1 Introduction

1.1 Data

Obtained from **kaggle**

https://www.kaggle.com/danielpanizzo/wine-quality#wineQualityWhites.csv

Number of Instances - red wine

1599 cases

Missing Attribute Values

None

1.2 Description of attributes

column	attribute	Attribute interpretation
1	fixed acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2	volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3	citric acid	found in small quantities, citric acid can add 'freshness' and flavor to wines
4	residual sugar	the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
5	chlorides	the amount of salt in the wine

column	attribute	Attribute interpretation
6	free sulfur dioxide	the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7	total sulfur dioxide	amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8	density	the density of water is close to that of water depending on the percent alcohol and sugar content
9	pH	describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10	sulphates	a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11	alcohol	the percent alcohol content of the wine
12	quality	score between 0 and 10

2 Data Processing

Question I: How to predict red wine quality
2.1 Principal Component Analysis

2.1.1 Pre-processing Data

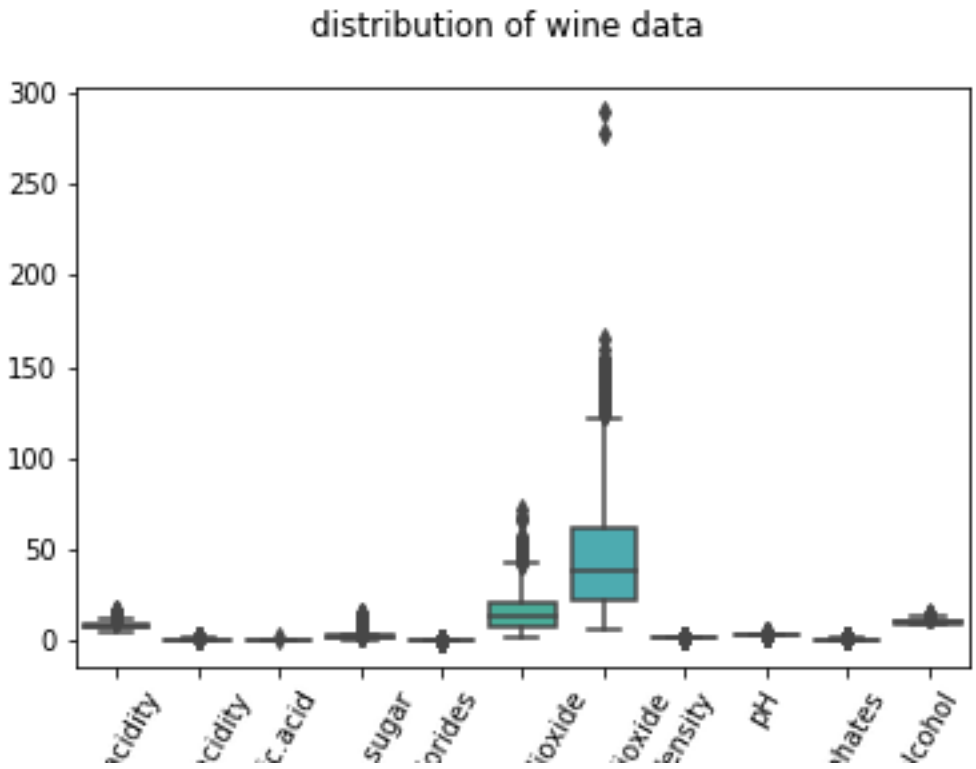

Figure 1 A box plot showing the range of values of the raw data

According to box plot of wine data, each attribute scale of data is at different unit, especially that of **free sulfur dioxide** and **total sulfur dioxide.** Therefore, it is necessary to standardize the data.

