Investigating Properties of Forest Fires from a 13-Dimensional Data Set

An Assignment Report of
CS212
Intelligent Data Analysis

BY

Xiangyi Yan

11510706@mail.sustc.edu.cn

UNDER THE GUIDANCE OF

Peter Tiňo

AND

Guoji Fu

Department of Computer Science and Engineering

Southern University of Science and Technology

Shenzhen, China, June 2017

Abstract

The purpose of this assignment was to utilize dimension-reducing techniques such as **principal component analysis (PCA)** and **Self-organizing Maps (SOM)** to explore relationships between varying attributes of **forest fires**. It was found that some of these attributes didnt correlate with the rest of the data, though when reduced enough, certain attributes were found to hold relationships.

Code & Notable Files:

`https://github.com/yanxiangyi/forest_fire_pca`

Data Set Introduction

13 Dimensions

- X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
- Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
- month - Month of the year: 'jan' to 'dec'
- day - Day of the week: 'mon' to 'sun'
- FFMC - Fine fuel moisture code index from the FWI system: 18.7 to 96.20
- DMC - Duff moisture code index from the FWI system: 1.1 to 291.3
- DC - Drought code index from the FWI system: 7.9 to 860.6
- ISI - Initial spread index from the FWI system: 0.0 to 56.10
- temp - Temperature in Celsius degrees: 2.2 to 33.30
- RH - Relative humidity in %: 15.0 to 100
- wind - Wind speed in km/h: 0.40 to 9.40
- rain - Outside rain in mm/m2 : 0.0 to 6.4
- area - The burned area of the forest (in ha): 0.00 to 1090.84

Data Set Source:

`https://archive.ics.uci.edu/ml/datasets/Forest+Fires`

Preprocessing

Replacement

The **month** and **day** dimensions contain text elements. The first step of data preprocessing is to replace text elements with corresponding integers.

Normalization

It's important to shift every attributes of the data to be centered:

$$E[X_i] = 0, i = 1, 2, ...d.$$

It's also important to scale the density of every attributes of the data so the standard deviation (or variance) need to be rescaled:

$$Var[X_i] = 1, i = 1, 2, ...d.$$

Now every attribute is standardized to $\frac{X_i - \mu}{\sigma}$.

Drop Special Features

After viewing the whole **rain** attribute, I find 99.6% (515/517) forest fires happened with no rain. Which is to say, only in 2 of 517 cases, forest fires happened while raining. We need to observe this feature independently, instead of using data analysis methods such as PCA or SOM. Therefore, the **rain** attribute is considered as a special feature and is decided to be dropped.

Research Questions

For my project, I tried to focus on the following two research questions:

How does the rest of the dimensions influence:

- How does the area of the forest fires influenced by all the forest fires index?
- Does the initial spread index of the forest fires influenced by any other features?

Chapter 1

Area

1.1 Labelling

According to **McArthur Forest Fire Danger Index**, forest fires can be labelled by area as:

Fire danger rating

Category	Fire Danger Index	
	Forest	Grassland
Catastrophic (Code Red)	100 +	150 +
Extreme	75–99	100–149
Severe	50–74	50–99
Very High	25–49	25–49
High	12–24	12–24
Low–Moderate	0–11	0–11

Figure 1.1.1: Fire danger rating

1.1.1 Labelling Strategy 1

For sake of lacking data, forest fires are labelled as three following classes (in ha):

- Class 1: No Fire: 0.
- Class 2: Low-Moderate: (0, 11].
- Class 3: High-Catastrophic: (11, ∞)

The final labelling result is:

- Class 1: 247 cases.
- Class 2: 183 cases.
- Class 3: 87 cases.

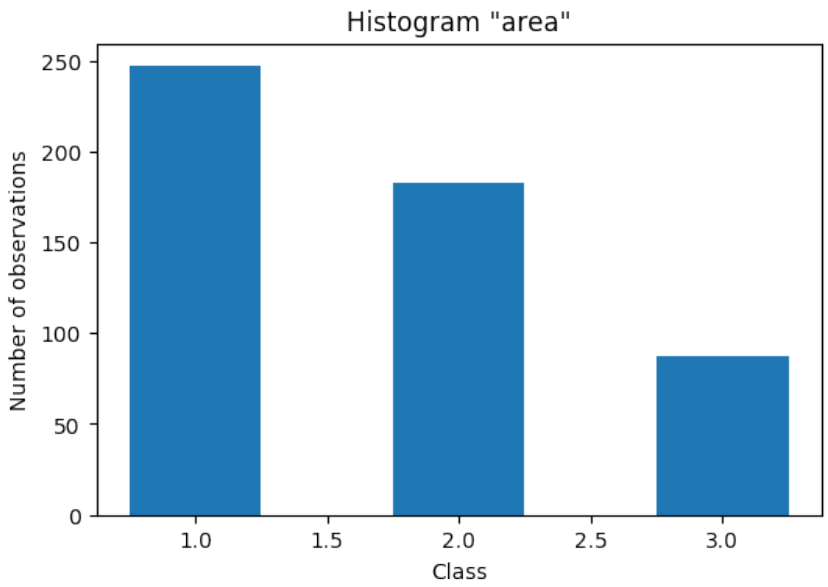

Figure 1.1.2: Histogram of area (Labelling Strategy 1)

1.1.2 Labelling Strategy 2

Labelling strategy 2 aims to soften the restriction of the former one. I just put class 2 and class 3 together, which means there are only 2 classes:

- Class 1: No Fire: 0.
- Class 2: On fire: $(0, \infty)$

The final labelling result is:

- Class 1: 247 cases.
- Class 2: 270 cases.

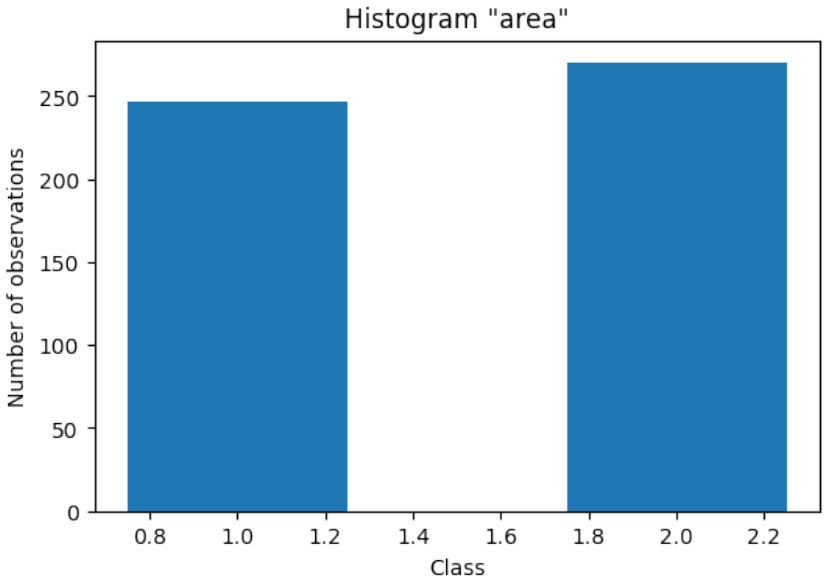

Figure 1.1.3: Histogram of area (Labeling Strategy 2)

1.2 Straightforward Co-ordinate Projections

In this section, let's just simply straightforward project the standardized data on a 2 dimensional plain. In **Figure 1.2.1**, the standardized data was projected on the **FFMC** and **wind** plain. In **Figure 1.2.2**, the standardized data was projected on the **FFMC** and **DMC** plain.

It's obviously not a good idea to do so, because we lost lots of information of the dropped dimensions. There's no apparent relationship that can be found in the following two figures.

