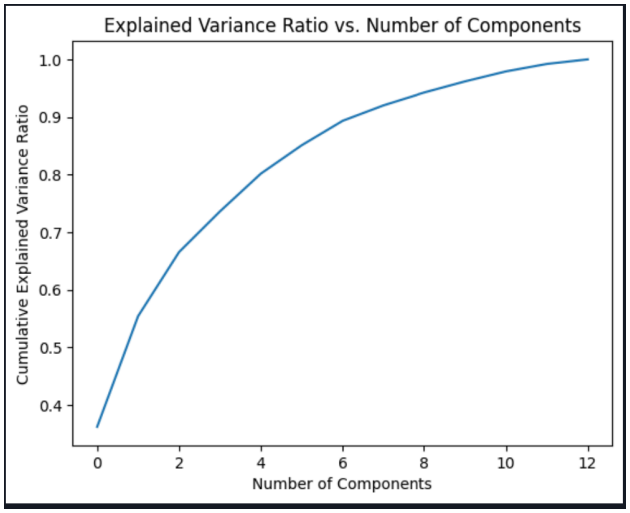
1: Background

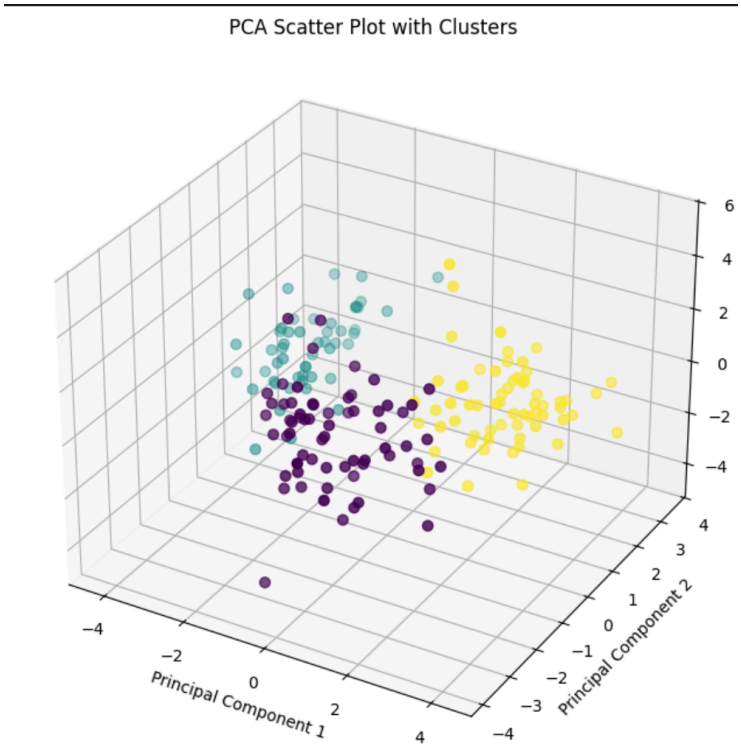
The data that was used was sourced from Kaggle. The data set was downloaded as a csv file through the terminal and pandas was used to read the data. The dataset comprises different wines made from 3 different regions of Italy. The data was collected from a chemical analysis in which 13 different attributes were formed to analyze the types of wines. These attributes are {Alcohol, Malic Acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavonoids, Non Flavonoid phenols, Proanthocyanins, Color intensity, Hue, OD280 of diluted wines and Proline}. The defining locations about where the wines were made were removed for the purpose of utilizing clustering to solve the problem. Thus, the main motive of this task is to find if there are any different classifications of wines that can be made from the sample which infer different locations and methodologies of wine production across the 3 Italian locations.

2: Methods

The methodologies that will be used are those that were mentioned in class, specifically K-Means clustering and PCA dimensional analysis. PCA dimensional analysis will be used to preprocess the data to reduce computational time and complexity, and to eliminate noise and redundancies. In order to use PCA, we need to know how many dimensions we should use. In order to do this, we can use the cumulative explained variance ratio to measure how many dimensions are needed to cover most of the variance. We plotted a graph that displayed the cumulative explained variance ratio vs the amount of components. What we found was that a good threshold balance we can use is 75% variance retention with 3 components.