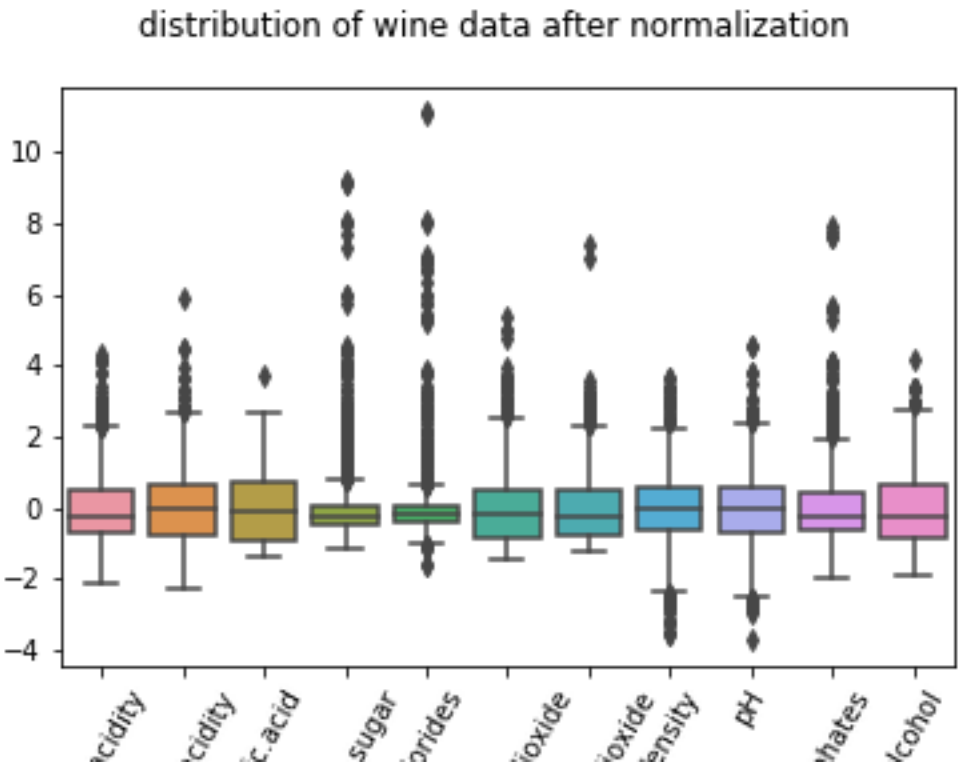

Figure 2 A box plot showing the range of values of the standardized data

2.1.2 Labeling Data

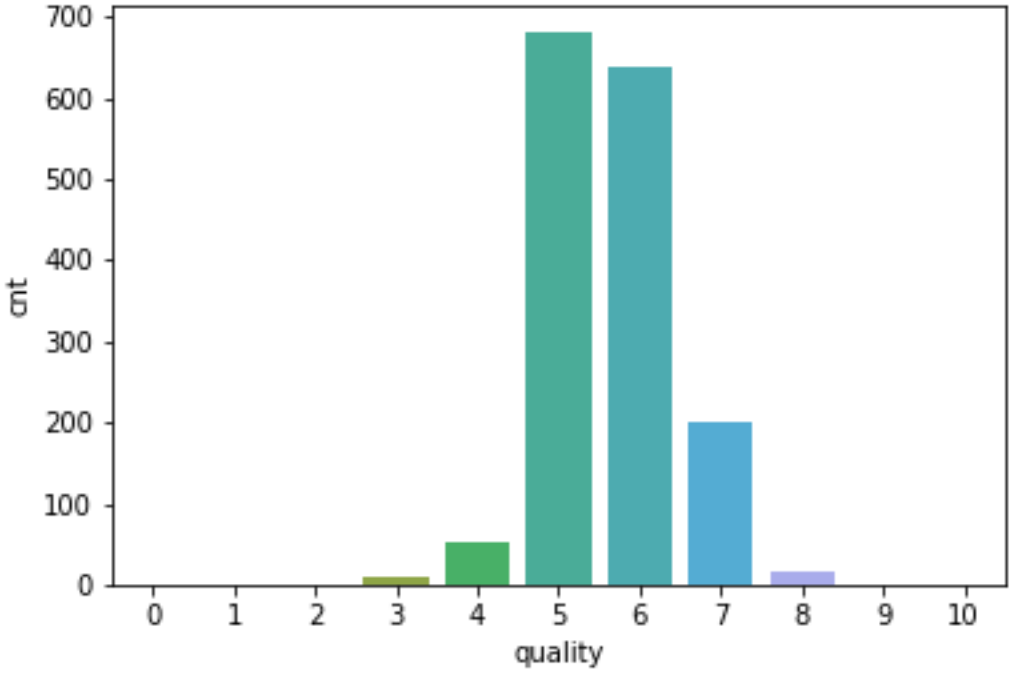

Figure 3 Quantitative Statistics for Different Quality

Table 1: Labels for Alcohol Content of Wines

Label	Quality (p)
Bad	3 <= p <= 4
Middle	5 <= p <= 6
Good	7 <= p <= 8

2.1.3 Dimensionality Reduction

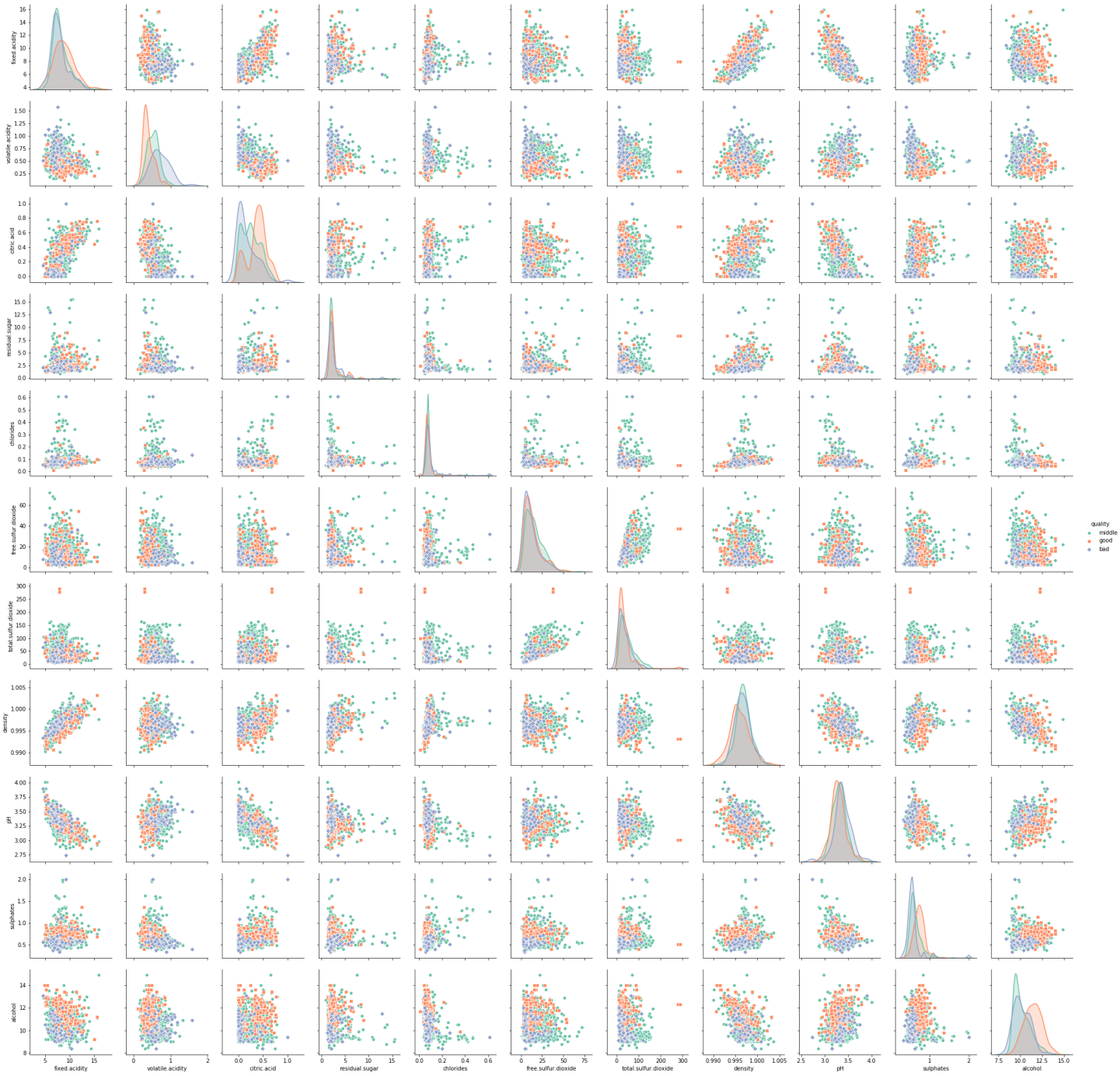

Figure 4 A pairwise plot showing the values of the standardized data

After normalizing the row data, Figure.3 shows a grid of axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis

across a single column. We can roughly see the correlation between any two attributes. For example, there is a strong correlation between **citric acid** with **fixed acidity**, **density** and **citric acid**, **free sulfur dioxide** and **total sulfur dioxide**. Medium quality wine account for the vast majority scattering in various parts of space. Good or bad quality wine well gather together.

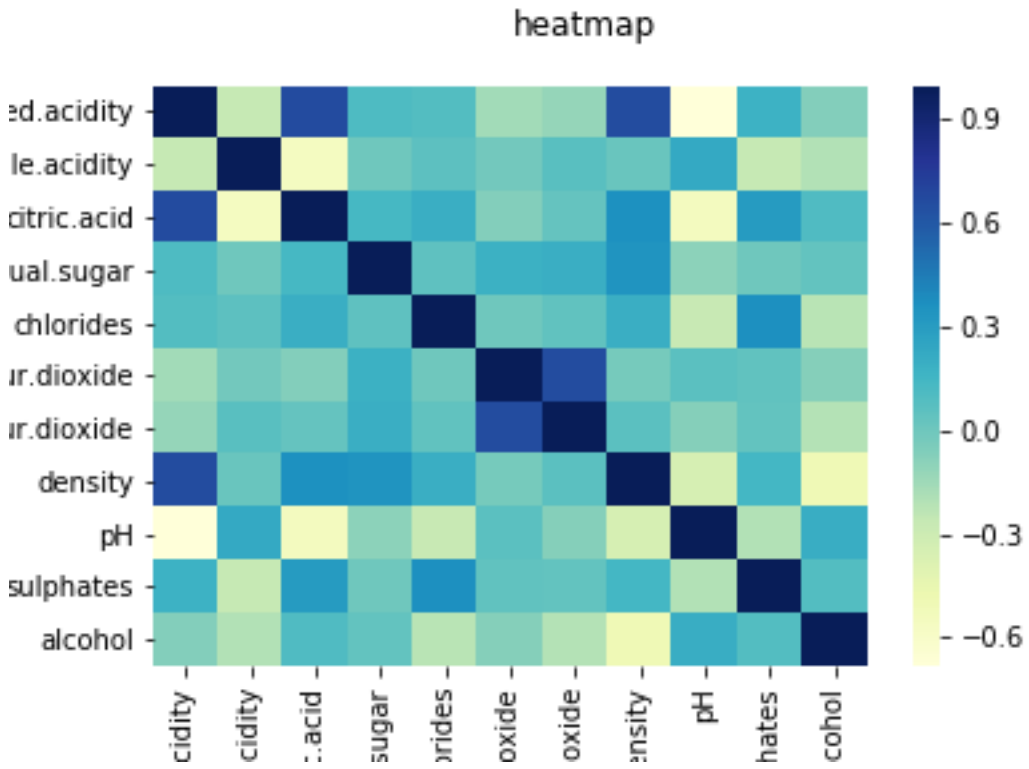

Figure 5 A heatmap showing the values of the covariance matrix

The correlation between any two attributes also can be showed using heatmap of covariance matrix. The darker the color, the stronger the correlation. The diagonal Axes are variances of each attributes, which equals to 1.

