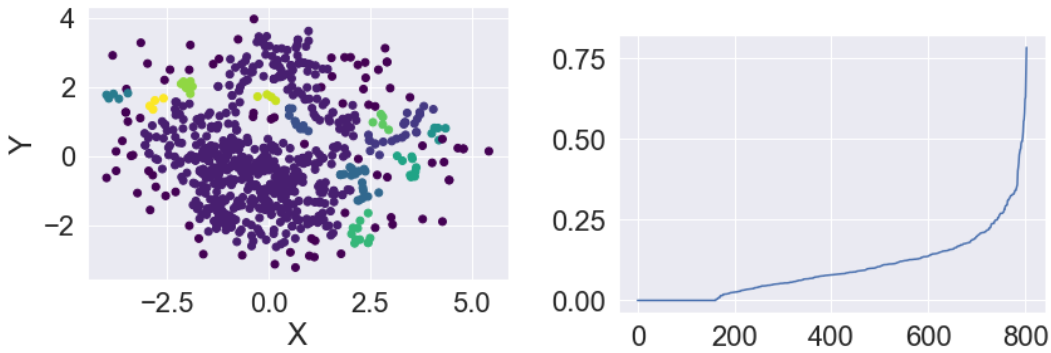


Figure 4: Portugal Wine Elbow Method

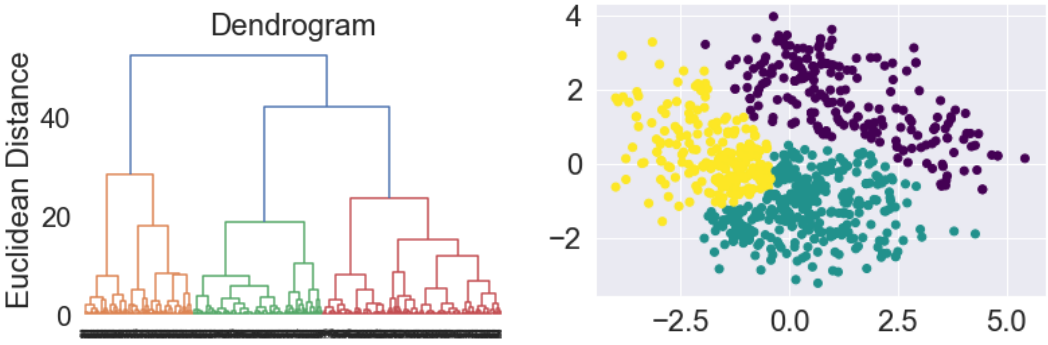
Density Based Clustering does not require the user to input a number of clusters, but optimizing the eps and minpts values is important. The min pts value, in this case, will be four since the dimension of the dataset is two. The eps value was optimized by plotting the nearest neighbor curve, represented in figure 5b. When determining the optimized eps value, it is important to identify the maximum curvature value. In this case, the optimized eps value for the Portugal Wine dataset is 0.3. Figure 5a shows the result of Density-Based Clustering. The purple represents the core point where the dataset is heavily dense. The colors outside the core point could be the boundary point and noise from the dataset.



(a) (b)

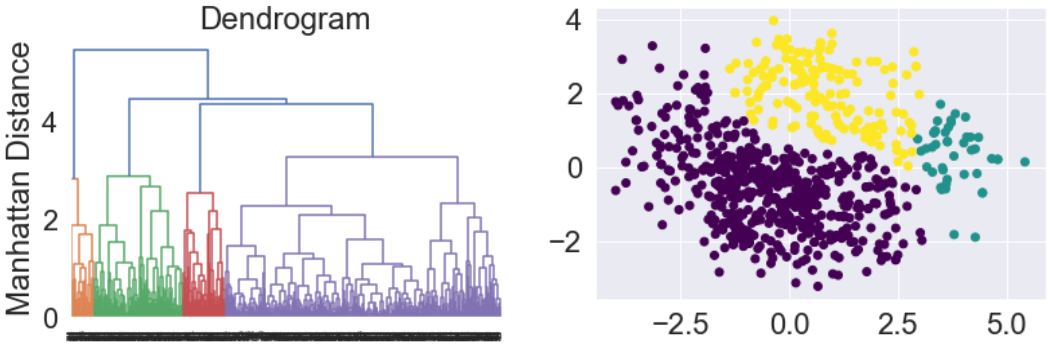
Figure 5: Portugal Wine Density Based Clustering