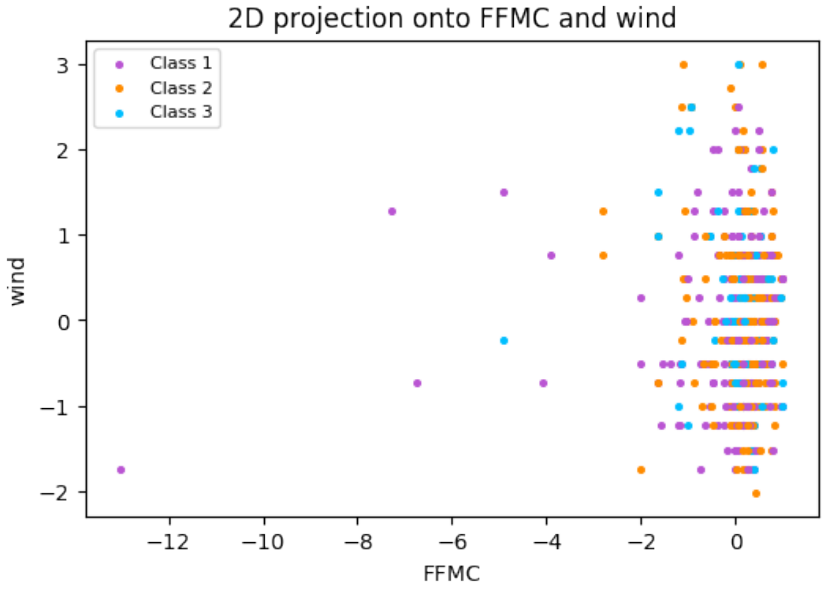

Figure 1.2.1: Projection on FFMC and wind

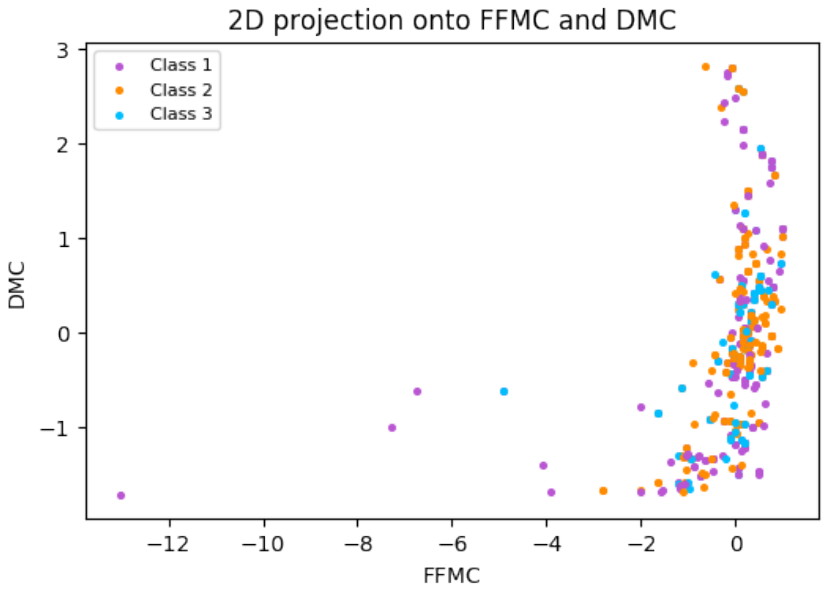

Figure 1.2.2: Projection on FFMC and DMC

1.3 Visualization

1.3.1 Principal Component Analysis

After normalization, the co-variance of matrix X can be estimated as:

$$Cov[X] \approx \widehat{Cov[X]} = \frac{1}{N} X X^T$$

In $Cov[X]$, most of the non-diagonal elements are positive numbers, so most of them are positive related, which means most of dimensions move together in the same directions.

After $Cov[X]$ calculated, linalg.eig() function in numpy was called to generate the eigenvalues, eigenvectors of the co-variance matrix.

After SVD decomposition, $\widetilde{Cov[X]}$ can be calculated as a matrix whose diagonal elements are exactly the eigenvalues of $Cov[X]$.

Before results from PCA were projected, I plotted the eigenvalue spectrum and cumulative eigenvalue plot to see how much can the first two principal components be used to represent the original relationship after the PCA projection.

The result is that the sum of first two largest eigenvalues is **0.44034744** and the sum of first three largest eigenvalues is **0.57060657**, which is not very ideal to project the data on a 2-dimensional or 3-dimensional plot, because we will lose about **56%** and **43%** variance of the original data, which will definitely influence the properties of the original data.

That's why in **Figure 1.7** and **Figure 1.8** the plot is a little bit messy. In principle component analysis, there are 11 dimensions used (without **area** being results and **rain** dropped).

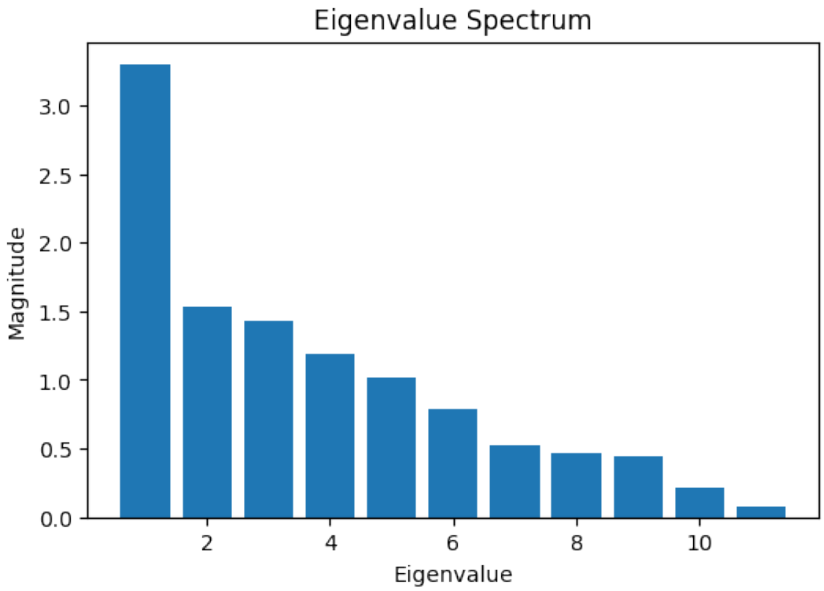

Figure 1.3.1: Eigenvalues of co-variance matrix

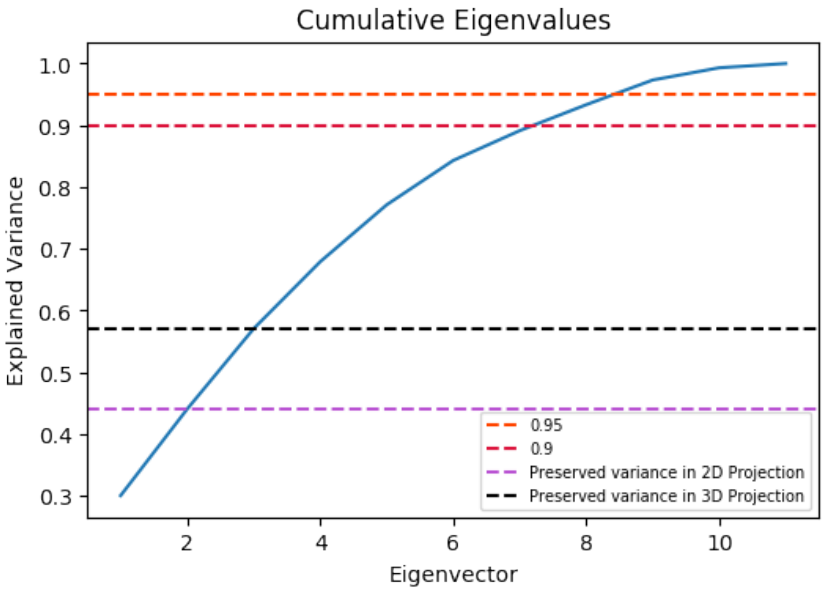

Figure 1.3.2: Cumulative eigenvalues of co-variance matrix

In the following table, I list the value of the first 2 eigenvectors of $Cov[X]$. The first column here shows coefficients of linear combination that defines principal component 1, and the second column shows coefficients for principal component 2. In the new axis, the new coordinates of the data are linear combination of the original coordinates and the coefficients are exactly the values in the table.