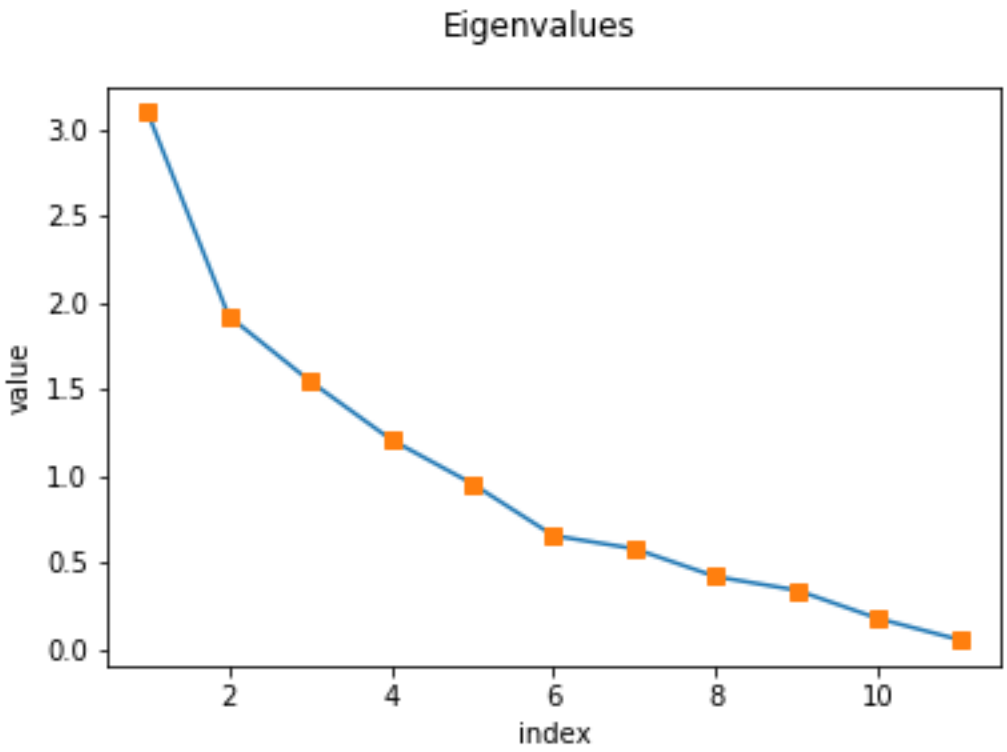

Figure 6 Eigenvalues of covariance matrix

Calculating the eigenvalues of Co-variance matrix and sorting them by big to small, so that transforming a high dimensional vector space into a low dimensional space with storing enough data information (approximately 80%).

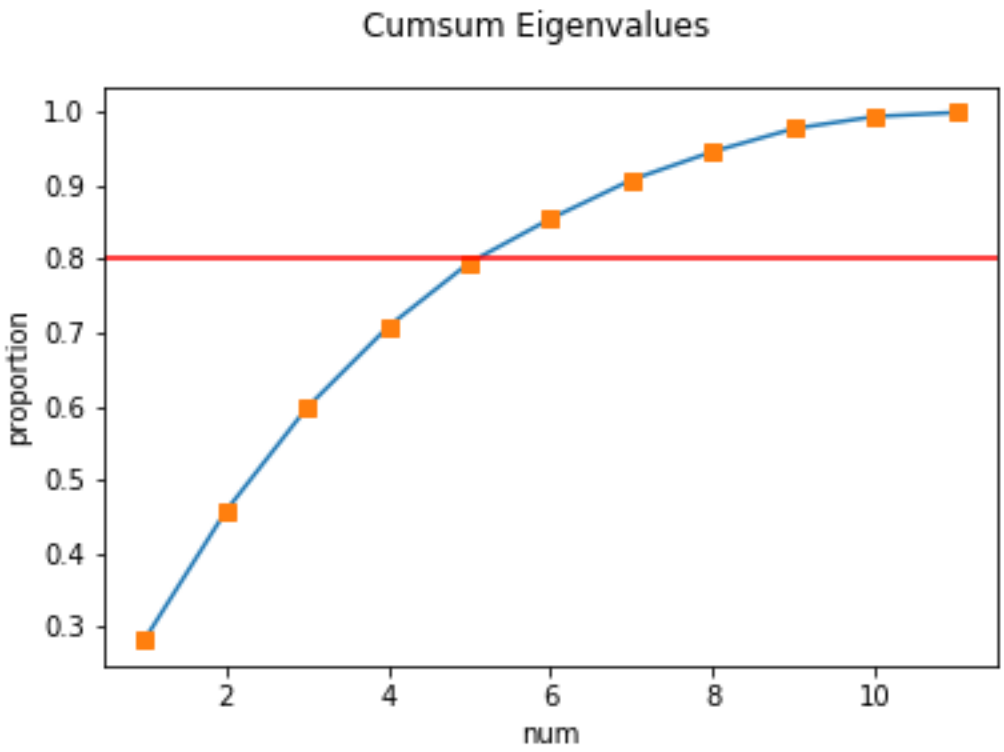

Figure 7 A cumulative total of eigenvalues

	0	1	2	3	4
0	0.489314	-0.110503	-0.123302	-0.229617	-0.082614
1	-0.238584	0.274930	-0.449963	0.078960	0.218735
2	0.463632	-0.151791	0.238247	-0.079418	-0.058573
3	0.146107	0.272080	0.101283	-0.372793	0.732144
4	0.212247	0.148052	-0.092614	0.666195	0.246501
5	-0.036158	0.513567	0.428793	-0.043538	-0.159152
6	0.023575	0.569487	0.322415	-0.034577	-0.222465
7	0.395353	0.233575	-0.338871	-0.174500	0.157077
8	-0.438520	0.006711	0.057697	-0.003788	0.267530
9	0.242921	-0.037554	0.279786	0.550872	0.225962
10	-0.113232	-0.386181	0.471673	-0.122181	0.350681

Figure 8 Eigenvectors correspond to sorted eigenvalues

From the Figure.6, it is able to observe that picking top 5 of those eigenvalues can preserve a proportion of about 80 percent of variability. And the eigenvectors correspond to the sorted eigenvalues. Sorting each eigen-vectors by absolute value of eleven attributes, the greater the absolute value, the greater the impact on this direction.

