Leveraging Feature Dimensionality Reduction and Clustering to Identify Transient Relationships between Wines

Zachary Tsouprakos
Northwestern University
MSDS – School of Professional Studies

1 Introduction

In this paper, we will highlight the research that was leveraged in order to support a wine manufacturer's initiative of increasing wine sales. From an analytics perspective, our goal is to create a scalable approach that can draw similarities across wines, and that is agnostic to changes introduced into the wine market. From a consultative perspective, our goal is to provide recommendations to our client which will first, yield high wine sales, second, inform which wines should be bought and manufactured after, and third, create opportunities for future growth. We will highlight an overview of the data set that was leveraged, along with the methodology and approach around data pre-processing, creating an orthogonal data set, and modeling.

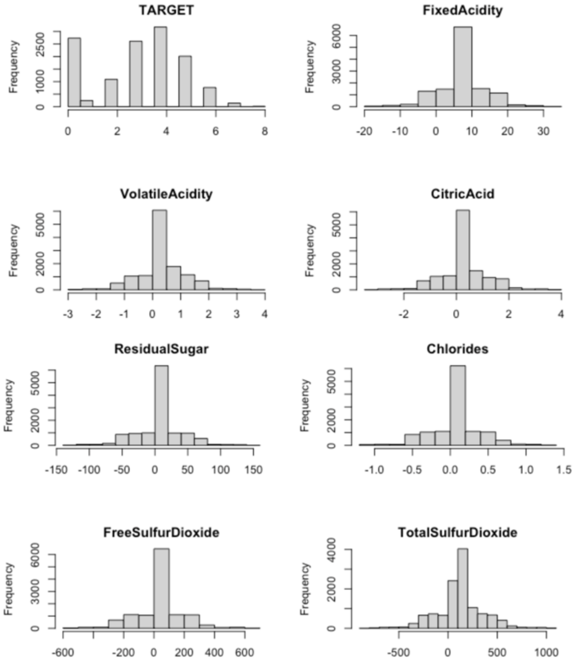
2 Methods and Findings of Exploratory Data Analysis (EDA)

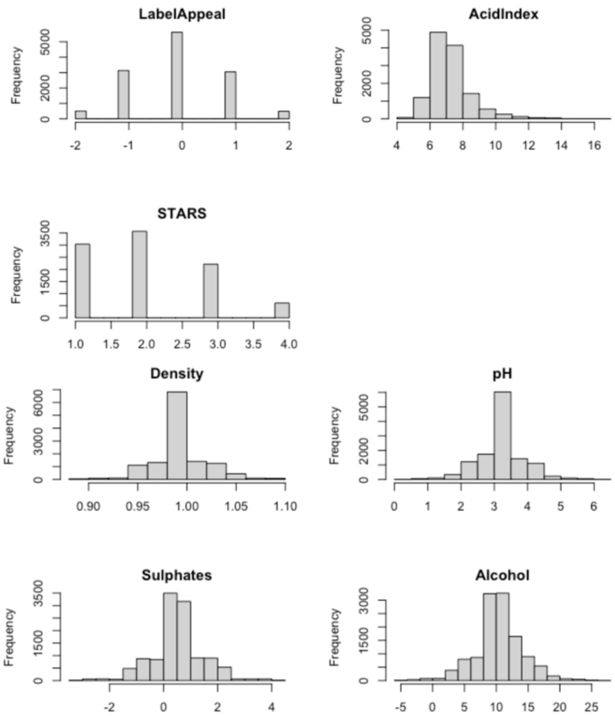
2.1 Data Set Introduction

For this analysis, we leveraged a data set of 12,795 wines that are commercially available today. Each record within the utilized data set represents a unique wine and has the following features associated to them: *TARGET (cases sold), fixed FixedAcidity, VolatileAcidity, CitricAcid, ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, Density, pH, Sulphates, Alcohol, LabelAppeal, AcidIndex, STARS.* These chemical properties will be leveraged as inputs to our model which will help draw similarities across wines, as well as to highlight potential opportunities the manufacturer can explore. We are fortunate that this data set contains all continuous variables; however, each feature has a relatively unique scale, while some are ordinal in nature (stars, acidIndex, etc.). These unique measures will be dealt with while discussing pre-processing. Prior to pre-processing we will need to make a decision around which variables will be included and removed from our data set.