We use K-Means with 3 clusters to represent the three regions of wine production. K-Means works by creating three centroids and assigning data points to each centroid to represent the mean of the points. The centroids are then moved around to better represent the three clusters. In python, we used the K-Means import from Scikit-learn to do the K-Means clustering for us with the processed PCA data. We then graphed the results on a 3-D graph using the Matplotlib library.



Cluster 1: Turquoise | *Cluster 2*: Purple | *Cluster 3*: Yellow

This shows the three clusters in relation to the three features created by the PCA dimensionality reduction.

3: Results

After viewing the graphs, we printed out the weighted values of the attributes that influenced the PCA dimensions using python.

PCA Weightings:

	Component 1	Component 2	Component 3
Alcohol	0.144329	0.483652	-0.207383
Malic_Acid	-0.245188	0.224931	0.089013
Ash	-0.002051	0.316069	0.626224
Ash_Alcanity	-0.239320	-0.010591	0.612080
Magnesium	0.141992	0.299634	0.130757
Total_Phenols	0.394661	0.065040	0.146179
Flavanoids	0.422934	-0.003360	0.150682
Nonflavanoid_Phenols	-0.298533	0.028779	0.170368
Proanthocyanins	0.313429	0.039302	0.149454
Color_Intensity	-0.088617	0.529996	-0.137306
Hue	0.296715	-0.279235	0.085222
OD280	0.376167	-0.164496	0.166005
Proline	0.286752	0.364903	-0.126746

Component 1:

- Flavonoids (0.4229)
- Total_Phenols (0.3947)
- OD280 (0.3762)
- Proanthocyanidins (0.3124)

Due to the weightings of the Flavonoids and the Total_phenols, clusters that score high on Component 1 tend to have high phenolic content. This means that they will be darker and red in colour (which means that they are made with the entire grape) and will feature health benefits such as reduced risk of cancer and cardiovascular disease [1].

Component 2:

- Colour Intensity (0.53)
- Alcohol (0.4837)
- Proline (0.3649)

Clusters that score high on Component 2 indicate wines that are high on Alcohol and Proline. Higher Proline means more viscosity and sweetness. High colour intensity indicates that wines that fall under this category tend to be more opaque [2].

Component 3:

- Ash (0.62620)
- Ash Alkalinity (0.6120)

Clusters that score high on Component 3 indicate wines tend to have more Ash content. Ash is a form of inorganic salt which has the potential of giving wine a fresh taste [3].

4: Conclusions/Findings

From our findings, when looking at the graph, *Cluster 1* which is represented by Turquoise on the graph scores high on Component 2 and low on Components 1 and 3. This means that wines on *Cluster 3* are sweeter, more viscous, and more opaque. *Cluster 2* is represented by purple and scores very high Component 2, and higher on Component 3 compared to *Cluster 1*. This means that *Cluster 2* also features more sweeter and viscous attributes, however it also has higher ash percentages. *Cluster 3* scores high on Component 1 and somewhat high on Component 3. This means that wines classified on this cluster have more phenolic acids contributing to a distinct red colour and perceived health benefits.

In conclusion, one can see that when using the K-Means methods, there are 3 clusters that are formed which are influenced by the 3 dimensions of the PCA analysis. The clusters possess different qualities and attributes which indicates that they were made in 3 distinct locations in Italy across varying cultures.

Sources

[1] https://www.livestrong.com/article/422207-wines-with-high-levels-of-polyphenols/
[2] https://www.awri.com.au/wp-content/uploads/2024/01/s2388-proline.pdf
[3]https://webofproceedings.org/proceedings_series/ESSP/ETMHS%202019/ETMHS19309.pdf