2.1.4 Projection and Analysis

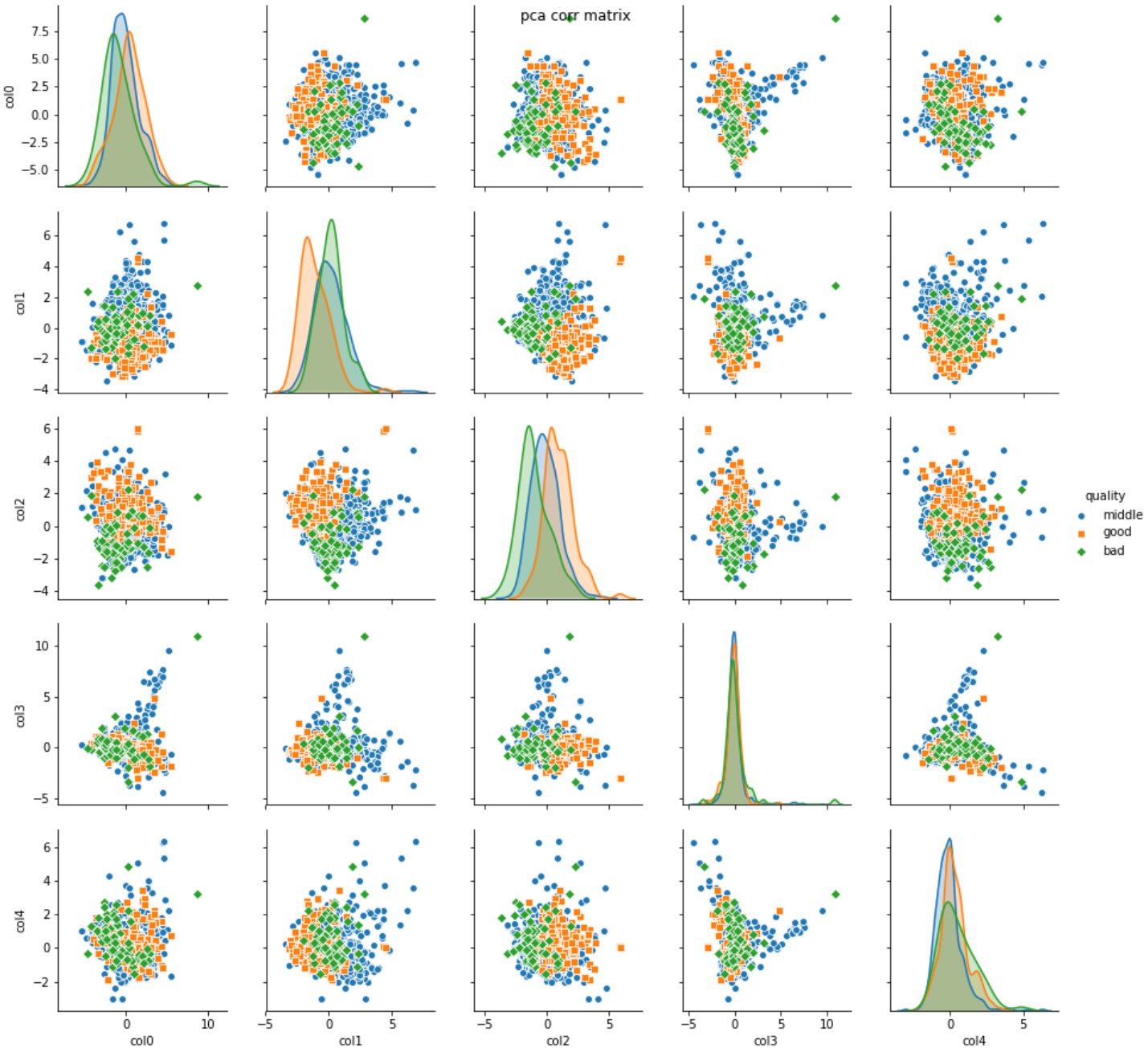

Figure 9 A 5D projection of the PCA results

The diagonal draws a plot to show the univariate distribution of the data for the variable in that column. As the data is labeled, it is easy to see different quality wines are gathered under each single variable distribution, without obvious

difference. But in col0, col1 and col2, they have slight different. Therefore, even using PCA technology to reduce data dimensions, it is hard to distinguish wine quality by a small portion of the properties of the wine. Therefore, I think the quality of the wine is determined by the combination of the attributes of the wine.

2.2 Clustering

2.2.1 How Many Codebook Vectors Do We Need

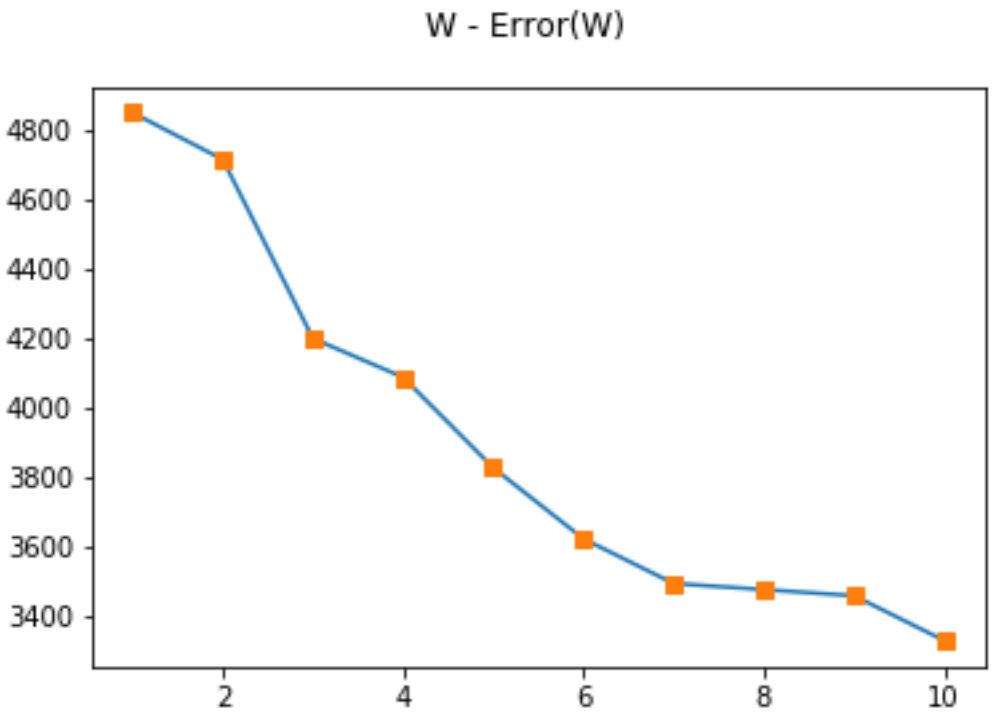

Figure 10 Error E(M) of different M

To find the optimal number of codebook vectors, I investigate the quantization error E(M) as M varies when we employ the algorithm with the same parameter (tau, eta, etc.). The Figure.9 plots the smallest quantization error for each value M, and we can identify the “knee” (M^*， $E(M^*)$) in the plot (M, E(M)) is (3, 4198.75).

2.2.2 Clustering Result Analysis

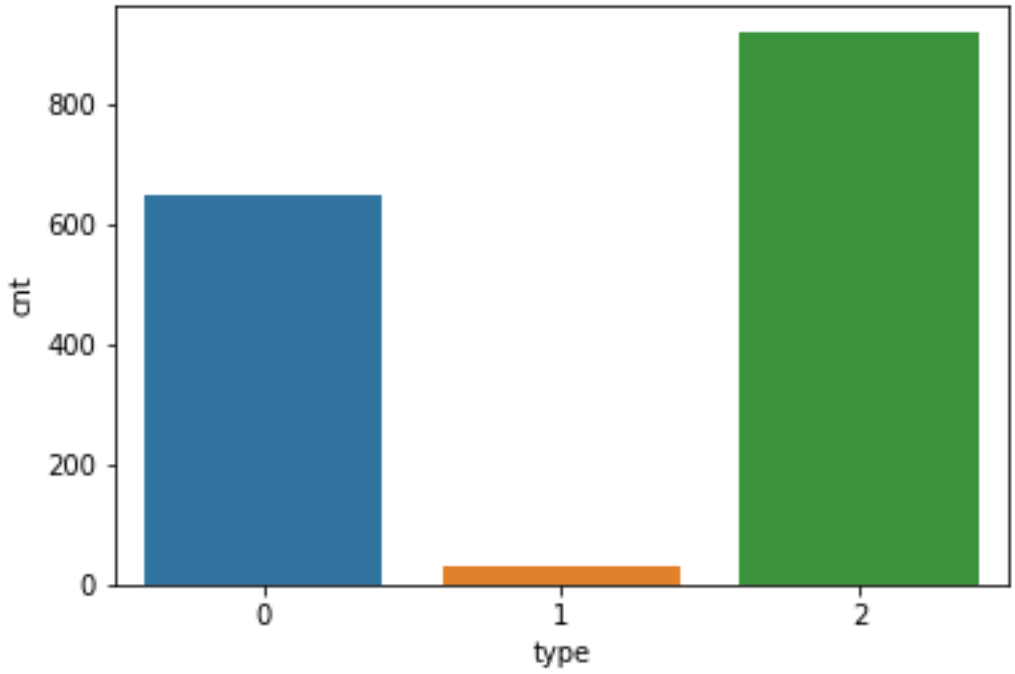

Figure 11 Different Types of Quantities after Clustering

According to Figure.10, I decide to set M = 3 (the number of codebook vectors) which is the ‘knee’ in plot (M, E(M)). After doing another cluster, it counts the number of each type. It is obvious that the number of type [1] is far less than other two types.