2.2 EDA

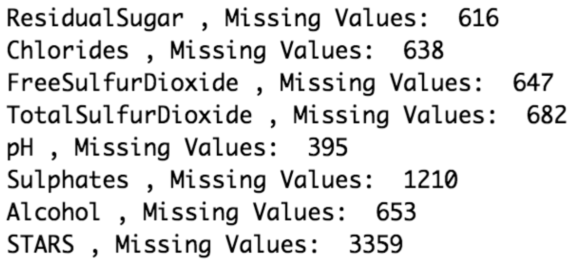
Our objective while conducting exploratory data analysis was to truly understand our data set, and document how we need to manipulate our data set to align with our use-case. The first step was to understand the distribution of each variable; this will tell us whether a feature is skewed and if it is necessary to handle any outliers. The below visuals represent histograms for each feature





The first takeaway from the above histograms highlights the over-index on zero for the target variable. This tells us that there are many wines that have not been sold or bought by consumers. This is not necessarily an issue due to our use-case; the target variable will be removed prior to modeling as we do not want the cases of wine sold to influence our outcome. Instead, we want to use target as a descriptor of each cluster that is formed. The other variable that raises a bit of question is AcidIndex; we see a bit right skewed distribution. The outliers are not drastic enough to be considered extreme *(>= (3 * IQR))*, allowing the data to be left as is. Speaking for the other variables in this data set, we see relatively normal distributions that do not raise any concerns.

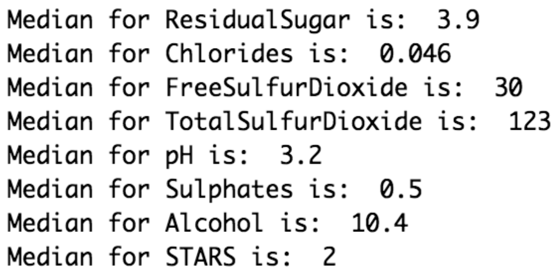
The next step of our EDA process explores missing values and highlights which features will imputing. This step. Is quite critical as it is necessary to impute these missing values prior to dimensionality reduction. A simple loop through our features highlighted that we are missing values for the following variables:


```
ResidualSugar , Missing Values:  616
Chlorides , Missing Values:  638
FreeSulfurDioxide , Missing Values:  647
TotalSulfurDioxide , Missing Values:  682
pH , Missing Values:  395
Sulphates , Missing Values:  1210
Alcohol , Missing Values:  653
STARS , Missing Values:  3359
```

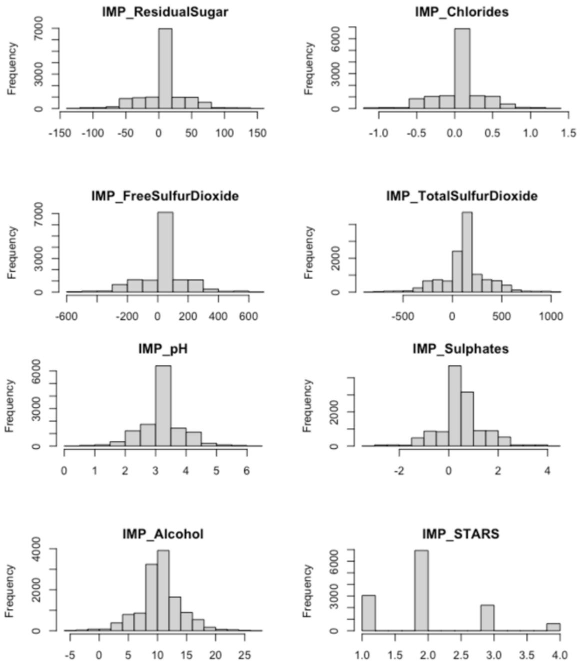
The above prepares us to move into our next phase of this analysis where we massage our data to prior to modeling.