If the value is positive, then a higher score on that variable is associated with a higher score on the component, if the value is negative, then a higher score implies a lower score on the component.

In new axis 1, **CD** index is the most important part because the coefficient is **0.46574435**, because the value is the largest. Similarly, **Y** is the most important part of new axis 2.

Also, in the following table, we can find that **day** is the most useless features, because the values of **day** for principle component 1 and principle component 2 are **0.00153541** and **0.0034352**. These values are too low to influence the projection. Therefore, **day** is the most useless dimension of all 11 dimensions.

Top 2 largest eigenvalues and eigenvectors			
1st largest: 3.29706752		2nd largest: 1.53738519	
Dimensions	Values	Dimensions	Values
DC	0.46574435	Y	-0.6927839
temp	0.42160692	X	-0.69267032
month	0.41058638	RH	-0.12190801
DMC	0.40663441	DMC	-0.11283833
FFMC	0.3602159	ISI	-0.09167051
ISI	0.29445686	FFMC	-0.05470367
RH	-0.1802439	temp	-0.02405163
wind	-0.11291306	DC	-0.01721328
X	-0.07257493	month	-0.01517577
Y	-0.06655858	wind	0.01230403
day	0.00153541	day	0.0034352

Table 1.3.1: Top two eigenvectors

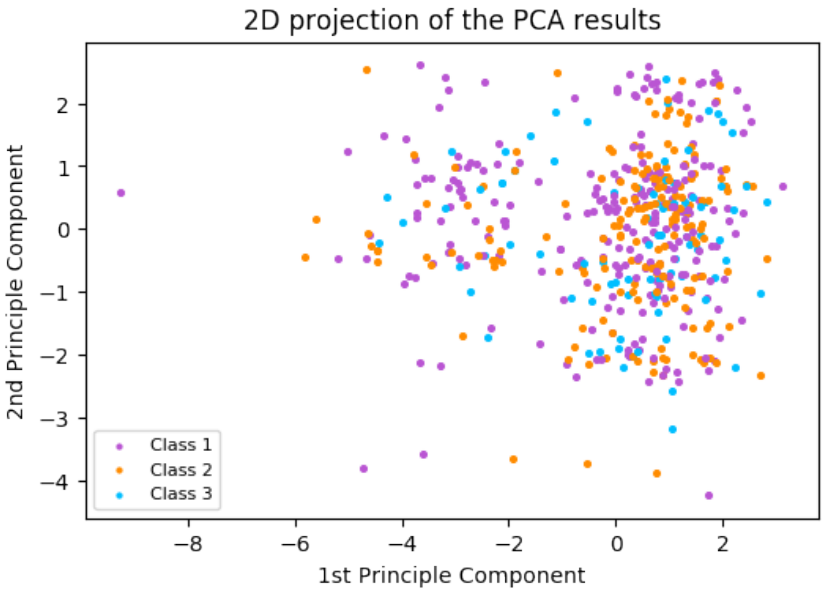

Figure 1.3.3: 2D PCA projection of area

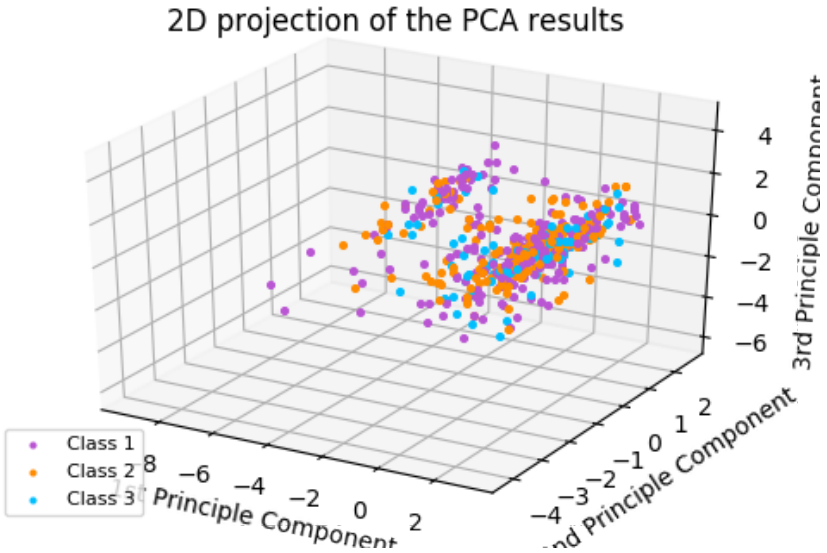

Figure 1.3.4: 3D PCA projection of area

1.3.2 Self-organizing Map

First, I just use SOM to fit the 11 dimensional data straightforwardly.

Obviously, plotting this is going to fool myself, because the 2D fishing net is very possible to bend a lot to fit 11D data set. The SOM would be like totally a mess.

After checking out the normal vectors of the grids in the dark center area, I confirmed my thought was right. Most of the cosine value of the normal vectors are negative values or very small values, which means that the angle between each grid is pretty large and the fishing net really curves a lot.

It's worth to mention that in this and the following SOMs, the darker the background color is, the more the "hole" of the fishing net contracts.

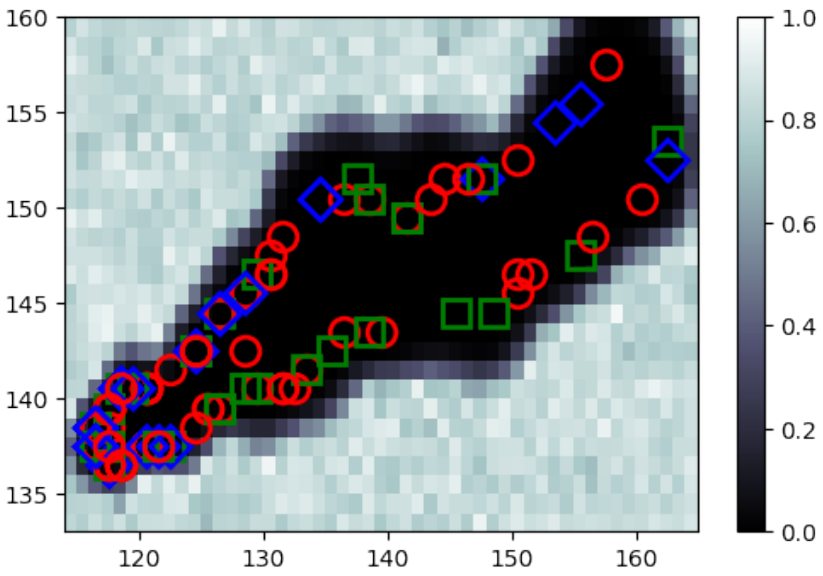

Figure 1.3.5: Direct SOM of 11 Dimensional data

Next, I used PCA to make the data reduce to 3 dimensions and used label strategy 2.