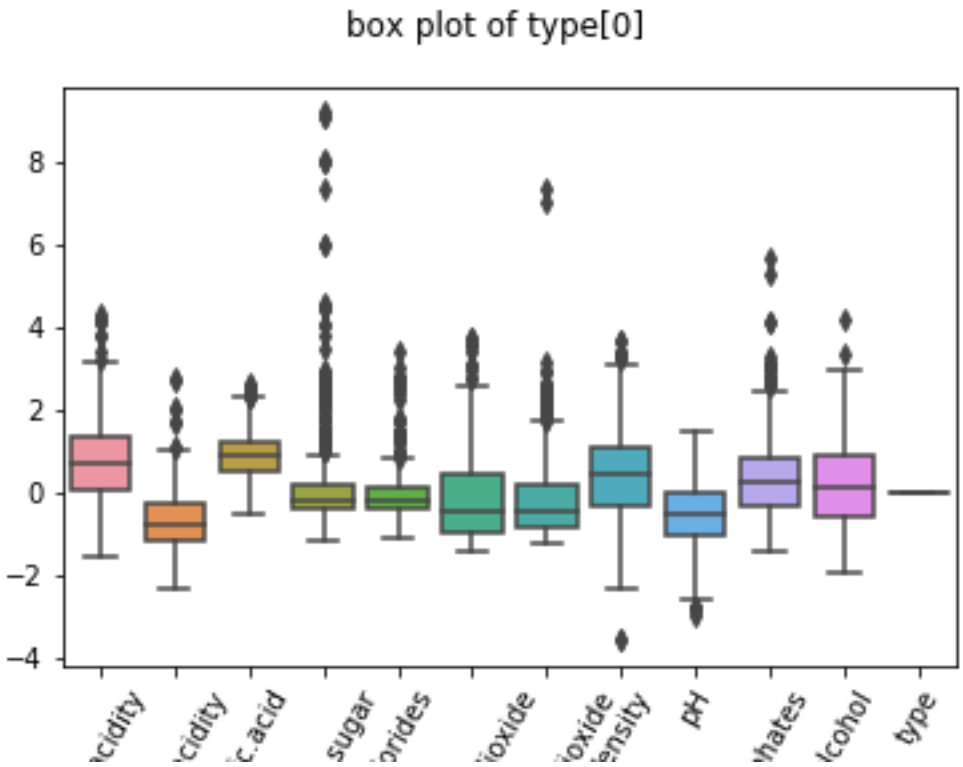

Figure 12 Box Plot of Type [0]

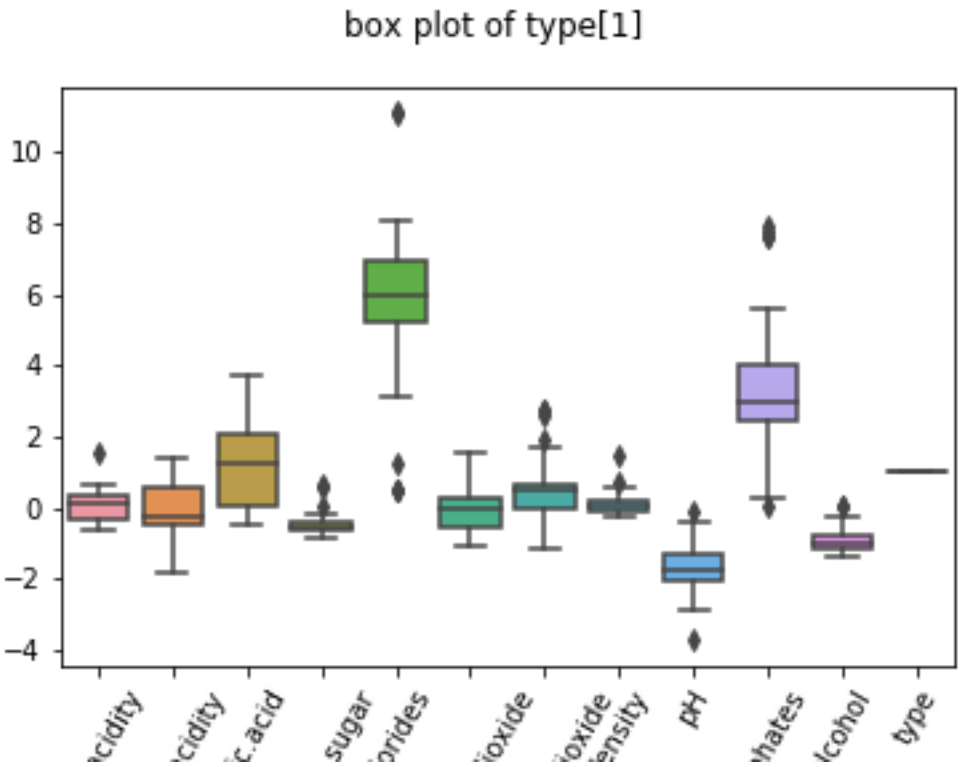

Figure 13 Box Plot of Type [1]

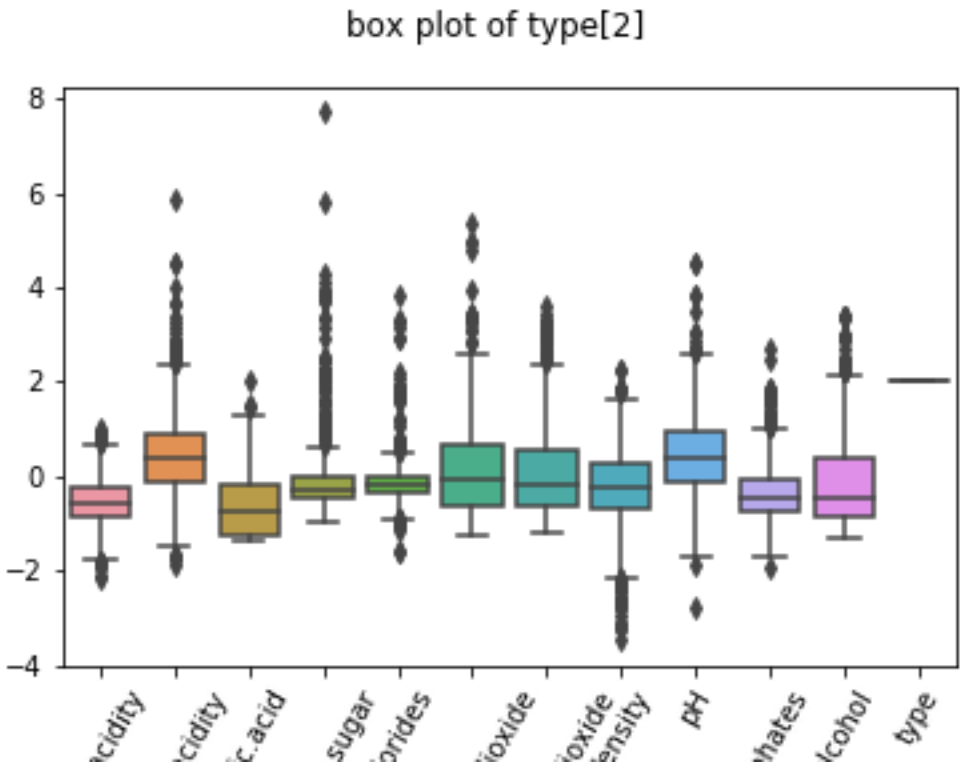

Figure 14 Box Plot of Type [2]

I plot the figure of each type of data. From box plots of three types data, it is obvious that type [1] includes the tightest points in data space according to its less outliers, and the values of **chlorides** and **sulphates** attributes of data are large. It only contains 29 data. The difference between type [0] and type [2] is reflected in **fixed acidity**, **volatile acidity**, **citric acid**, **density** and **pH** by means and outliers.

Comparing to PCA approach, the clustering method is mainly affected by the distance in the data space, even they are standardized. Therefore, aggregated class after applying clustering cannot distinguish the quality of the wine well.