3 Data Preparation

As we alluded to earlier, prior to dimensionality reduction and clustering, it is critical that we handle our missing values. The first step is to identify features that have missing values, which has been completed. Now we must determine how we will be imputing these missing variables. An industry-wide solution, which is very common across use cases, is to calculate the median of each feature and replace missing values accordingly. While this does not fit every use case, we have determined that this fits ours appropriately. The below highlights the calculated mean for each variable that contained missing values. Moving forward, one can assume that each value missing will be replaced by the below medians, respective of each feature.


```
Median for ResidualSugar is:  3.9
Median for Chlorides is:  0.046
Median for FreeSulfurDioxide is:  30
Median for TotalSulfurDioxide is:  123
Median for pH is:  3.2
Median for Sulphates is:  0.5
Median for Alcohol is:  10.4
Median for STARS is:  2
```

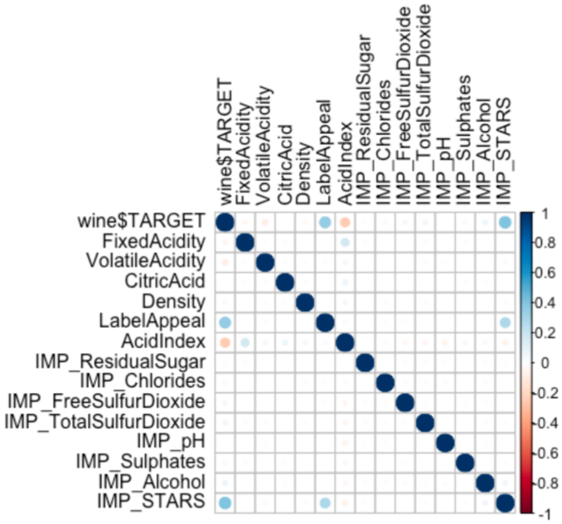
Prior to any imputations, we created a copy of each feature that contained missing values and pre-pended “IMP_” to the feature name. This was done in order to preserve the integrity of the original data set. This will also allow future researchers to understand that the feature they are interacting with has been imputed. Post data imputations, it was important that we preserved the normal distribution of each feature. The below histograms reflect the distribution of the features that were manipulated.



The above histograms confirm that the features continued to be normally distributed post imputations. IMP_STARS is an exception due to the ordinal nature of the feature.

4 Analysis Methods

Post EDA and data pre-processing, we move on to dimensionality reduction and clustering. We have 14 features and would like to avoid multicollinearity; prior to modeling we leveraged Principal Component Analysis (PCA) to create an orthogonal data set and reduce features in Principal Components that explain variance amongst the original data set. Prior to reducing features, it is important to highlight correlated features. The below is a visual that highlights values within a correlation matrix.



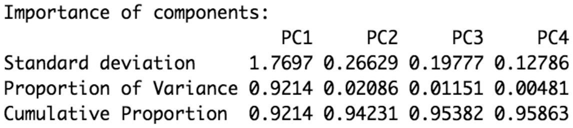
Most features within our data set are loosely correlated; however latent relationships can exist within our data.

4.1 Principal Component Analysis

As previously stated, our goal for PCA is to mitigate multicollinearity but also reduce our features. Prior to conducting PCA, it was determined that we will take as many principal components needed to explain a minimum of 90% variance. In other terms, we will take the number of components necessary to represent 90% of our data. Also, it was concluded that we will take at least 2 principal components to ensure a two-dimensional aspect to our data. With this line drawn in the sand, we moved forward with PCA.