Hierarchical clustering was performed with three different distance metrics and Agglomerative clustering. Based on the Elbow method, the optimized number of clusters is three; therefore, three clusters were utilized with Agglomerative clustering. The Euclidean Agglomerative clustering and dendrogram are represented in figure 6. When running this model, the ward linkage method was utilized. It is important to note that the default linkage setting is ward. Based on the Euclidean dendrogram, the longest Euclidean distance starts after 30, indicating that three clusters are ideal for this distance metric. The second Hierarchical model utilizes the Manhattan distance. As seen in figure 7a, the dendrogram is skewed to the left. Unlike the Euclidean distance, the Manhattan distance is shorter near the top, making it harder to differentiate the optimum number of clusters. The Agglomerative cluster in figure 7b shows heavy clustering in the purple region, which is not seen in K Means clustering. The third Hierarchical model utilizes the Cosine distance. As seen in figure 8a, the dendrogram shows that the distance is long after Cosine distance of 0.5. Thus, the optimum number of clusters is also three. The Agglomerative cluster using Cosine distance is similar to Euclidean distance and K means because the clusters are equally divided into three clusters.



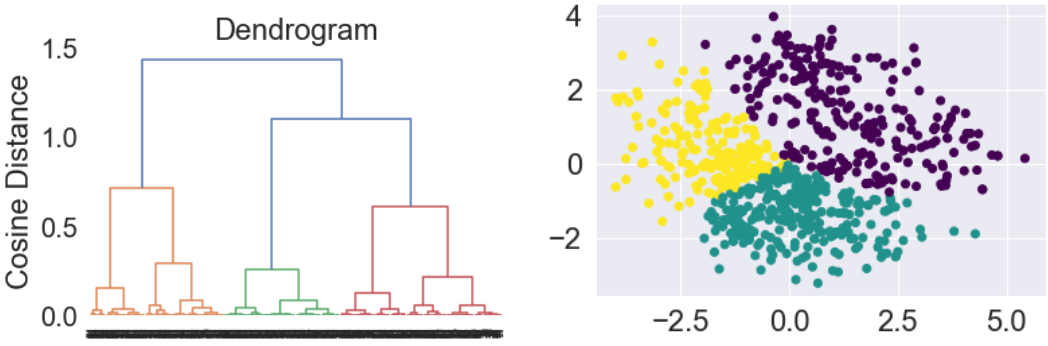
(a) (b)

Figure 6: Portugal Wine Dataset: Euclidean Distance



(a) (b)

Figure 7: Portugal Wine Dataset: Manhattan Distance

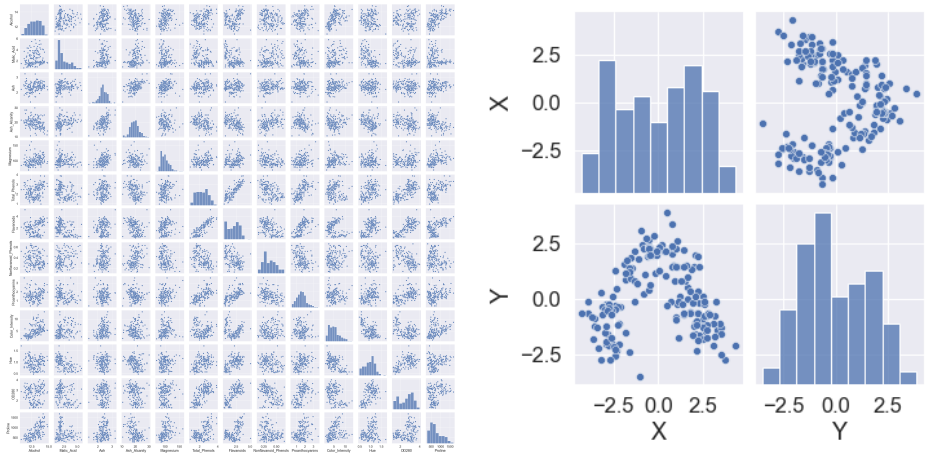


(a) (b)