2.3 Self-organizing Map

2.3.1 Run SOM on Model Data

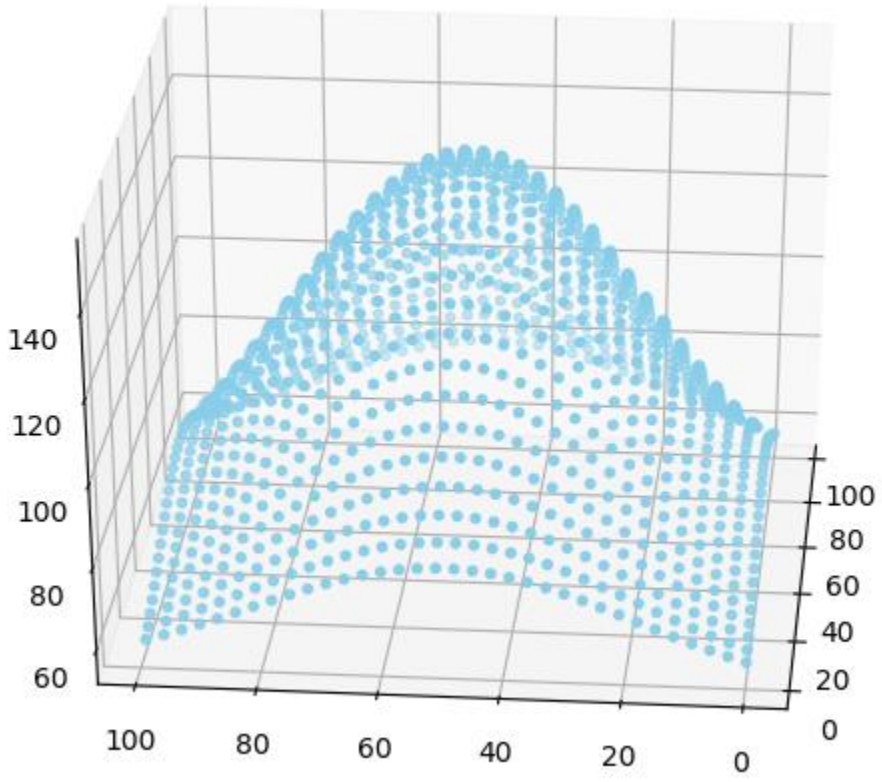

Figure 15 Model Data

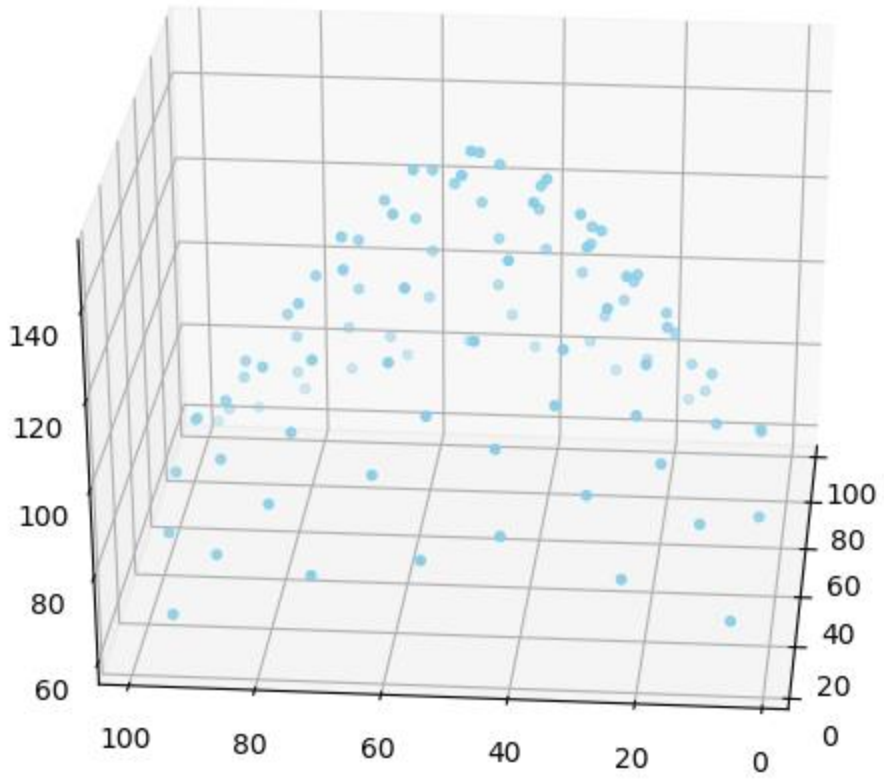

Figure 16 Running SOM on Model Data

Before running SOM on my own row data, I try to run SOM on a 3D data in order to test my SOM algorithm correctness by visualization. Figure.16 is showing 10*10 grid that fits Figure.15.

2.3.2 Run SOM on Wine Data

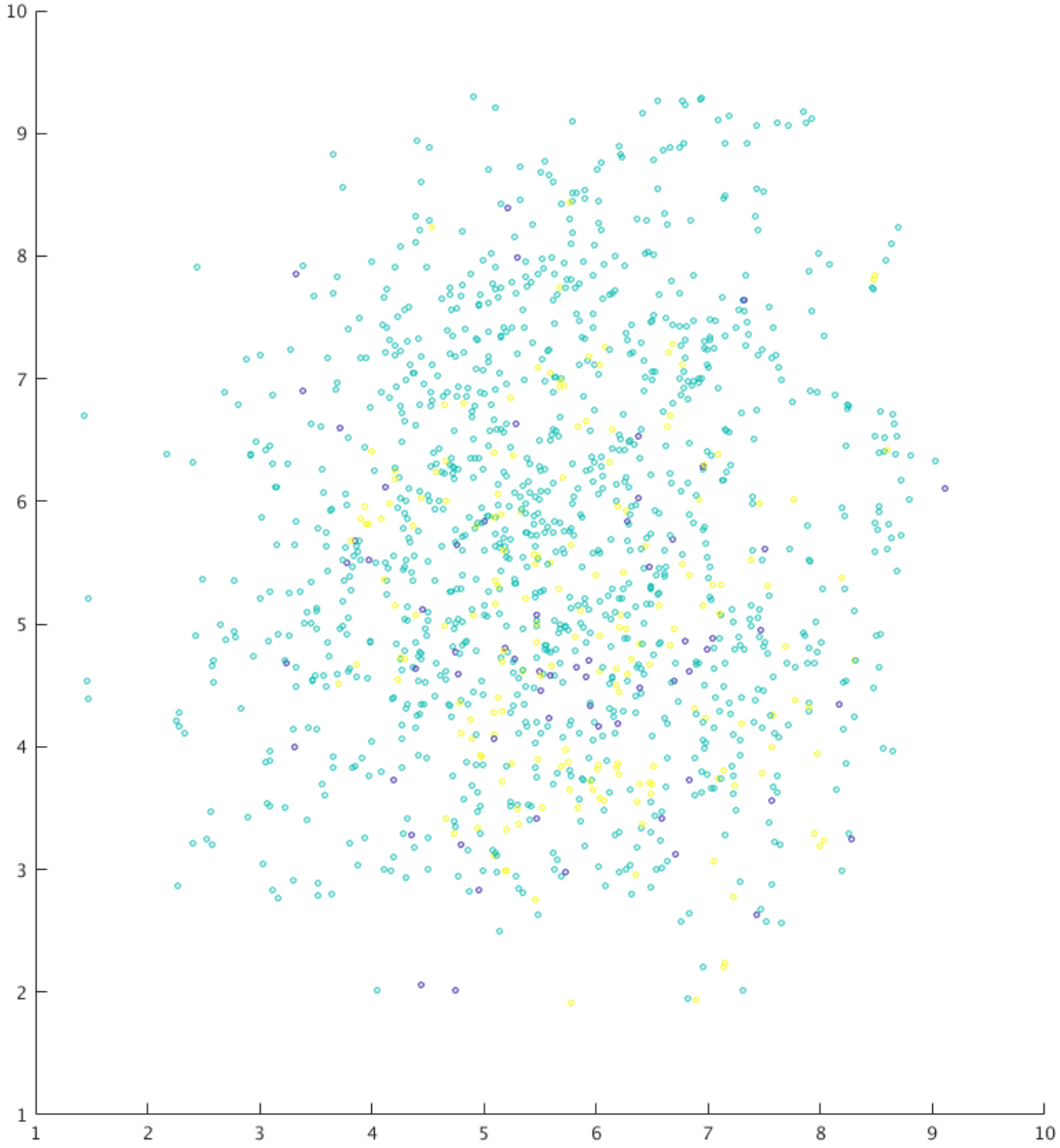

Figure 17 Running SOM on Wine Data

After wine data project on SOM 2D plane, it shows that the projected data did not cluster well by SOM, but data are well clustered in fact as what I show in

PCA and clustering parts. After analysis, I think maybe wine data cannot simply be fitted in non-linear two dimensions. The data in fact is in a high dimensional space, so the projection of three types of quantity of wine is depressed and loose.

2.4 Conclusion

After applying data analysis techniques including PCA, clustering and SOM, I cannot find an appropriate and easy way to distinguish the quality of wine.