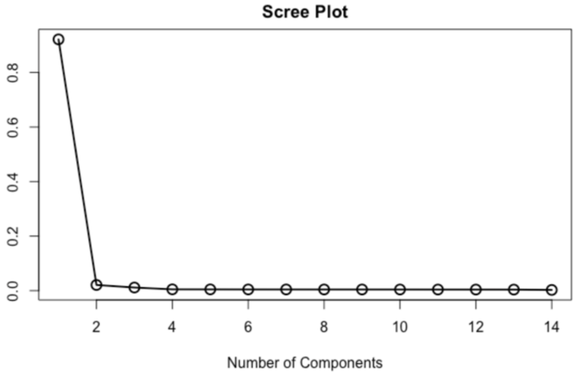
4.1.1 PCA Attempt #1

Our first attempt with PCA was an “all-in” approach. With the exception of the target variable, all features were passed into the PCA algorithm in order to reduce our features. While 14 components were a result of PCA, the below displays summary statistics for the first four principal components.

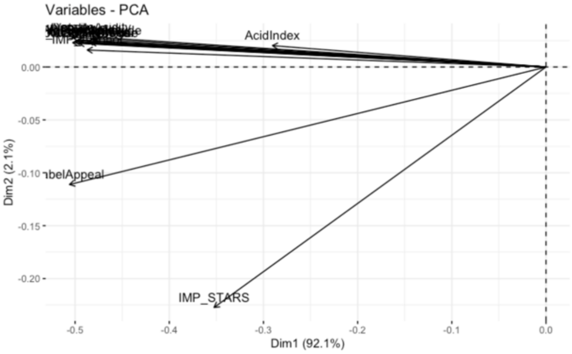


```
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.7697 0.26629 0.19777 0.12786
Proportion of Variance 0.9214 0.02086 0.01151 0.00481
Cumulative Proportion  0.9214 0.94231 0.95382 0.95863
```

This first attempted yielded promising results; the first components explain 92% of variance. The first two principal components yield a cumulative explained variance of 94%. This passes our initial requirement of explaining a minimum of 90% variability. To reinforce our decision of selecting the first two components, we leveraged a scree plot to visualize the appropriate number of components needed.



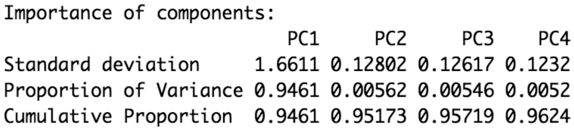
The above scree plot visualizes the explained variance for each principal component. After 2 components, we see that there is not much more variance that can be explained, supporting our notion that two principal components are appropriate for this analysis. One last item in regard to testing is the influence of features on each principal component.



The above map of feature influence shows an over index on two features: IMP_STARS and labelAppeal. These two features are influencing the first principal component much differently than the rest of the other features. With these findings, we will run another PCA model; however, this time we will remove IMP_STARS and labelAppeal.

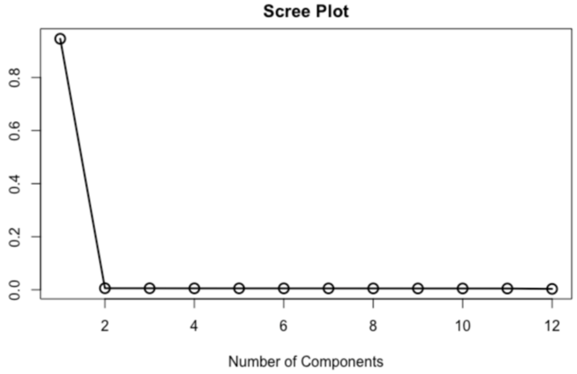
4.1.2 PCA Attempt #2

In our second attempt at PCA, we are looking for an evenly distributed measure of influence for each feature on our principal components. The concern around over-influence is that our clusters can be shaped be the underlying influence of stars and label appeal. These two variables are good measures to compare across each cluster. We ran a similar approach as our first attempt highlight the summary of our first four principal components after the removal of these two features.