Figure 8: Portugal Wine Dataset: Cosine Distance

Chemical Analysis Dataset

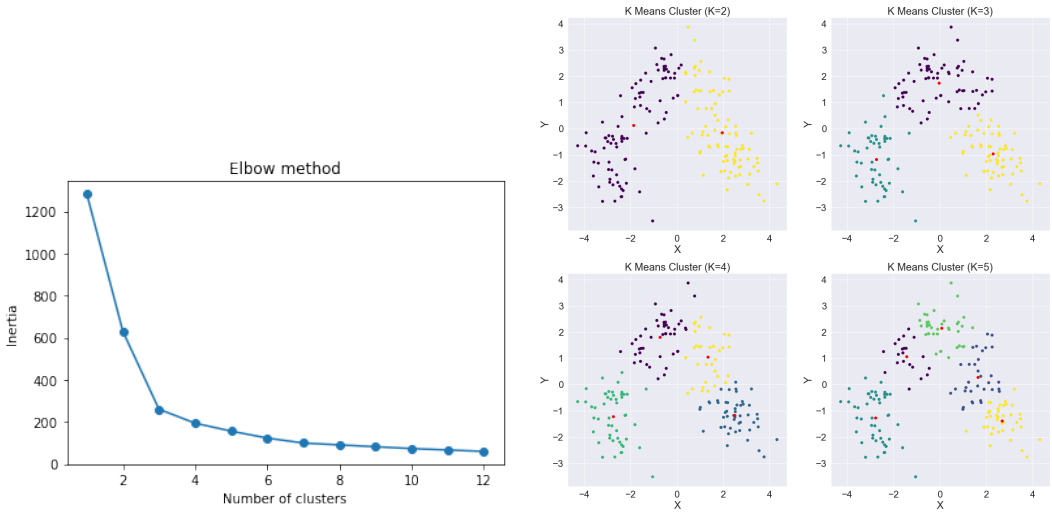
The Chemical Analysis dataset is a high-dimension dataset with 13 variables. The dataset was also prepared for clustering by scaling. Principle Component Analysis with a dimension of two was employed to reduce the dimensions, making it easier to visualize and perform clustering. The original dataset was explored using a pairwise plot. Figure 9b shows a pairwise plot of the transformed dataset. There is a pattern in this plot because it shows some curvature. K Means clustering will be performed using K=2, K=3, K=4, and K=5. The



(a) (b)

Figure 9: Chemical Analysis Dataset Initial Exploration

Elbow method will be employed to compare the inertia at varying K values. The K Means clustering models will also be compared using the Silhouette Score. Comparisons of the K Means cluster with varying K values are shown in figure 10b. In the figure, the red dot indicates the centroid of each cluster. The Silhouette score indicates how well the dataset clustered with respect to distance. In this comparison model, the Silhouette score using K=2 is 0.465, K=3 is 0.560, K=4 is 0.491, and K=5 is 0.441. Comparing these four K Means clustering models, K=3 is the better model since the Silhouette score is closer to one. Figure 10a shows the inertia at different numbers of clusters. Similar to the Silhouette score, the Elbow method also indicates that the optimized number of clusters is three.

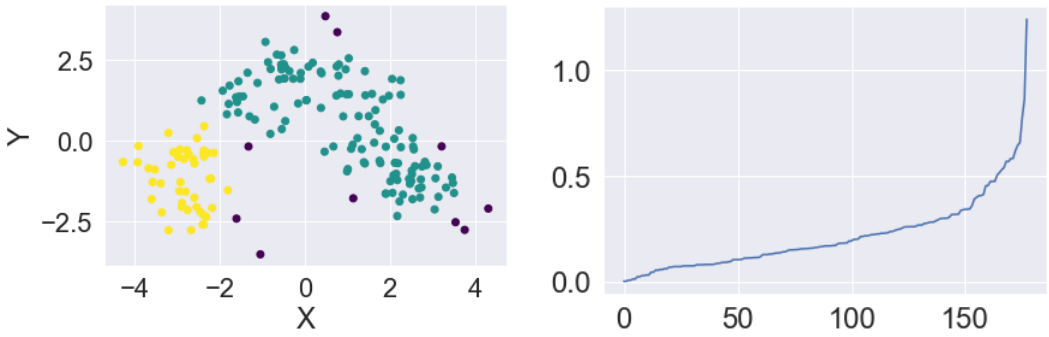


(a) (b)

