Wyatt Rasmussen

DSC 680

Wine Clustering

Business Problem

The business problem that I am trying to solve is taking a dataset of wines that contain the chemical makeup and group together similar wines. By understanding the chemical makeup of the wines we will be able to cluster together similar groups of wine. This could help us with understanding how the chemical makeup affects the flavors of wines. With the clusters that we have created we will be able to give recommendations of the wines based on the flavor profile that someone may enjoy.

Background/History

The wine making process is a science. In the process there are 5 main steps that introduce different flavors, colors, and other attributes. Those 5 steps in the process are the harvest, crushing and pressing, fermentation, clarification, and aging and bottling. The process that the wine makers go through will impact the chemical makeup of the wine. This introduces a lot of different features of wine.

Data Explanation

The data I found for this project was on Kaggle. It contains 13 different features and 178 different wines. There is one file supplied named, “wine-clustering.csv”. Here is the link to the Kaggle dataset:

https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering?select=wine-clustering.csv

These different features that are included in this dataset relate to the chemical makeup of the wine. As mentioned in the background and history section those differences come up through the wine making process.

Methods

For this project the primary method will be using K-Means Clustering. This method is commonly used in clustering unsupervised learning problems. K-Means clustering is an algorithm that tries to minimize the distance of the points to the cluster centroid. Before starting the K-Means clustering, I am using the standardizing the data using Standard Scaler. This will set the features to a standardized value based on the mean of the feature. To understand the best clusters, I am using a few different types of component analysis. Principle Component Analysis, or PCA, is a common technique used to minimize the features in the dataset. This technique is used to simplify the number of features in a dataset while also making the machine learning process simpler on the backend. T-distributed stochastic neighbor embedding, or tSNE, is the next feature reduction techniques that I will be using. tSNE is another commonly used technique that can help make sense out of a dataset with a large number of features. It seems to be most effective with hundreds of features. The last feature reduction technique will be Uniform Manifold Approximation and Projection, or UMAP. This technique is very similar to tSNE with a few smaller differences.

Using these different feature reduction techniques I will be performing K-Means clustering on the dataframes. This will give me different clusters of the data. Once I complete one of the techniques with K-Means I attach the cluster array to the dataframe as a column to use later in visualizing the clusters. This will give me 3 different sets of clusters in the dataset. With those 3 clusters I will then export the dataframe as a csv file and import that into Tableau for visualizing the clusters with the features in the dataset.

Analysis

Using the different feature reduction techniques I have created a few different visualizations of the components that were created into clusters. The first two figures will be using PCA for feature reduction and clustered using K-Means clustering. Those clusters are in Figure 1. Each cluster is represented by a different color with the 'X' signifying the centroid. In Figure 2, I was also able to build a three-dimensional plot of the K-Means clustering using PCA to get a better understanding of the clusters.

Next I clustered and plotted using a different feature reduction technique named tSNE. Figure 3 shows the two dimensional results of the K-Means clustering using tSNE. Each color is representative of a different cluster with the 'X' showing the centroid in that cluster. After clustering 2D, I created a three-dimensional plot in Figure 4. This shows the clusters of the tSNE using K-Means clustering. Each color represents a different cluster.

Lastly I clustered and plotted using my last feature reduction technique, UMAP. Figure 5 shows the two-dimensional K-Means clustering results using UMAP. Each color represents a different cluster and the ‘X’ shows the centroid of the cluster. In Figure 6 you can see a three-dimensional plot of the K-Means clustering using UMAP. Each cluster is represented in a different color.

Now that I have each set of data clustered into different sections I can compare the sizes of the clusters, using PCA cluster 0 has 49 wines, cluster 1 has 64 wines, and cluster 2 has 65 wines. Using tSNE, cluster 0 has 55 wines, cluster 1 has 65 wines, and cluster 2 has 58 wines. Using UMAP cluster 0 has 54 wines, cluster 1 has 65 wines, and cluster 2 has 59 wines.