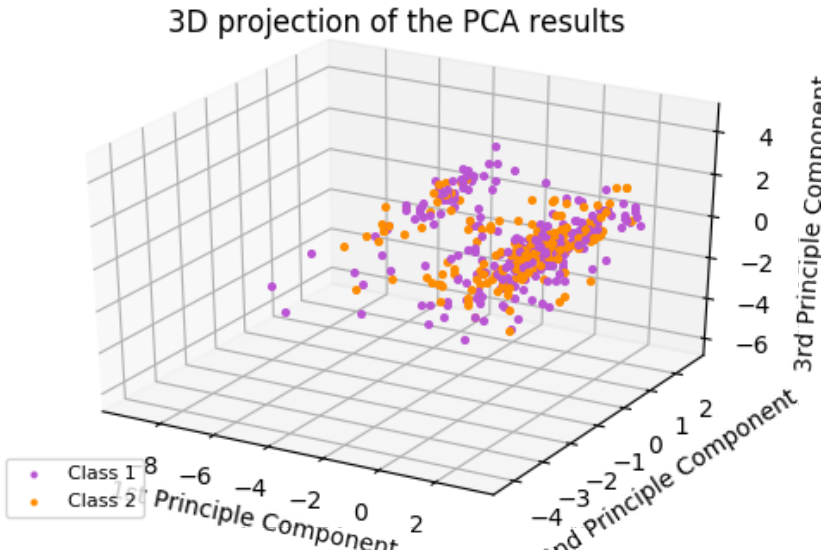

Figure 1.3.6: 3D PCA projection of area

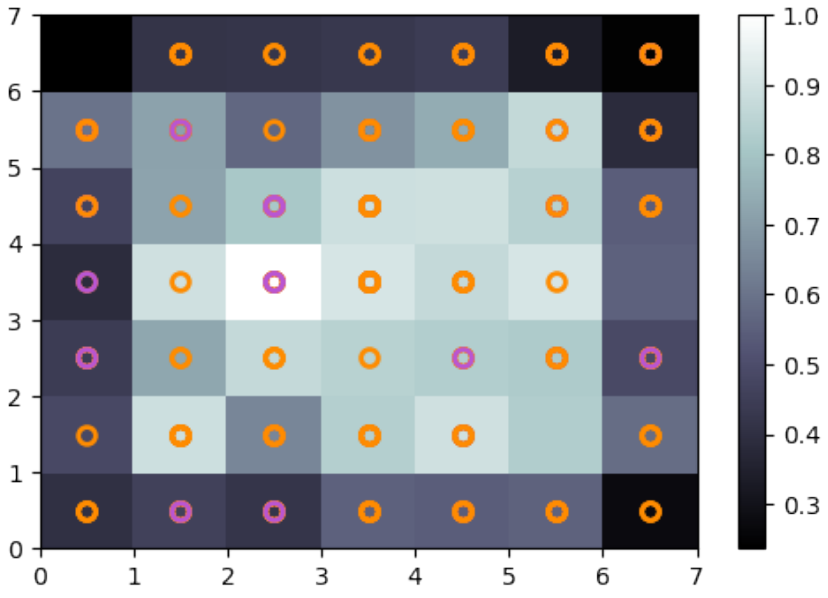

Figure 1.3.7: SOM of 3D PCA data

We can find that the SOM is tightly contracted on the edges, and a little bit loose in the center area.

Next, I firstly used PCA to make the data reduce to 3 dimensions and then did a 3-means clustering. After that, I plot a SOM for this 3D data.

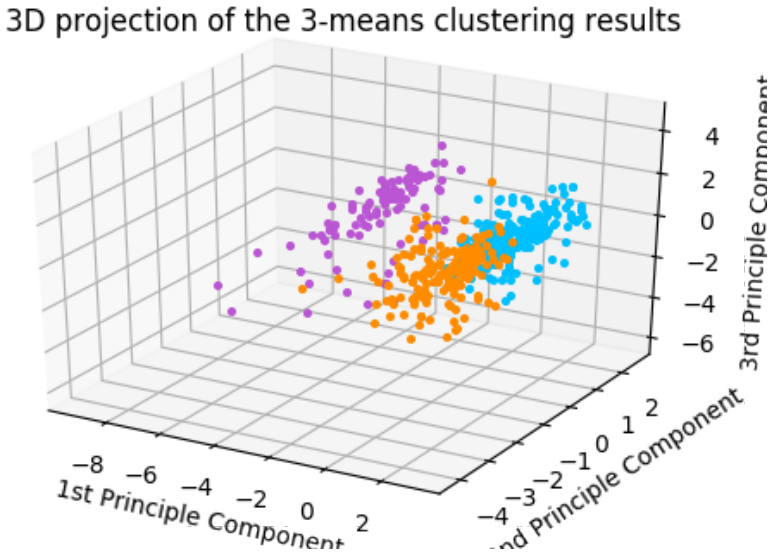

Figure 1.3.8: 3D Clustering of area

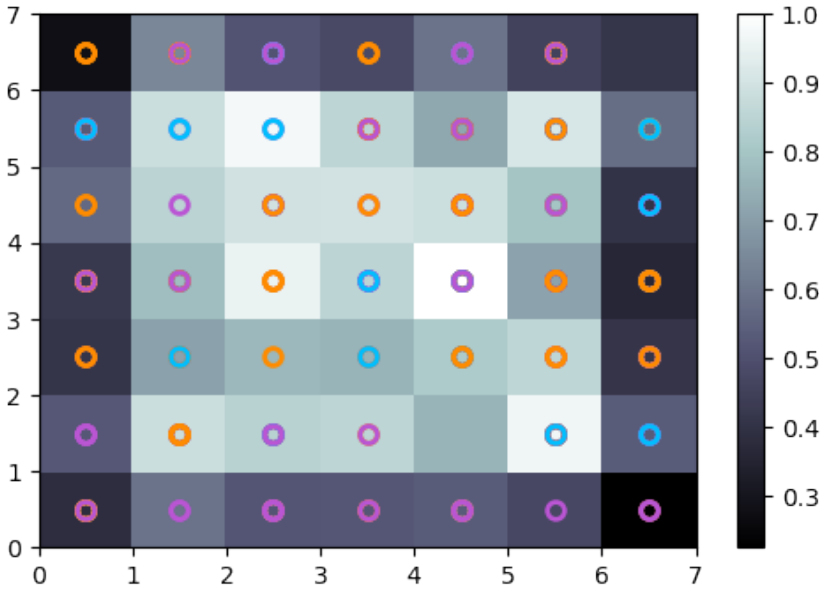

Figure 1.3.9: SOM of 3D clustering data

We can find that the SOM is tightly contracted on the edges, and a little bit loose in the center area.

1.3.3 Comparing PCA and SOM

After doing the above manipulation of the data, my conclusion is that:

- If the data is in a high dimension, don't use SOM. Because that will definitely fool yourself. And in this condition, PCA helps a lot to visualize the data.
- Try to combine these 2 techniques: do PCA first to reduce the dimension to an acceptable space(dimension) and then apply SOM to make the result of PCA easier to understand.

1.4 Clustering

1.4.1 Clustering After PCA

After visualization in top 3 principle components axis, we can obviously find that the data has about 2 clusters.

When I plotted the quantization error figure, the knee point is at K = 2, which means there should be 2 clusters and this can prove our former visual guess.