Figure 10: Chemical Analysis: Elbow Method and K Means Clustering

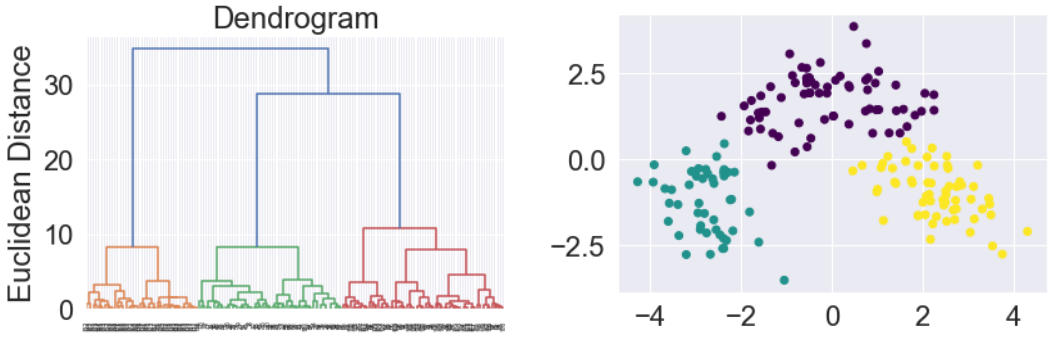
The min pts value for Density-Based clustering will be four since the dimension of the dataset is two. The eps value was optimized by plotting the nearest neighbor curve, represented in figure 11b. In this case, the optimized eps value for the Chemical Analysis dataset is 0.6. Figure 11a shows the result of Density-Based clustering. The blue represents the core point where the dataset is heavily dense. The yellow represents another cluster

in the dataset. The purple can represent boundary point and noise. Hierarchical Clustering was performed with three different distance metrics and Agglomerative clustering. Based on the Elbow method, the optimized number of clusters is three; therefore, three clusters were utilized with Agglomerative clustering. The Euclidean Agglomerative clustering and dendrogram are represented in figure 12. Based on the Euclidean dendrogram, the longest Euclidean distance starts after 10, indicating that three clusters are ideal for this distance metric. The second Hierarchical model utilizes the Manhattan distance. The Agglomerative cluster in figure 13 shows equal clustering similar to the K Means model. The third Hierarchical model utilizes the Cosine distance. As seen in figure 14, the dendrogram shows that the distance becomes longer after Cosine distance of 0.5. Thus, the optimum number of clusters is also three. The Agglomerative clusters show similar results to the K Means model in all three Hierarchical models. There is equal density in all three clusters.



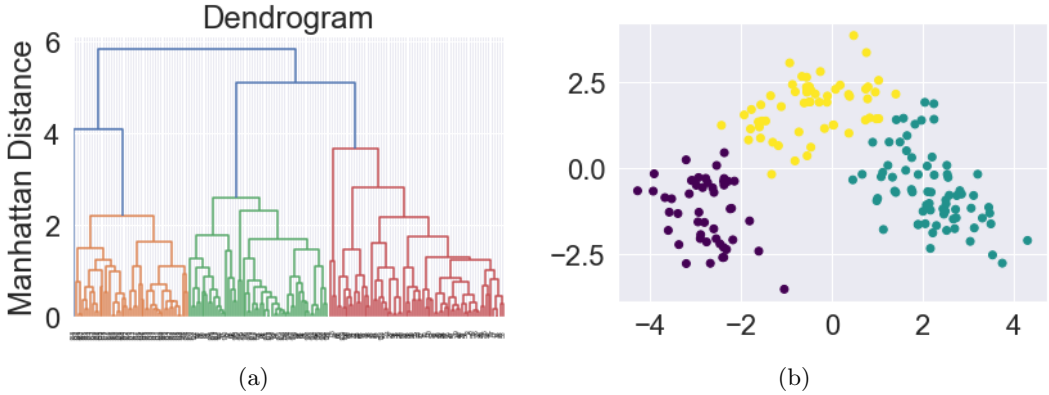
(a) (b)