Question II: Which wine properties determine the density of the wine

3.1 Principal Component Analysis

3.1.1 Labeling Data

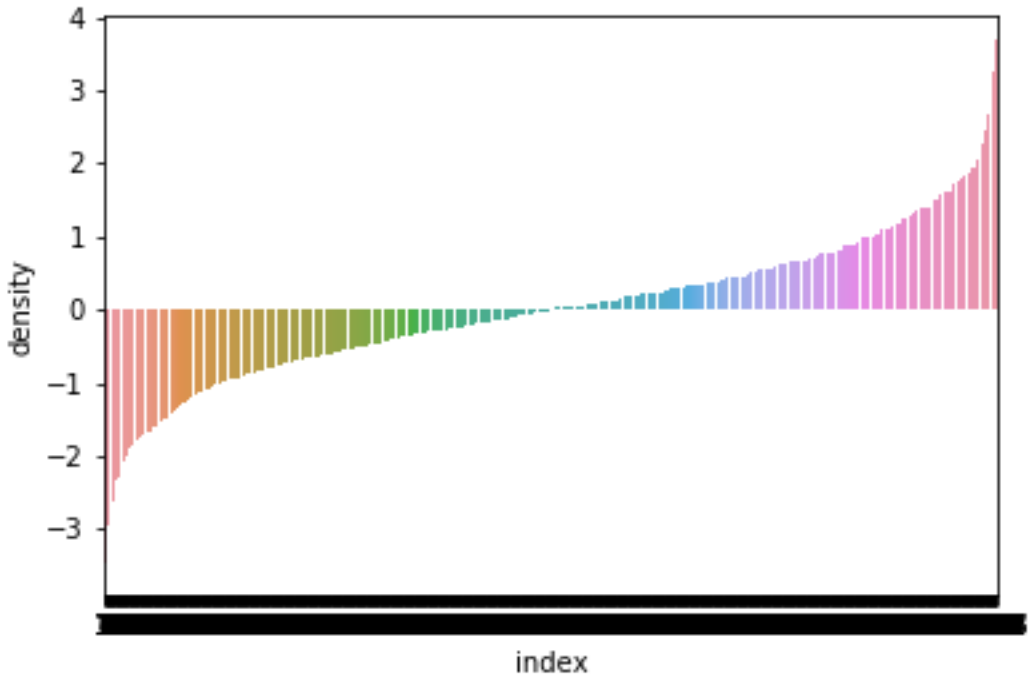

Figure 18 Distribution of Density

Table 2: Labels for Density Content of Wines

Label	Density(d)
Low	d < -0.5
Middle	-0.5 <= d <= 0.5
High	0.5 < d

3.1.2 Dimensionality Reduction

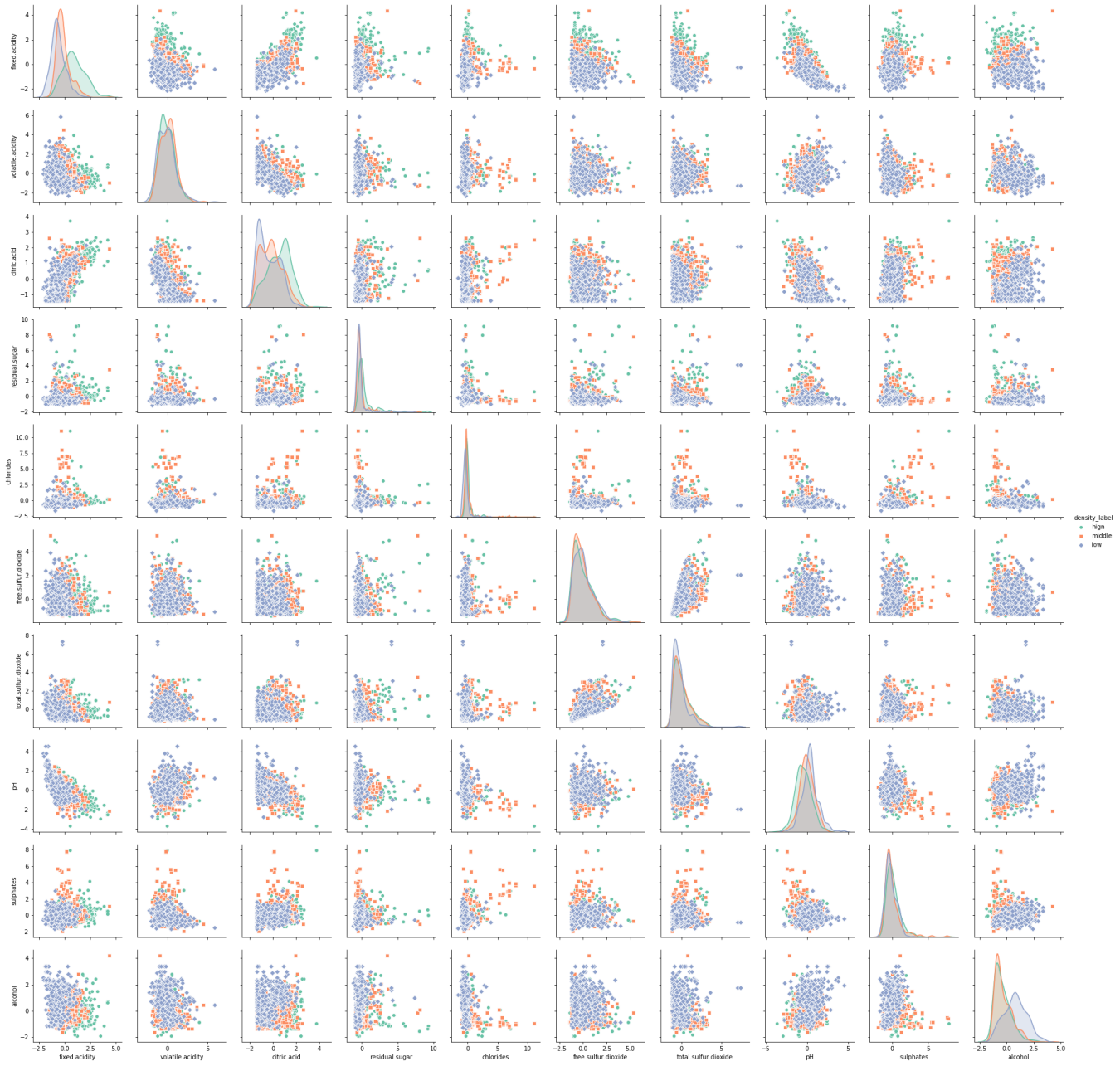

Figure 18 A pairwise plot showing the values of the standardized data

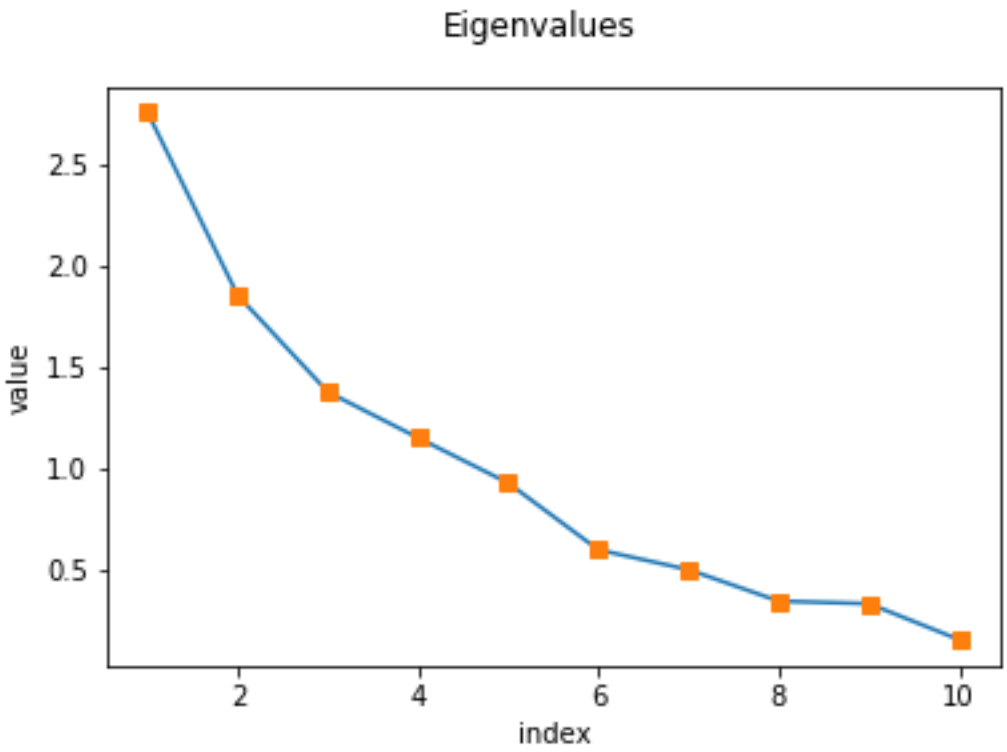

Figure 20 Eigenvalues of covariance matrix

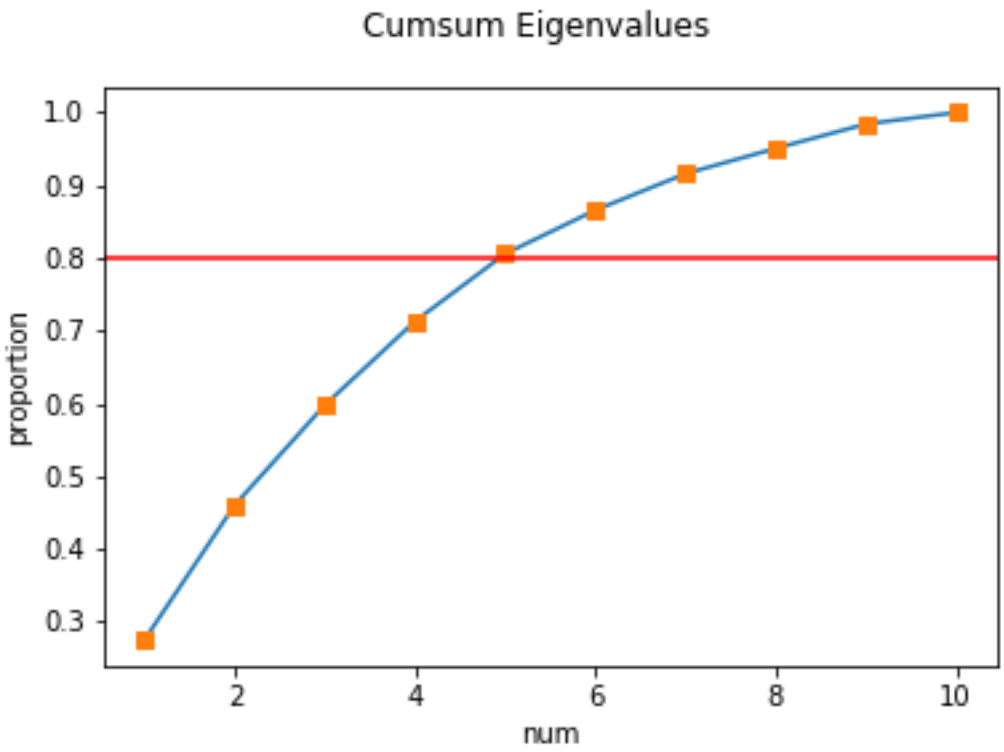

Figure 21 A cumulative total of eigenvalues

Similarly, calculating the eigenvalues of co-variance matrix and sorting them by big to small. From Figure.20, I decide to pick top 5 eigenvalues with their eigenvectors.

	0	1	2	3	4