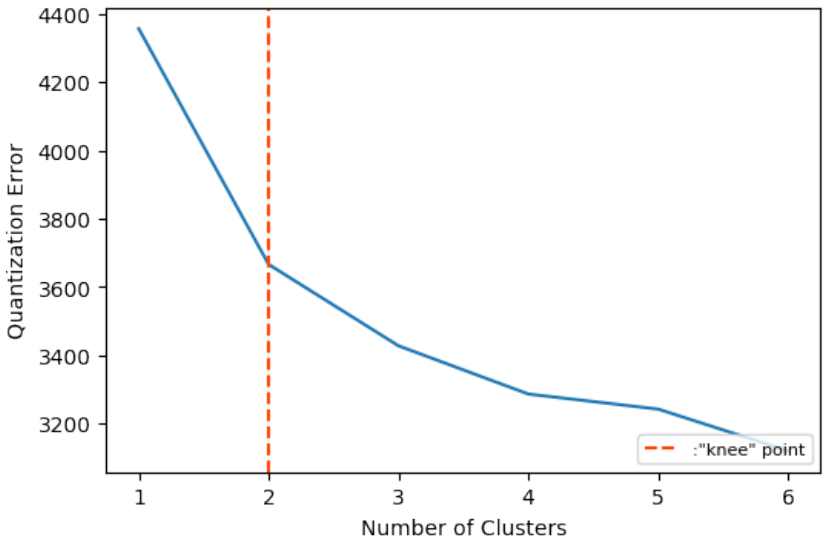

Figure 1.4.1: Quantization Error

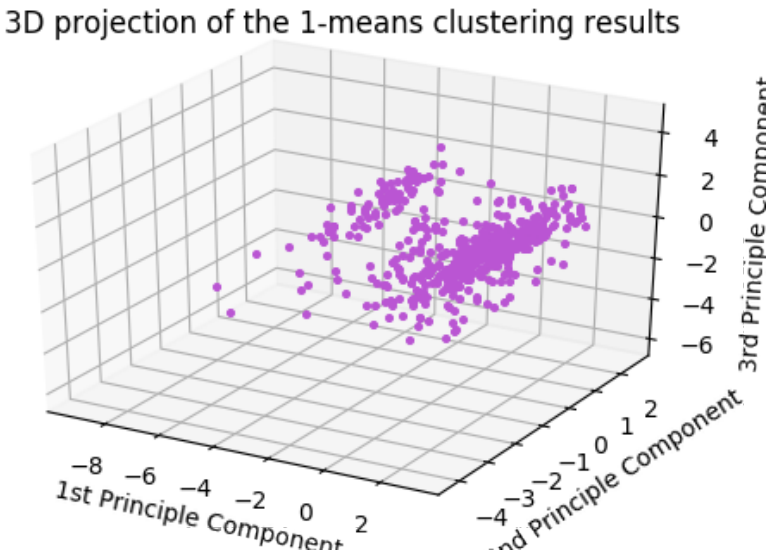

Figure 1.4.2: 1-Means Clustering

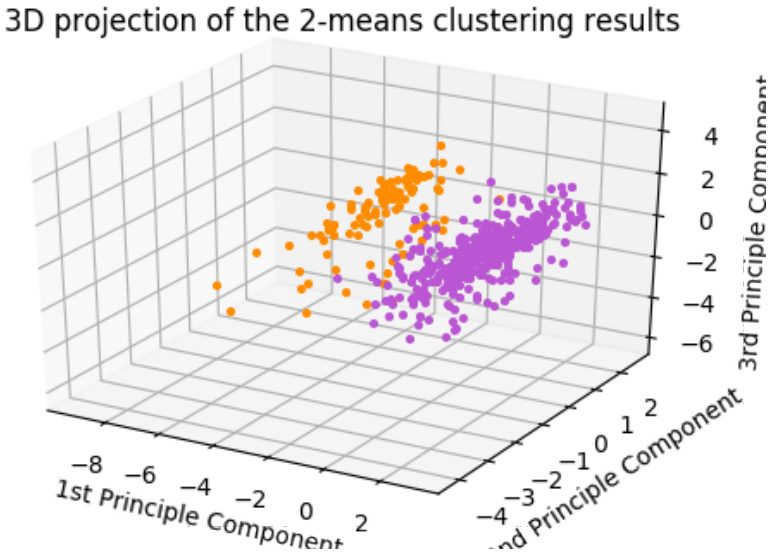

Figure 1.4.3: 2-Means Clustering

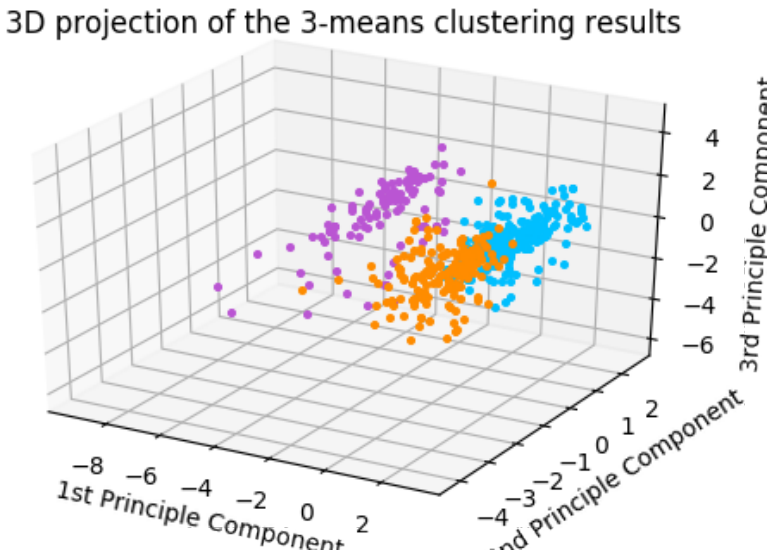

Figure 1.4.4: 3-Means Clustering

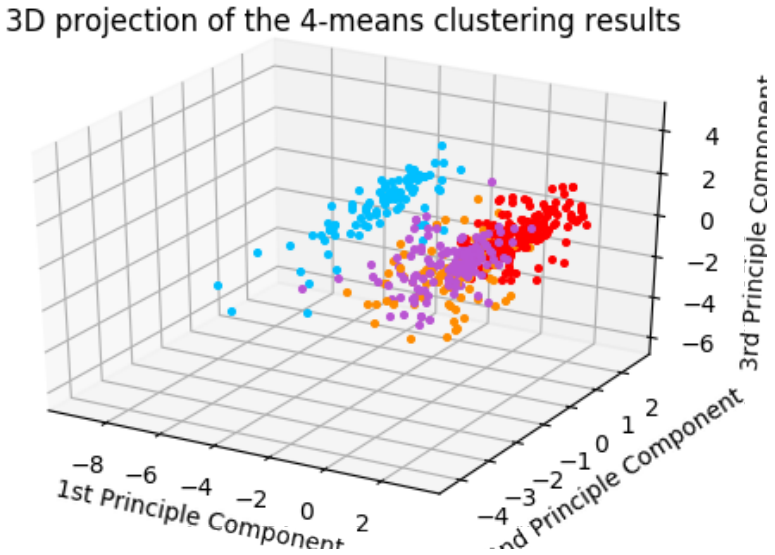

Figure 1.4.5: 4-Means Clustering

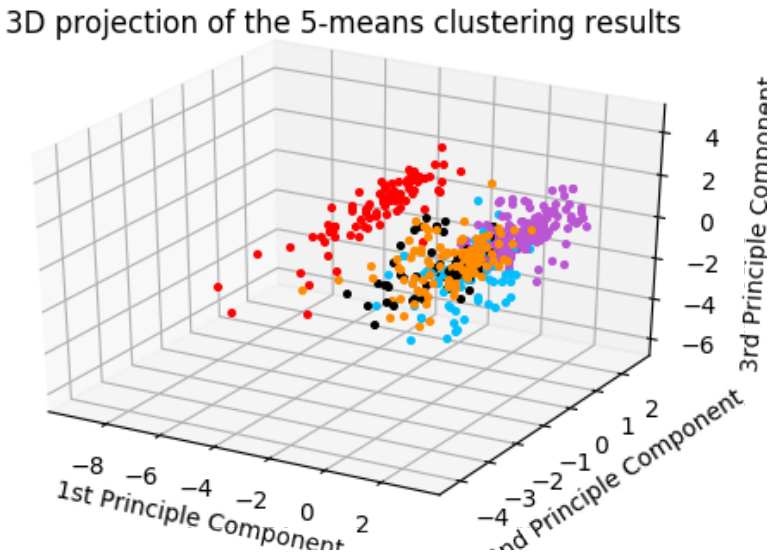

Figure 1.4.6: 5-Means Clustering

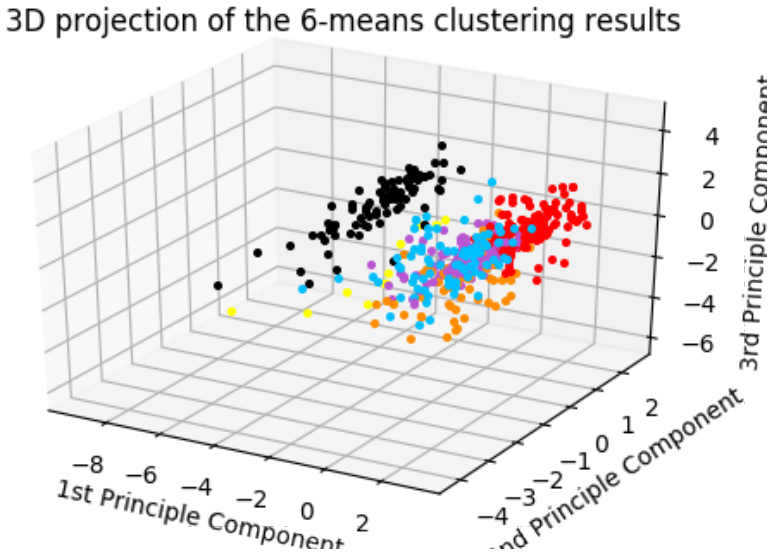

Figure 1.4.7: 6-Means Clustering

1.4.2 Clustering before PCA

Unlike the former subsection, we can't visually know how many cluster there in high dimensional space.