After attaching the cluster to each wine for each feature reduction technique, I pulled the data into Tableau to do another series of visualizations of the results. In this series I compared the original features using different colors for the cluster and stacked the visualizations in order to compare the ability of the clustering. This allows to see the difference in the clustering techniques. Figure 7 shows the comparison of the Proline-Flavanoids clusters. As you can see in general the clusters are very similar, with only minor differences. These features clustered together more than other features that I tested. In Figure 8, you can see the scatter plots of the ash and alcohol features. With this one there is less clearly defined clusters, but you can see the change in some of the techniques where individual points have a different cluster attached to them. Lastly, Figure 9 shows the scatter plots of the color intensity and hue features. This plot shows another successful clustering groups on the features. For the most part one cluster has a lower color intensity, another cluster has a higher hue and higher color intensity, and the last cluster has a lower hue and higher color intensity.

Conclusion

For this unsupervised learning problem, I believe I was able to accurately place these wines into different clusters. This will allow for a better understand of the chemical makeups of each of these and what chemicals affect the cluster the most. UMAP seemed to cluster the data the closest of the different feature reduction techniques.

Assumptions

A few assumptions were made for this analysis. The biggest assumption that was made is that these clusters should be grouped into three different clusters, 3 clusters seems to fit the data based on the PCA visualizations. Another assumption that was made was that the clusters are accurate, since this is unsupervised learning we don't have a validation technique.

Limitations

I think the primary limitations here is the amount of wines that we have, since we only have about 180 different wines that creates a problem of having very few actual data points. This could impact how things are clustered as we continue to collect the data on more wines.

Challenges

The biggest challenge I am facing is the validation techniques with the unsupervised learning. In supervised learning you can set aside a testing and validation dataset that will allow you to cross test the results of your modeling, with unsupervised learning that luxury isn't afforded to you. This presents me with the challenge of knowing accuracy.

I also am being challenged in how I present the data in their clusters, I could present the data simply in the clusters that I found using PCA, tSNE, or UMAP, or I could present it in clusters of each feature. This present an interesting challenge.

Future Uses/Additional Applications

Since this is unsupervised learning, we need to continue to evaluate the results that we have and tweak the algorithm as we gather more information going forward. This process could be applied to other alcohol types like whiskey, beer, etc.

Recommendations

My recommendations for these results would be to continue analyzing the data. We are seeing some promising results. This can really help us with recommending wines to customers who enjoy a certain flavor profile. By continually tweaking the algorithm we can potentially make the clusters a bit more pointed into potentially subclusters within the main cluster.

Implementation Plan

The implementation plan for this is a little tricky. This should continue to get more accurate as we continue to use this and gain more data. We should probably start with this more loosely aiding us until we gain more confidence in the results of the modeling.

Ethical Assessment

For this project the most important ethical consideration is that alcohol is illegal for minors so to not use this information in the context of minors.

Questions That Could Be Asked

1. How can you validate results?

   a. There isn’t exactly a great way to validate results in unsupervised learning. Going forward we would continue to evaluate as we got more data.

2. How do you know 3 clusters was the right amount?
   a. In general determining the cluster amounts is difficult, but three clusters seemed to fit the data the best based on the feature reduction techniques.
3. How can we improve?
   a. We can improve with more data. With having more data we would likely see larger clusters and sharper changes in the data.
4. Who is the primary audience for these results?
   a. The primary audience of these results would be wine makers who want to understand how the chemical makeup affects the wines.
5. What other areas can clustering be used?
   a. Clustering could also be used in different types of alcohols fairly easily here.
6. Why is unsupervised learning the right approach here?
   a. Unsupervised learning is the right approach here because we are taking the data and attempting to understand it better. Going in, we didn’t have the end results that would be required to do supervised learning.
7. Which visualizations do you think are most successful?
   a. I thought the feature reduction two-dimensional clusters were the easiest to see the differences, but I thought it was interesting to see the variety of results in the clusters being applied to the individual featues.