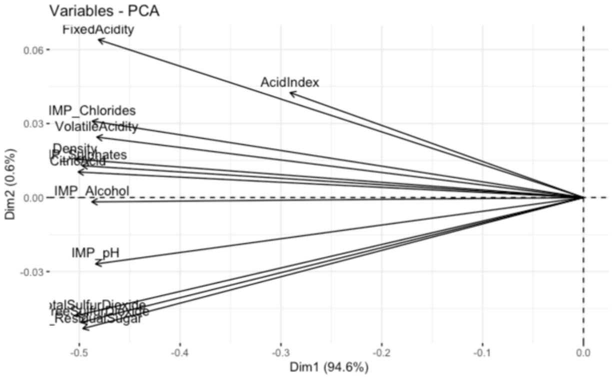


```
Importance of components:
                          PC1     PC2     PC3    PC4
Standard deviation     1.6611 0.12802 0.12617 0.1232
Proportion of Variance 0.9461 0.00562 0.00546 0.0052
Cumulative Proportion  0.9461 0.95173 0.95719 0.9624
```

Again, we have promising results as the first two principal components can explain for 95% of the data sets variability. The below screen plot shows a similar story and supports our claim of taking the first two principal components.



The final check is to confirm our hypothesis of evenly distributed influence on our principal components.



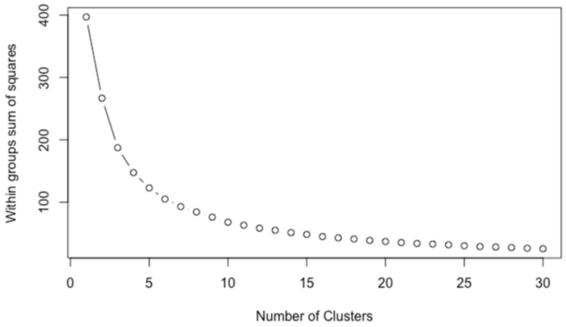
The above map of our feature influence confirms that one feature is not over influencing the output of our principal components.

Our second attempt at PCA yielded results that allowed for confidence prior to moving into clustering. Prior to clustering, the scores for the first two principal components were calculated in order to be passed into our clustering model.

4.2 K-Means Clustering

After successfully reducing our data set from 14 features to two principal components, we were ready to move forward with clustering. Leveraging the K-Means algorithm, we passed the scores of our first two principal components into this model. An input to K-Means that must be defined prior to modeling is the number of anticipated clusters. Similar to PCA, an additional threshold was set around the accuracy of our model; a minimum of 80% accuracy was required to feel comfortable with our results. We leveraged two techniques to determine the number of clusters needed for this use-case. Along with a simple equation to calculate accuracy.

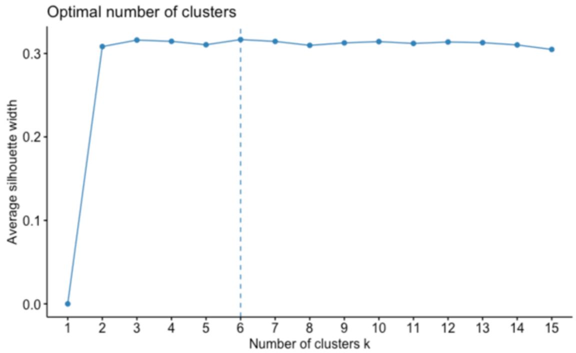
First, we began with producing an elbow plot which highlights the sum of squares (error) found within each cluster (group). We simulated this calculation in a for-loop to understand how the error changes as we introduce more clusters.



The above simulation shows the known error as we iterate through models containg zero to 30 clusters. The objective of the above test is to highlight where the change in error begins to stop changing drastically. This indicates that the variance in error is becoming small and we have reached an optimal number of clusters needed.

Variability in error begins to diminish between eight and ten clusters.

Calculating the silhouette score was the second test performed to understand the optimal number of clusters. The silhouette scores check the Euclidean distance of each observation from its assigned centroid. This metric helps highlight whether observations are being assigned appropriately, or if there are observations that are closer to other cluster centers than the centroid it is assigned to. Silhouette scores that are greater than zero represent clusters that have minimal overlap. As values begin to move under zero, one can infer that observations are not being assigned to clusters appropriately.