Therefore, we can only know from the quantization error figure, in which the knee point is also at K = 2, which means it's more possible that there are 2 clusters in the 11 dimensional space.

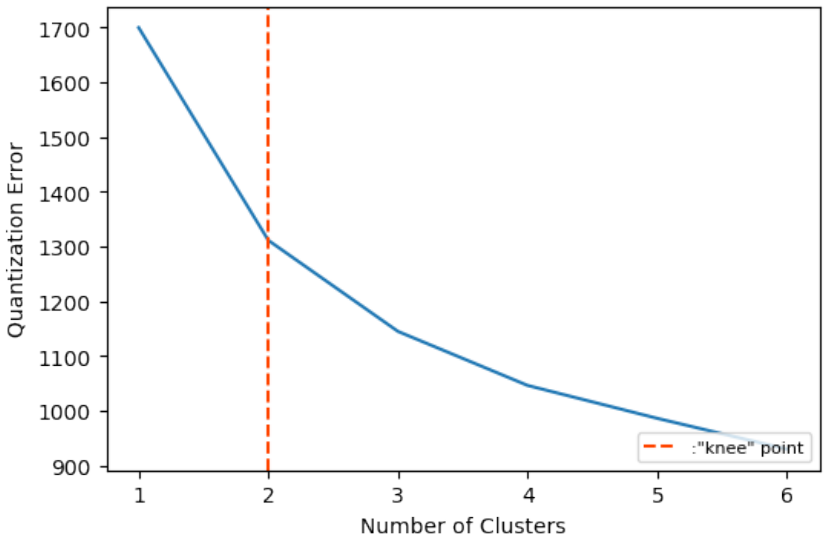

Figure 1.4.8: Quantization Error

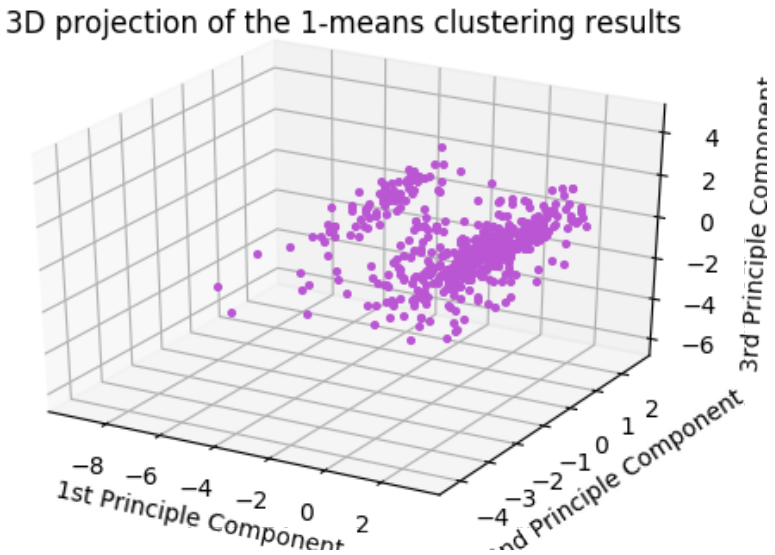

Figure 1.4.9: 1-Means Clustering

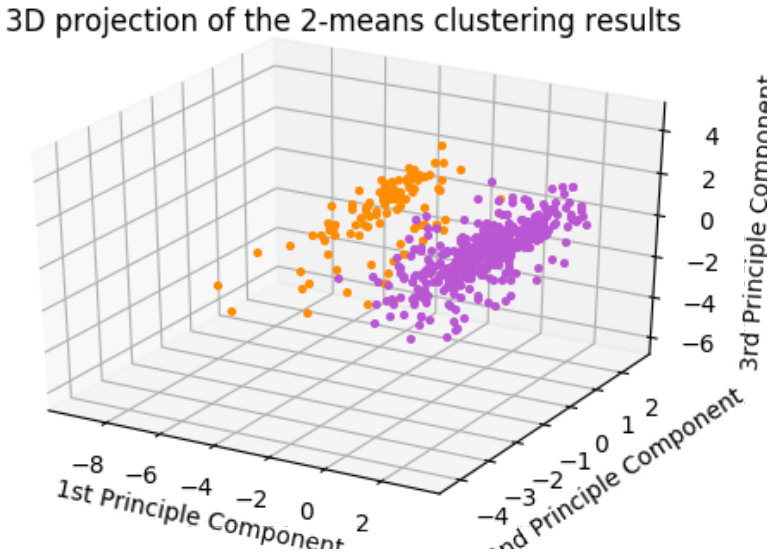

Figure 1.4.10: 2-Means Clustering

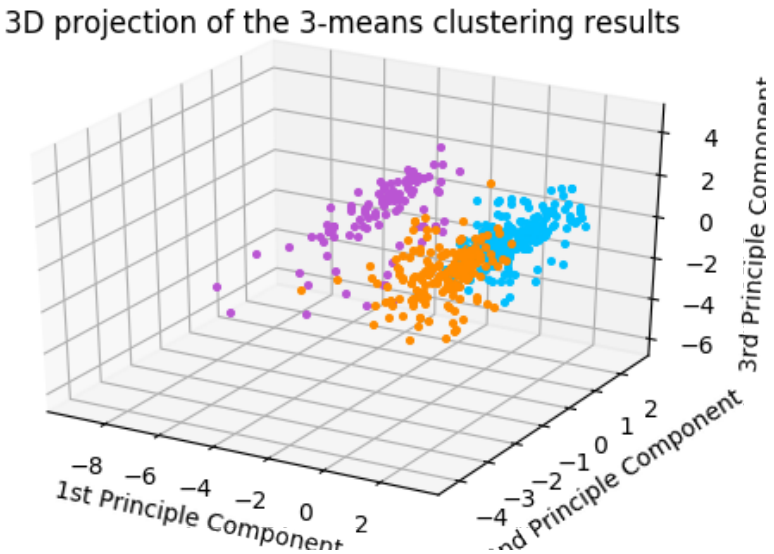

Figure 1.4.11: 3-Means Clustering

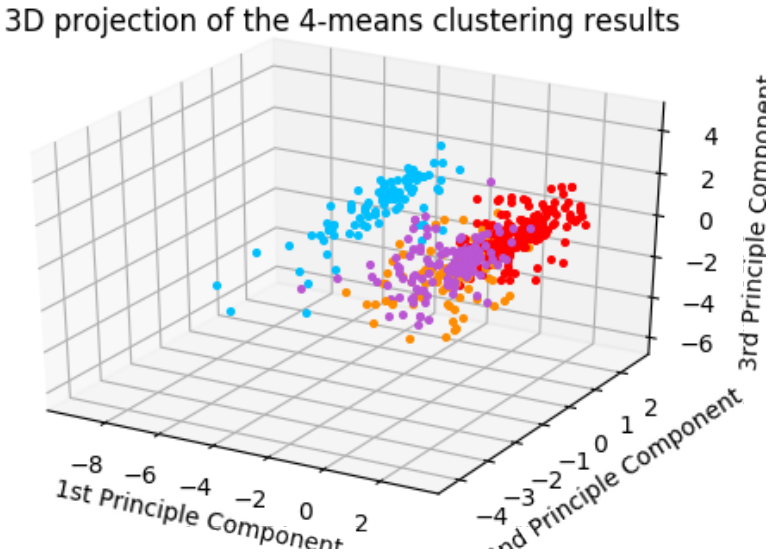

Figure 1.4.12: 4-Means Clustering

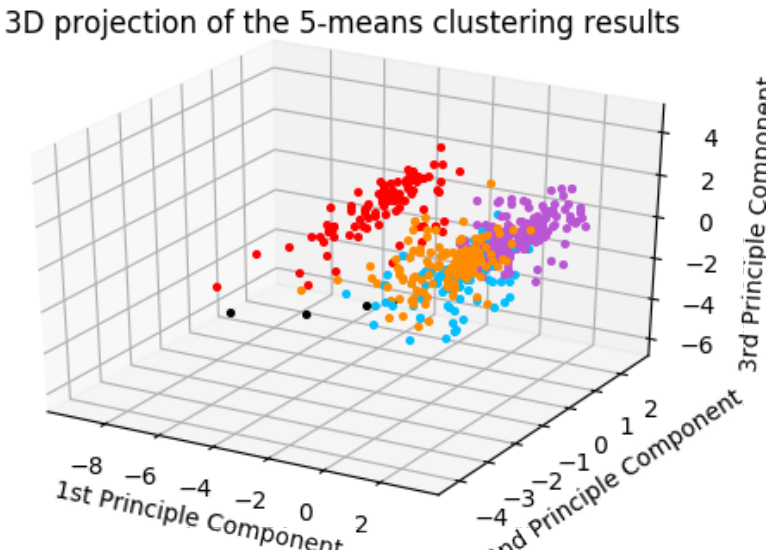

Figure 1.4.13: 5-Means Clustering

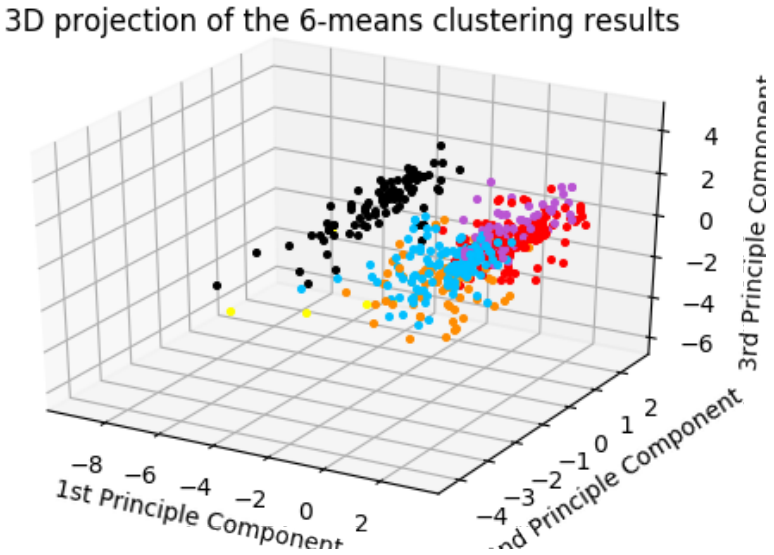

Figure 1.4.14: 6-Means Clustering

1.4.3 Comparing the sequence of PCA

Comparing **Section 1.4.1** and **Section 1.4.2**:

The knee point is not so obvious (the angle is not that sharp) in **Figure 1.4.8** like the one in **Figure 1.4.1**, which means that clustering in 11 dimensional space is vaguer than the one in the former 3 dimension.