8. What other visualization techniques could you use other than scatter plots here?
    a. I could have also used box and whisker plots as well as histograms to show the distribution of each of the features since they were all numerical.
9. Why do you use standardization techniques in this problem?
    a. Standardization techniques helps to reduce the variety of numbers and centralizes all of the features around 0 and the variation is based on standard deviation.
10. Are there other feature reduction techniques that could have been used?
    a. There are quite a few other techniques such as linear discriminant analysis (LDA) or latent semantic analysis (LSA).

References

https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering?select=wine-clustering.csv

https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

https://www.nvidia.com/en-us/glossary/data-science/clustering/

https://help.tableau.com/current/pro/desktop/en-us/clustering.htm

https://towardsdatascience.com/cluster-analysis-in-tableau-1f19acd0c647

https://umap-learn.readthedocs.io/en/latest/clustering.html

https://distill.pub/2016/misread-tsne/

Figures

Figure 1

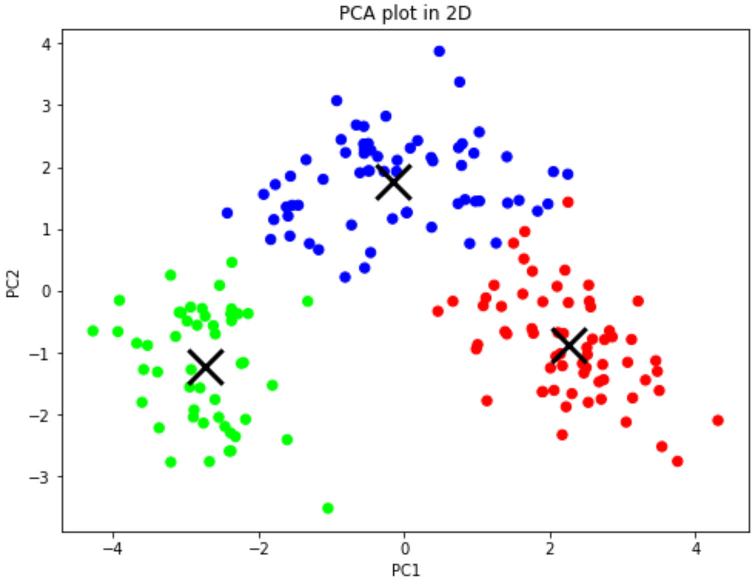


Figure 2

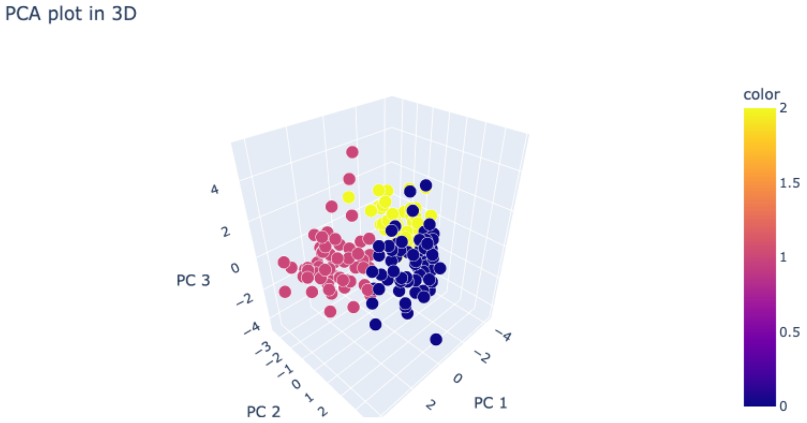


Figure 3

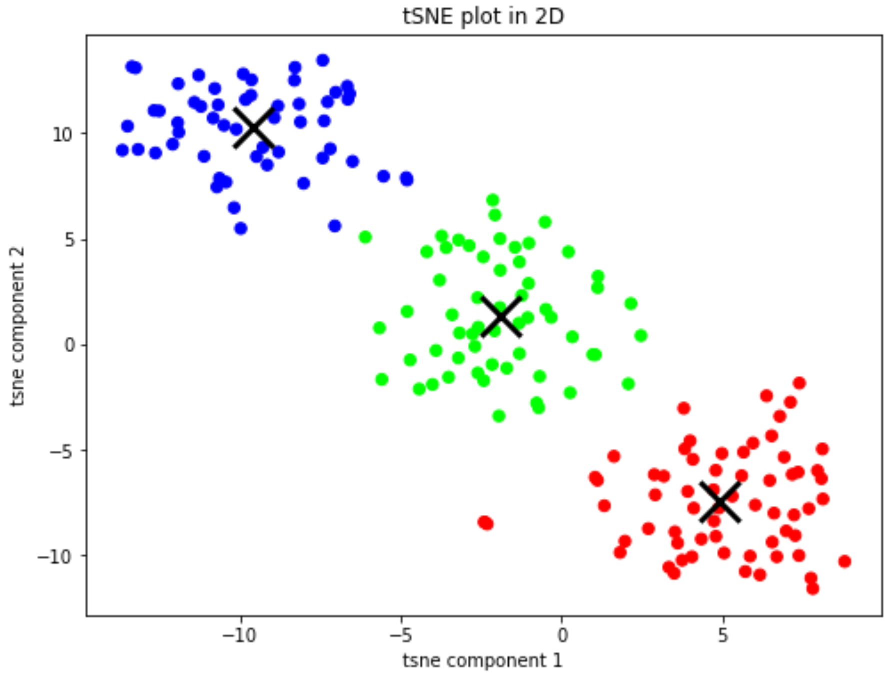


Figure 4

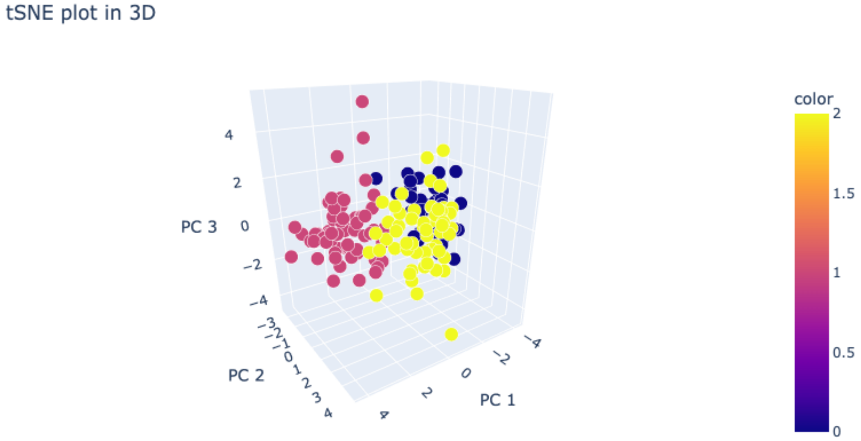


Figure 5

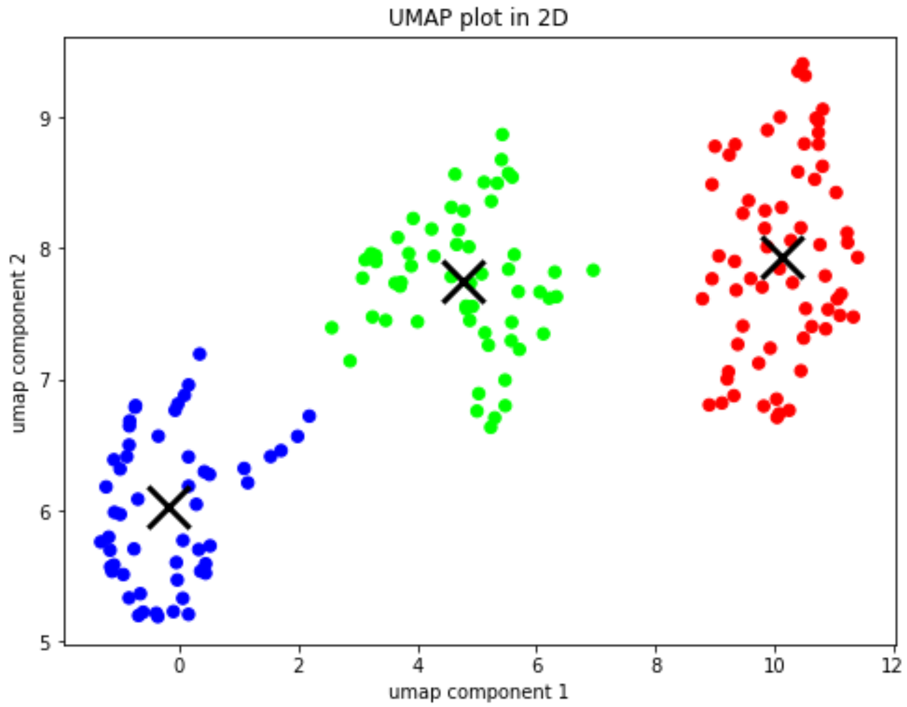


Figure 6

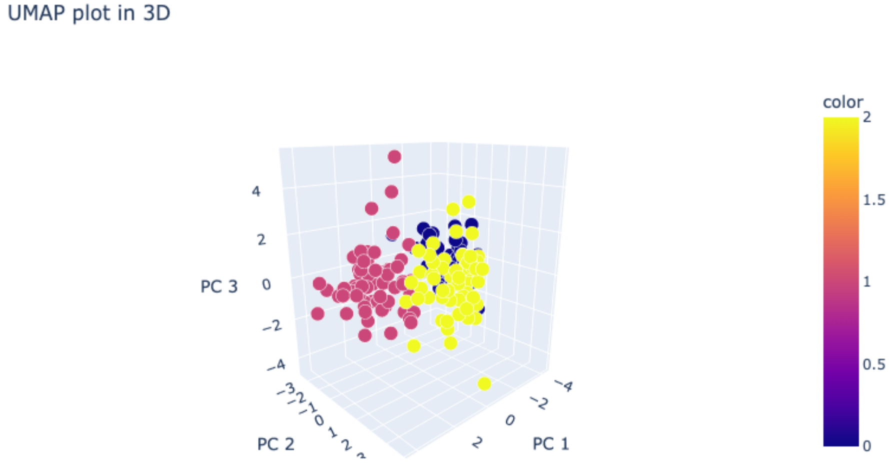
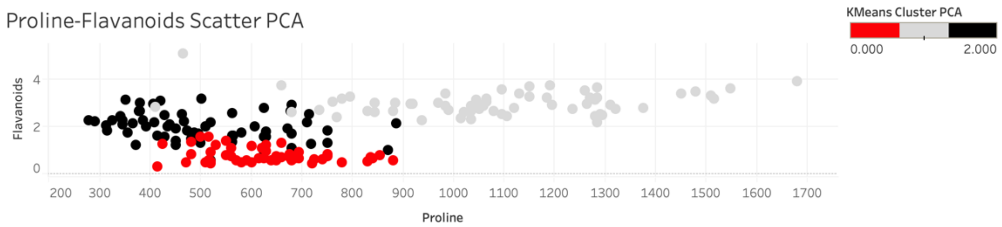
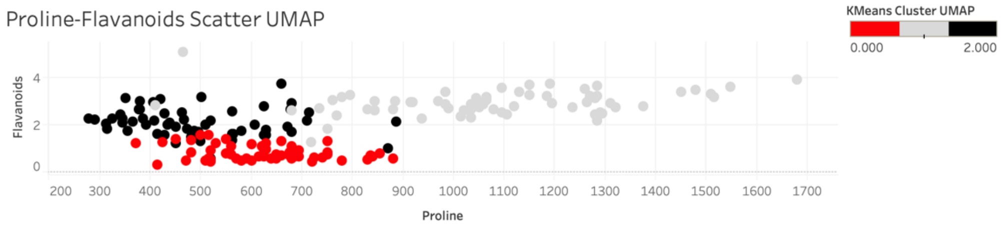