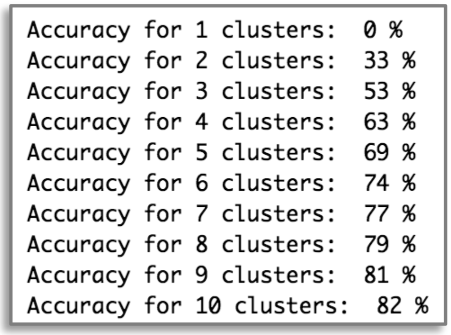


Similar to our elbow plot, we simulated different K-Mean models ranging from one to 15 clusters. The above silhouette plot highlights six clusters as the most optimal; however, ten resembles a very similar score and aligns well with the elbow plot we previously highlights.

Our final test was to calculate accuracy for each model simulation. As previously mentioned, our accuracy must be a minimum of 80%. Our method behind calculating accuracy is as follows:

$$Accuracy = \frac{Sum\ of\ Squares\ Between\ Clusters}{Total\ Sum\ of\ Squares}$$

This ratio provides a descriptor of fit and accuracy which allowed us to understand the appropriate number of clusters necessary and will help support or reject our elbow and silhouette plots. Taking a similar approach, we iterated through K-Mean models that had clusters ranging from one to ten.

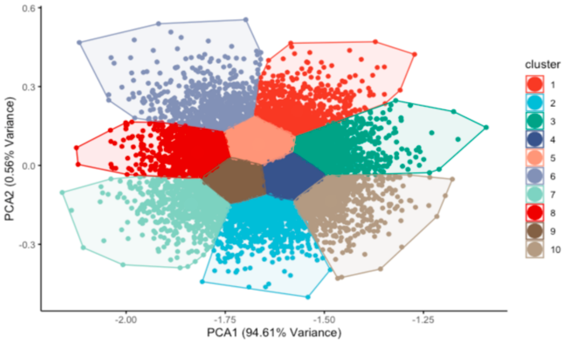


```
Accuracy for 1 clusters:  0 %
Accuracy for 2 clusters:  33 %
Accuracy for 3 clusters:  53 %
Accuracy for 4 clusters:  63 %
Accuracy for 5 clusters:  69 %
Accuracy for 6 clusters:  74 %
Accuracy for 7 clusters:  77 %
Accuracy for 8 clusters:  79 %
Accuracy for 9 clusters:  81 %
Accuracy for 10 clusters:  82 %
```

Abiding by our minimum requirement of 80% accuracy (fit), we can select any number of clusters, greater than 9.

5 Final Model Presentation

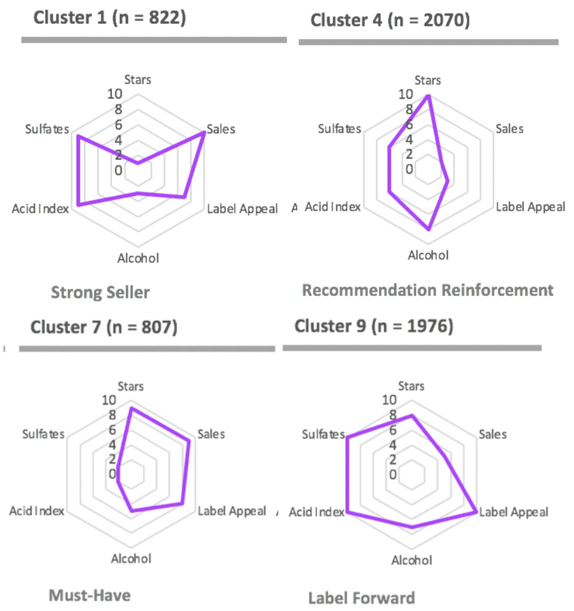
Thus far, we have spoken about exploratory data analysis, data pre-processing, feature dimensionality reduction, and K-Means model stress testing. The inputs have allowed us to create the most optimal model for or use-case by leveraging our first two principal components and a K-Means model that produces ten clusters.