So the my conclusion is that K-means clustering may not be a good way to analysis this data set.

Chapter 2

Initial Spread Index

2.1 Labelling

According to **McArthur Forest Fire Danger Index**, forest fires can be labelled by area as:

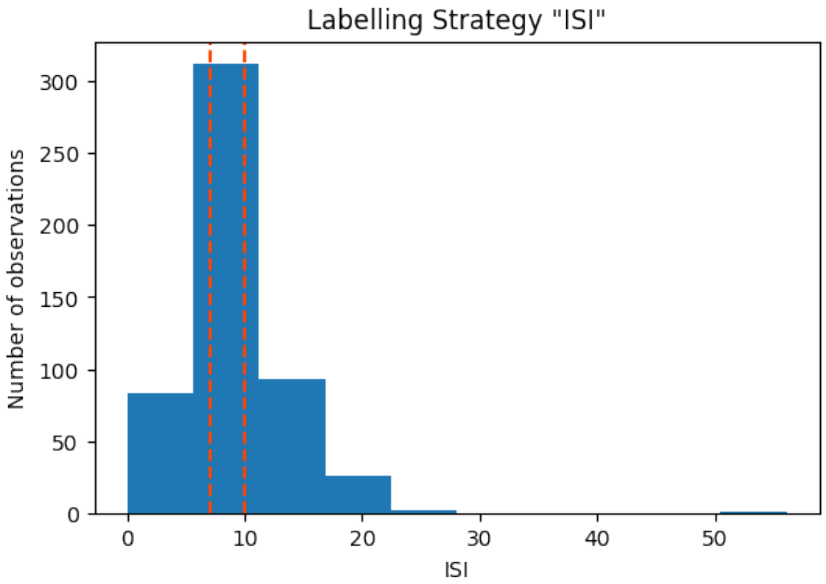

Figure 2.1.1: Labelling Strategy

For sake of lack of data, I label forest fires with three following classes (in ha):

- Class 1: Low ISI: (0, 7]
- Class 2: Moderate ISI: (7, 10]
- Class 3: High ISI: (10, ∞)

The final labelling result is:

- Class 1: 145 cases.
- Class 2: 217 cases.
- Class 3: 155 cases.

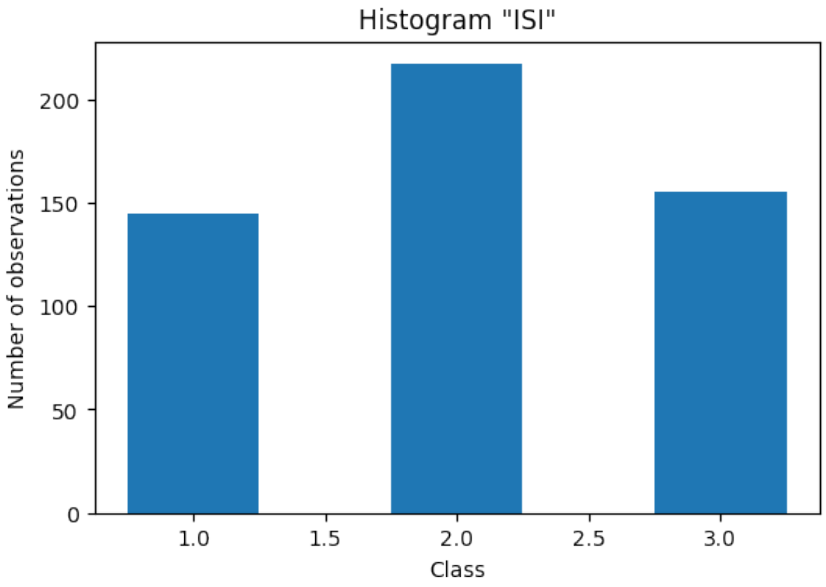

Figure 2.1.2: Histogram of ISI

2.2 Straightforward Co-ordinate Projections

In this section, let's just simply straightforward project the standardized data on a 2 dimensional plain. In **Figure 2.2.1**, the standardized data was projected on the **FFMC** and **wind** plain.

From **Figure 2.2.1**, we can approximately find the lower FFMC is, the lower the ISI is, which makes great sense because the more fine fuel in the air, the faster the forest fire can spread.