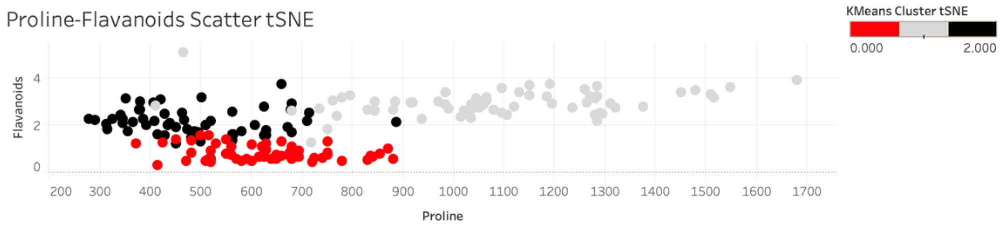
Figure 7



Proline vs. Flavanoids. Color shows details about KMeans Cluster PCA. Details are shown for KMeans Cluster PCA.

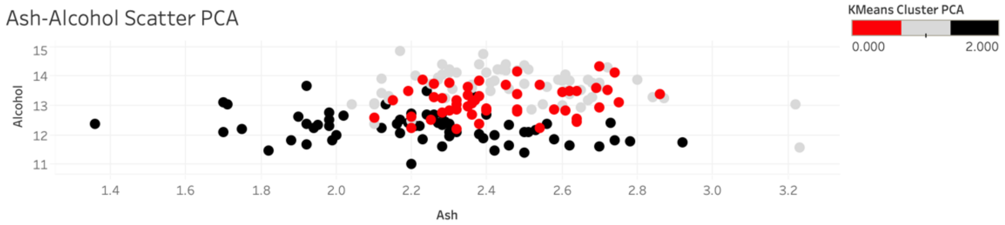


Proline vs. Flavanoids. Color shows details about KMeans Cluster UMAP. Details are shown for KMeans Cluster UMAP.

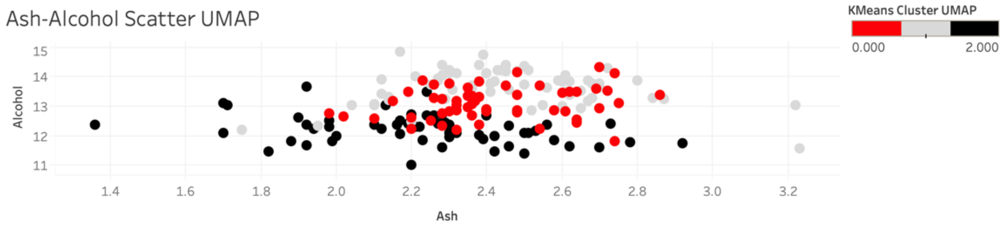


Proline vs. Flavanoids. Color shows details about KMeans Cluster tSNE. Details are shown for KMeans Cluster tSNE.