Figure 11: Chemical Analysis Dataset Density Based Clustering



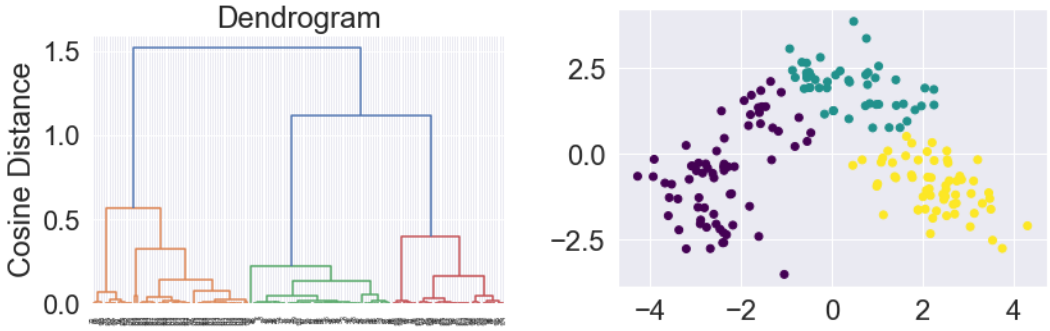
(a) (b)

Figure 12: Chemical Analysis Dataset: Euclidean Distance



(a) (b)

Figure 13: Chemical Analysis Dataset: Manhattan Distance



(a) (b)

Figure 14: Chemical Analysis Dataset: Cosine Distance

Based on all three clustering methods, it is clear that the Chemical Analysis dataset can be divided into three clusters. When comparing the Hierarchical clustering method, it should be noted that the Manhattan distance metric does not perform well compared to other metrics.

Results: Association Rule Mining

Association Rule Mining was performed using the Apriori algorithm in R. The support was set to 0.01, and the confidence was 0.7. Anything with a confidence value greater than 70% was included in the rules. No rules were found when changing the support variable to greater than 0.02; therefore, the support was set to 0.01. Figure 15 and 16 represents the top 20 support, confidence, and lift rules. When examining the lift report, the greatest lift is the rule pinot with noir and noir with pinot. This is an obvious association since pinot noir is a type of wine.

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{started}	=>	{now}	0.01742160	1.00	0.01742160	20.500000	5
[2]	{york}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[3]	{avoid}	=>	{want}	0.01045296	1.00	0.01045296	28.700000	3
[4]	{dragon}	=>	{house}	0.01045296	1.00	0.01045296	47.833333	3
[5]	{varietals}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[6]	{cookbook}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[7]	{grapes}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[8]	{white}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[9]	{tasting}	=>	{wine}	0.01045296	0.75	0.01393728	5.817568	3
[10]	{impressive}	=>	{make}	0.01045296	1.00	0.01045296	16.882353	3
[11]	{pasta}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[12]	{travelers}	=>	{many}	0.01045296	1.00	0.01045296	19.133333	3
[13]	{noir}	=>	{pinot}	0.01045296	1.00	0.01045296	71.750000	3
[14]	{pinot}	=>	{noir}	0.01045296	0.75	0.01393728	71.750000	3
[15]	{fresh}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[16]	{late}	=>	{summer}	0.01045296	0.75	0.01393728	14.350000	3
[17]	{lunch}	=>	{dinner}	0.01045296	1.00	0.01045296	41.000000	3
[18]	{meal}	=>	{dinner}	0.01045296	0.75	0.01393728	30.750000	3
[19]	{started, wine}	=>	{now}	0.01045296	1.00	0.01045296	20.500000	3
[20]	{now, wine}	=>	{started}	0.01045296	1.00	0.01045296	57.400000	3

(a)