0	0.491120	-0.075992	-0.101647	-0.328083	-0.013380
1	-0.337379	0.109292	-0.454117	-0.077904	-0.321277
2	0.527330	-0.008585	0.170949	-0.063655	0.034481
3	0.096588	0.282069	0.188839	-0.250064	-0.824106
4	0.212124	0.187917	-0.428829	0.521640	-0.254375
5	-0.065624	0.603236	0.279005	0.018201	0.193290
6	-0.018215	0.645658	0.125922	-0.070397	0.190667
7	-0.472160	-0.076362	0.245468	0.238973	-0.080283
8	0.291508	0.087838	0.064323	0.670863	-0.032495
9	-0.004818	-0.270217	0.613658	0.187476	-0.265218

Figure 22 Eigenvectors correspond to sorted eigenvalues

3.1.3 Projection and Analysis

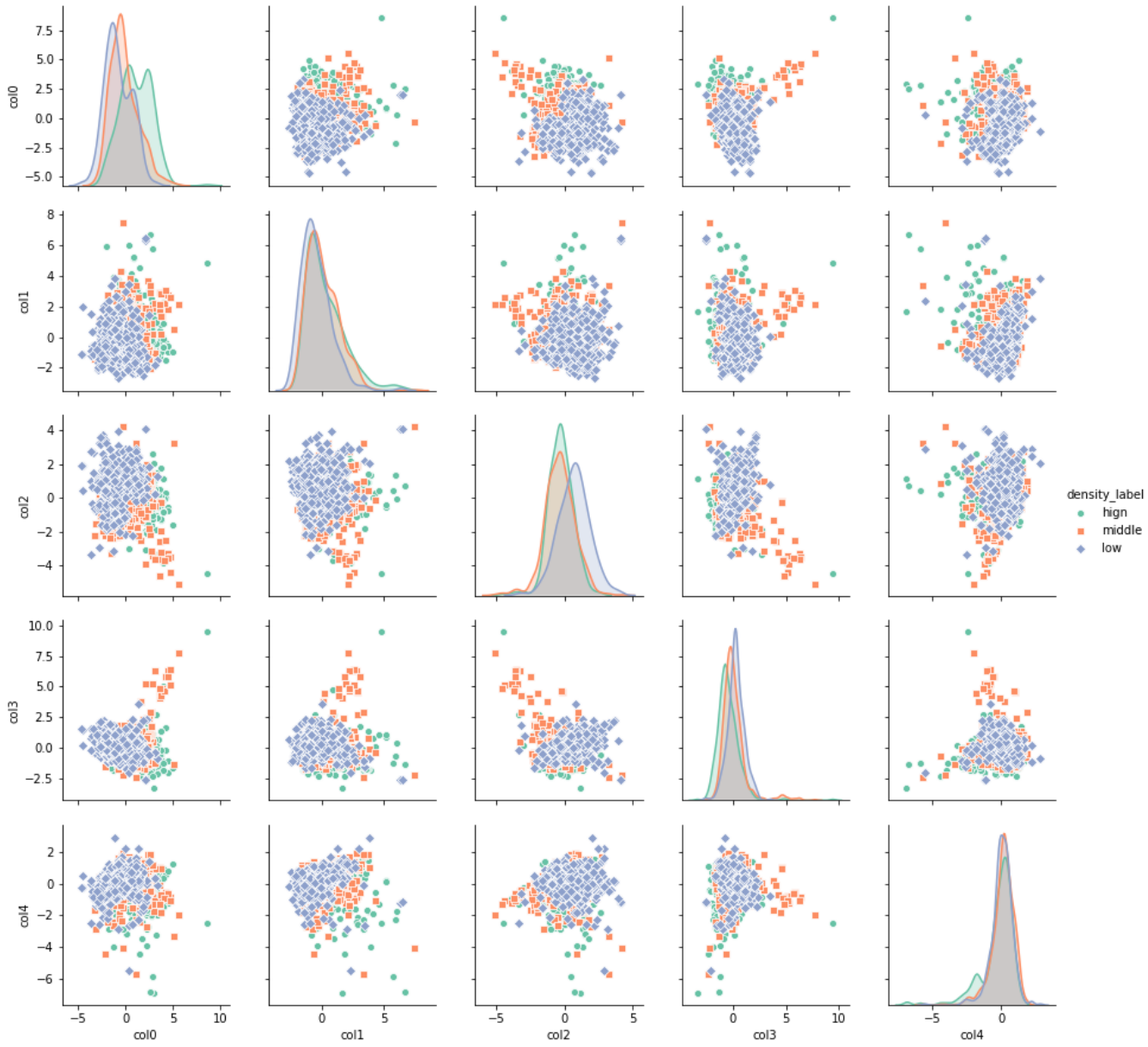

Figure 23 A 5D projection of the PCA results

The distribution of data is the same as last question. Therefore, it is hard to distinguish wine desity by a small portion of the properties of the wine.