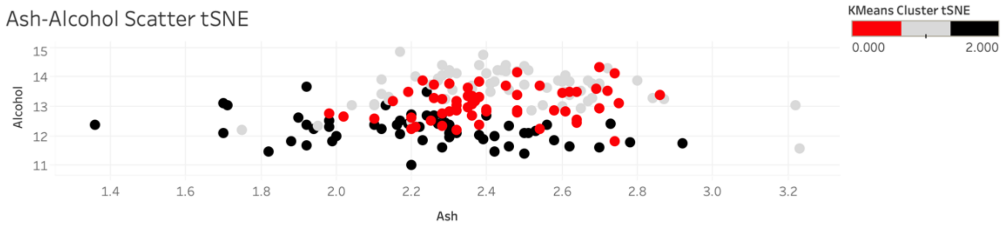
Figure 8



Ash vs. Alcohol. Color shows details about KMeans Cluster PCA. Details are shown for KMeans Cluster PCA.

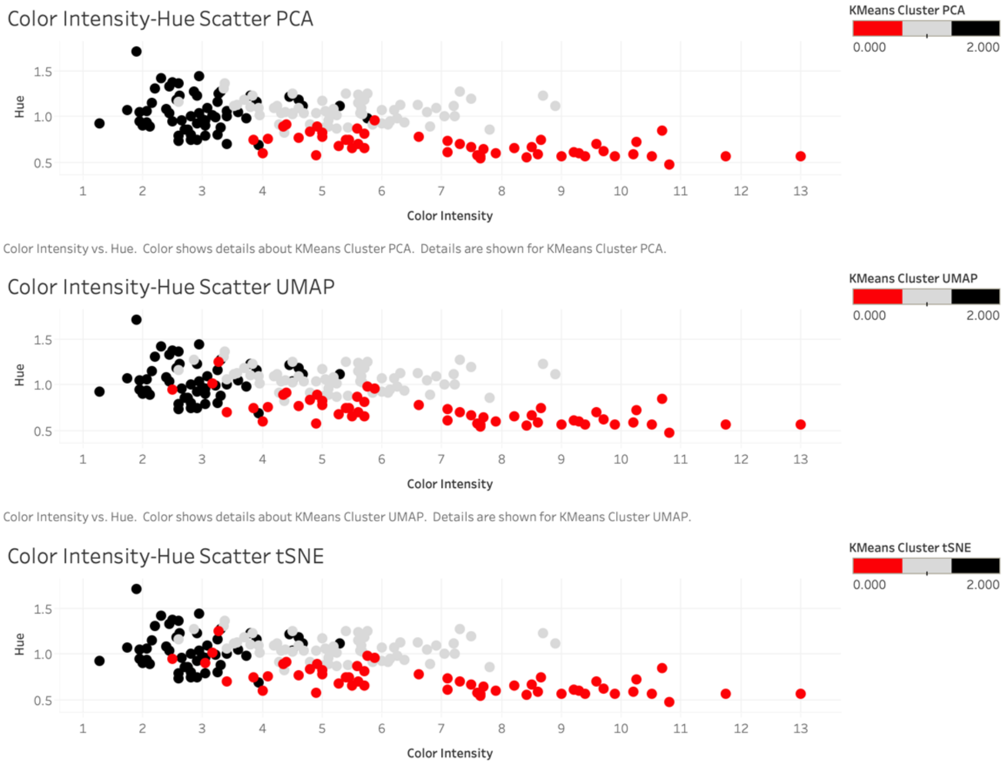


Ash vs. Alcohol. Color shows details about KMeans Cluster UMAP. Details are shown for KMeans Cluster UMAP.



Ash vs. Alcohol. Color shows details about KMeans Cluster tSNE. Details are shown for KMeans Cluster tSNE.

Figure 9



Color Intensity-Hue Scatter PCA

Color Intensity vs. Hue. Color shows details about KMeans Cluster PCA. Details are shown for KMeans Cluster PCA.

Color Intensity-Hue Scatter UMAP

Color Intensity vs. Hue. Color shows details about KMeans Cluster UMAP. Details are shown for KMeans Cluster UMAP.

Color Intensity-Hue Scatter tSNE

Color Intensity vs. Hue. Color shows details about KMeans Cluster tSNE. Details are shown for KMeans Cluster tSNE.