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{noir}	=>	{pinot}	0.01045296	1.00	0.01045296	71.750000	3
[2]	{pinot}	=>	{noir}	0.01045296	0.75	0.01393728	71.750000	3
[3]	{now, wine}	=>	{started}	0.01045296	1.00	0.01045296	57.400000	3
[4]	{dragon}	=>	{house}	0.01045296	1.00	0.01045296	47.833333	3
[5]	{lunch}	=>	{dinner}	0.01045296	1.00	0.01045296	41.000000	3
[6]	{pasta}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[7]	{fresh}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[8]	{meal}	=>	{dinner}	0.01045296	0.75	0.01393728	30.750000	3
[9]	{avoid}	=>	{want}	0.01045296	1.00	0.01045296	28.700000	3
[10]	{started}	=>	{now}	0.01742160	1.00	0.01742160	20.500000	5
[11]	{started, wine}	=>	{now}	0.01045296	1.00	0.01045296	20.500000	3
[12]	{travelers}	=>	{many}	0.01045296	1.00	0.01045296	19.133333	3
[13]	{impressive}	=>	{make}	0.01045296	1.00	0.01045296	16.882353	3
[14]	{late}	=>	{summer}	0.01045296	0.75	0.01393728	14.350000	3
[15]	{york}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[16]	{varietals}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[17]	{cookbook}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[18]	{grapes}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[19]	{white}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[20]	{tasting}	=>	{wine}	0.01045296	0.75	0.01393728	5.817568	3

(b)

Figure 15: Support(a) and Lift (b)

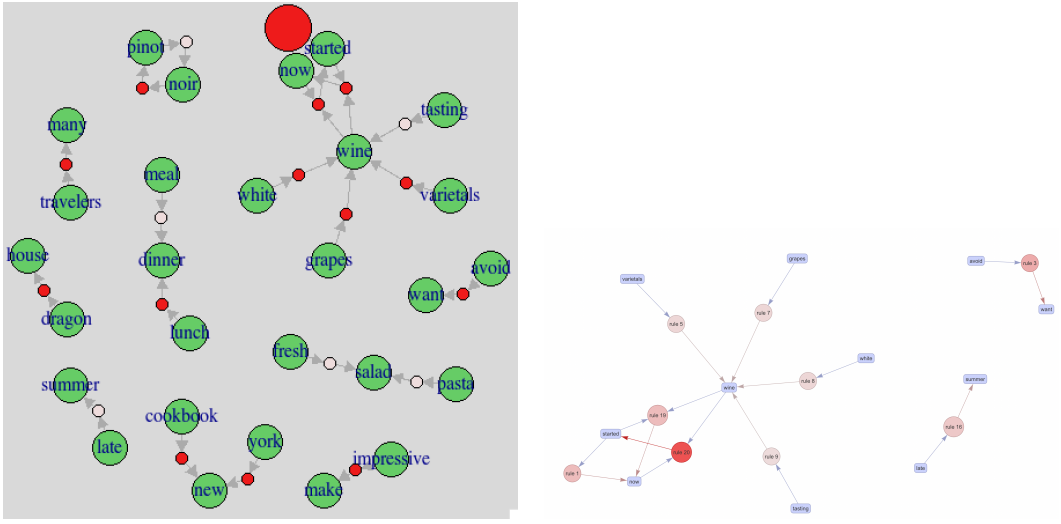
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{york}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[2]	{avoid}	=>	{want}	0.01045296	1.00	0.01045296	28.700000	3
[3]	{dragon}	=>	{house}	0.01045296	1.00	0.01045296	47.833333	3
[4]	{varietals}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[5]	{cookbook}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[6]	{grapes}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[7]	{White}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[8]	{impressive}	=>	{make}	0.01045296	1.00	0.01045296	16.882353	3
[9]	{travelers}	=>	{many}	0.01045296	1.00	0.01045296	19.133333	3
[10]	{noir}	=>	{pinot}	0.01045296	1.00	0.01045296	71.750000	3
[11]	{lunch}	=>	{dinner}	0.01045296	1.00	0.01045296	41.000000	3
[12]	{started}	=>	{now}	0.01742160	1.00	0.01742160	20.500000	5
[13]	{started, wine}	=>	{now}	0.01045296	1.00	0.01045296	20.500000	3
[14]	{now, wine}	=>	{started}	0.01045296	1.00	0.01045296	57.400000	3
[15]	{tasting}	=>	{wine}	0.01045296	0.75	0.01393728	5.817568	3
[16]	{pasta}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[17]	{pinot}	=>	{noir}	0.01045296	0.75	0.01393728	71.750000	3
[18]	{fresh}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[19]	{late}	=>	{summer}	0.01045296	0.75	0.01393728	14.350000	3
[20]	{meal}	=>	{dinner}	0.01045296	0.75	0.01393728	30.750000	3