The above visual representation highlights the scores of each observation plotted against our first two principal components. Each observation's color represents the cluster it belongs to; at first glance, we were not able to see clusters that overlapped with one another. This reinforced our silhouette analysis and provided confidence around the number of clusters selected.

Post-clustering, we must re-visit our objective of understanding the differences between wines and which yield the most sales. Along with highlighting which wines performed the best, we must understand where there is opportunity for our manufacturer. Prior to analyzing our clusters, each cluster number was appended back to our imputed data set, allowing for a comparison of wines across each cluster.

While all clusters were carefully compared, the below four show the most room for opportunity.



Clusters one and seven contain 822 and 2,070 wines respectively. These two clusters are wines that contain a combination of chemical properties that perform the best in regard to sales. Interestingly enough, wines found in Cluster one does not have great reviews (stars) from customers; however, this could be a cheap wine that is affordable and accepted be the general wine community. Cluster seven on the other hand has very high reviews (stars), the second highest sales, and the third highest label appeal. A recommendation to our manufacturer would be to continue to produce these wines or mirror future wines off of the chemical properties found in this cluster. While there are macro-factors that impact sales, from the data we have available, mirroring these two clusters can be considered a low-risk initiative by the manufacturer.

Clusters four and nine offer our manufacturer areas to explore. Cluster four has the highest reviews (stars) and the lowest sales. Our data set is limiting in the sense that we can only conduct quantitative studies. I would recommend our manufacturer to partner with a qualitative research group and subject matter experts to understand why sales are low, but customer reviews are high. Speaking directly with consumers can create a feedback loop to improve wines that are being sold today. Marrying quantitative and qualitative research can empower our manufacturer to optimize this cluster to result in high sales and satisfied customers.

Cluster nine is another unique group of wines that allows for further research and potential opportunity. These wines are very appealing to the consumer eye and has the highest label appeal of all clusters. Along with label appeal, this cluster has the third highest reviews (stars) and mimics the chemical properties of wines found in cluster one, which are the highest sellers. I would recommend to our manufacturer to conduct a marketing campaign in order to get this cluster of wines “to the front of the store” in order to increase the number of consumers seeing these wines. There is a potential that this cluster of wines, today, is not marketed properly, resulting in the lack of sales.

6 Conclusion and Reflection

This project provided a use-case that allowed for multiple avenues that could have been explored. Along with these hypotheses, there were also multiple machine learning techniques that could have been utilized. Due to timing constraints, there was a natural limitation around exploring different techniques; however, PCA and K-Means resulted in results that yielded actionable recommendations for our manufacturer client.

While the results provided insights that were worthy of recommending, if time was not a constraint, there are a couple of items that could have been improved. First, our data set. Our data was not clean (most data sets are not), and assumptions were made without an industry expert providing guidance. In a perfect world, a subject matter expert would have advised on how to impute missing values, and whether provided more information around each feature. Moving on to the feature dimensionality reduction algorithm, PCA. Our data set’s features were not heavily correlated; however, there could have been latent relationships that were not detected by normal correlation. In future iterations, additional algorithms can be testing such as t-SNE or autoencoders.

K-Means clustering produced digestible results and actionable clusters. The overall performance of K-Means satisfied our requirements; however, there are other clustering options that could have been tested (hierarchical). Also, there are different K-Mean initialization methods that could have been tested, such as K-Means ++, Bradley and Fayyad’s, and more.

There were several items that were reiterated as a result of conducting this analysis. First, stakeholder alignment. Prior to diving deep into analysis, understanding what the client or business needs will help guide the approach, as well as answer any preconceived assumptions. Second, always understand the data set being utilized. Wine is a common household item; however, its metadata is quite esoteric. Third, set an objective prior to modeling. It is more than critical to start the modeling process with objectives defined and thresholds documented (accuracy, error, etc.). Finally, understand the audience. Not every stakeholder involved in research or an analysis will be technical. Breaking down these large cumbersome tasks into digestible insights and action items is critical to ensure stakeholders truly understand the findings at-hand.