(a)

Figure 16: Confidence

Furthermore, it makes sense that pinot noir is associated with new articles because it is one of the most famous and popular wine varieties in the United States. As the lift decrease, food associations are observed. Foods such as pasta and salad are popular associations. When examining the support, the greatest support association is new with york and started with now. The support is a list of rules with the most frequent occurrences. In

this case, the most popular associations in news articles are started with now and new with york. Based on this result, it would be safe to assume that New York was highly associated with wines from August through September. When exploring the confidence list, the greatest confidence is new with york and varietal with wines. Like the support list, New York is also high on the confidence list. The second highest association is varietals and wine. This is also expected since wine types can be divided into different varietals. The News API data gained from August through September show high associations between wine and travel, food, New York, cookbook, tasting, white wine, and red wine.



(a) (b)

Figure 17: Overall Network (a) and "Wine" Network (b)

Figure 17 represents the network graphs of the 20 association rules generated by apriori and the network graph of the word wine. It is clear that wine is associated with white, tasting, varietal, and grapes. Those are expected associations with wine. In addition to the obvious associations, interesting associations include travelers and many.

Conclusion

When exploring the clustering results of the Portugal Wine dataset, it is clear that the dataset has three different clusters. Although the dataset is heavily clustered near the center, the K means and Hierarchical clustering methods indicate that three different clusters are present. These three different clusters could represent the "good", "bad", and "average" quality wines. If the three clusters do represent the three different qualities, then it would be interesting to explore further the optimum concentrations of chemical compounds that make up the "good" ,"average, and "bad" quality wines. The Chemical Analysis dataset clustering also indicates three different clusters. It is important to note that the Chemical Analysis dataset only contains quantitative information about chemical compounds. Therefore, this clustering method highlighted that the dataset has three independent groups with similar chemical makeup. In both datasets, the Manhattan distance metric performed unsatisfactorily because it was hard to determine the optimum number of clusters. Both results from the Manhattan distance metric showed skewed dendrogram results, and the distances were relatively short near the top. However, the Cosine and Euclidean distance metrics were consistent in that equal cluster group areas were present. The clustering results gained from the Portugal Wine and Chemical Analysis datasets revealed that wine quality depends on the chemical compound concentrations.

The attitude of news outlets and the public towards wine was analyzed through Association Rule Mining. From August through September, the transactional data indicates that news outlets showed frequent relationships between wine and food, traveling, and varietals. Words such as varietal, grape, white, and tasting were frequently associated with wine, indicating the public interest in wine. Furthermore, cooking and food were commonly associated with wine, highlighting frequent associations between food and wine pairings. It is important to note that traveling and summer are common words from August through September, which could justify the frequent associations of "summer" words with wine. The Association Rule Mining results seem obvious since the associations discovered through this process are common relationships seen with wine. Therefore, news outlets and the public seem to have somewhat normal associations with wine.

When performing clustering and Association Rule Mining, a few interesting facts were discovered during the process. The Portugal Wine and Chemical Analysis datasets are high-dimensional with more than ten variables.

Therefore, it is important to perform correct dimensionality reduction techniques. Principle Component Analysis was performed on both datasets. However, it would be interesting to see how Linear Discriminant Analysis or other classification methods impact the outcome of clustering methods. When dealing with text data, the first task was to transform the chunk of text data into transactional data. Through exploring the transactional data, it was evident that News API gathers information from less credible sources. Although it was assumed that news outlets would have correct spelling and level of professionalism, it was clear that some news outlets had multiple spelling mistakes and unprofessional language. In addition, the phrase "continue reading" was common in the transactional dataset. Therefore, it would be interesting to see how transactional data would change if news data were gained from a few well-known news outlets.