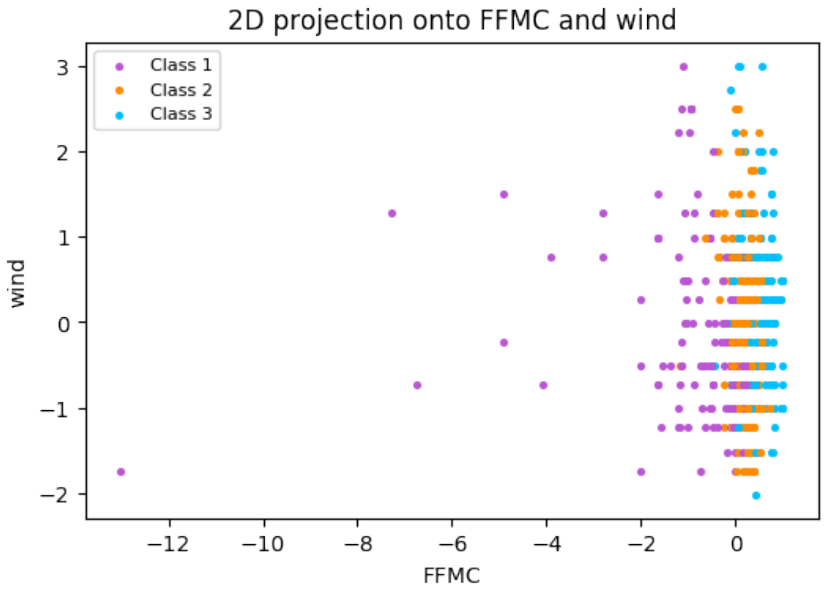

Figure 2.2.1: Projection on FFMC and wind

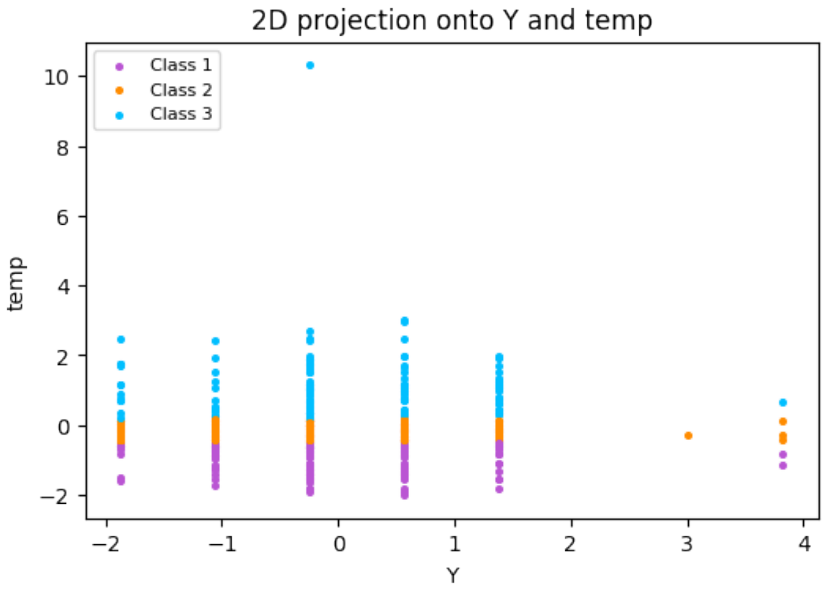

Figure 2.2.2: Projection on Y and temp

2.3 Visualization

2.3.1 Principal Component Analysis

Like what I did in **Section 1.3**, I used the same technique to complete this part.The result is that the sum of first two largest eigenvalues is **0.46279004** and the sum of first three largest eigenvalues is **0.5997552**, which is also not very ideal to project the data on a 2-dimensional or 3-dimensional plot. And that's why in **Figure 2.3.1** and **Figure 2.3.2** the plot is a little bit messy. In principle component analysis, there are 11 dimensions used (without **ISI** and **area** being results and **rain** dropped).

Dimension **Y** is the most important dimension, because in new axis 1 and 2, the value of **Y** is **-0.68206234** and **-0.68659901**. Dimension **temp** is the 2nd most important dimension, because in new axis 1 and 2, the value of **temp** is **0.62039458** and **0.64185592**.

Initial spread speed of forest fires can be mostly affected by **Y** and **temp**, maybe because if the latitude is low (the coefficient is negative), it will be very hot and there will be more rotten plants and woods that can be ignited quickly, and that will accelerate the initial spread speed. When I looked back to **Figure 2.2.2**, the figure proved my thought was right.

Top 2 largest eigenvalues and eigenvectors			
1st largest: 3.29706752		2nd largest: 1.53738519	
Dimensions	Values	Dimensions	Values
Y	-0.68206234	Y	-0.68659901
temp	0.62039458	temp	0.64185592
DC	0.3283881	DC	-0.28270006
month	0.13147483	month	0.14181775
X	-0.08656373	X	-0.07572487
RH	-0.0844196	wind	-0.05896374
FFMC	-0.07047972	DMC	-0.05662879
DMC	-0.07024343	RH	0.05064913
wind	0.01595388	day	0.03558612
day	0.00107296	FFMC	0.01793505

Table 2.3.1: Top two eigenvectors

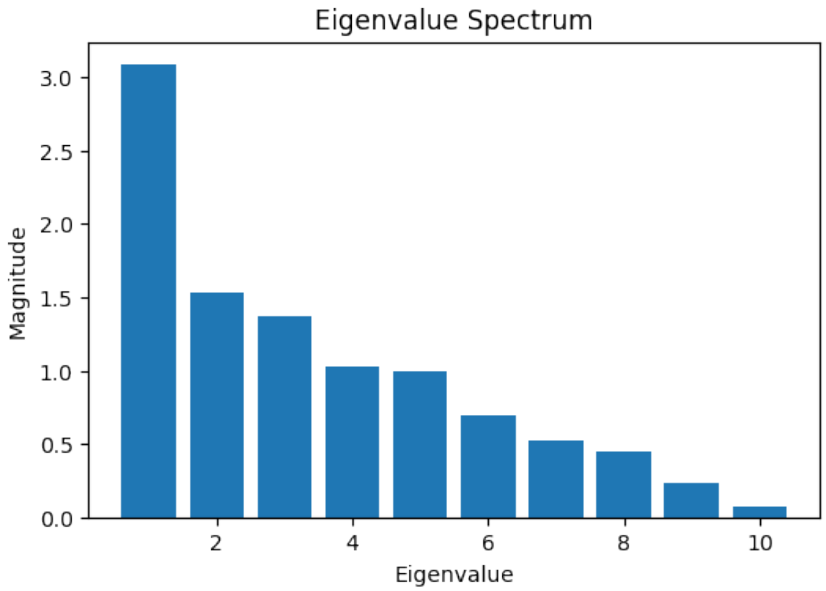

Figure 2.3.1: Eigenvalues of co-variance matrix

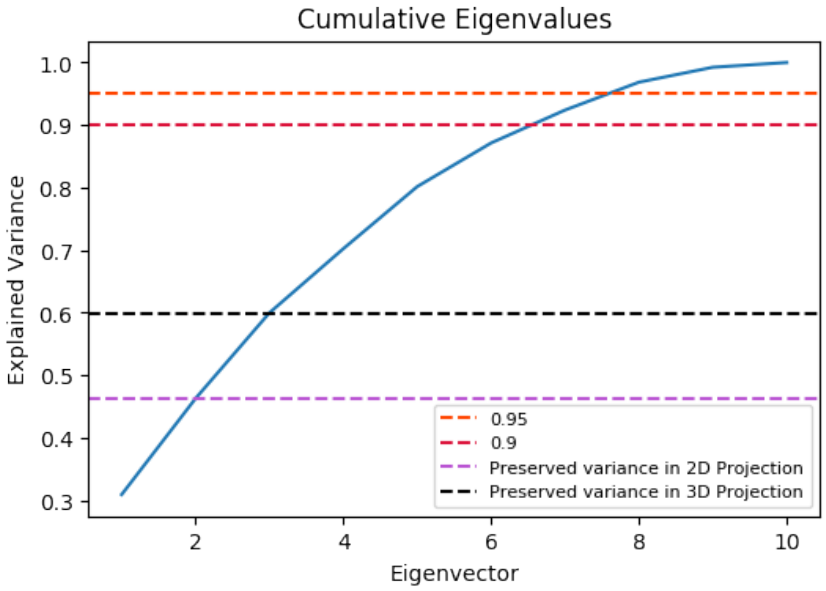

Figure 2.3.2: Cumulative eigenvalues of co-variance matrix

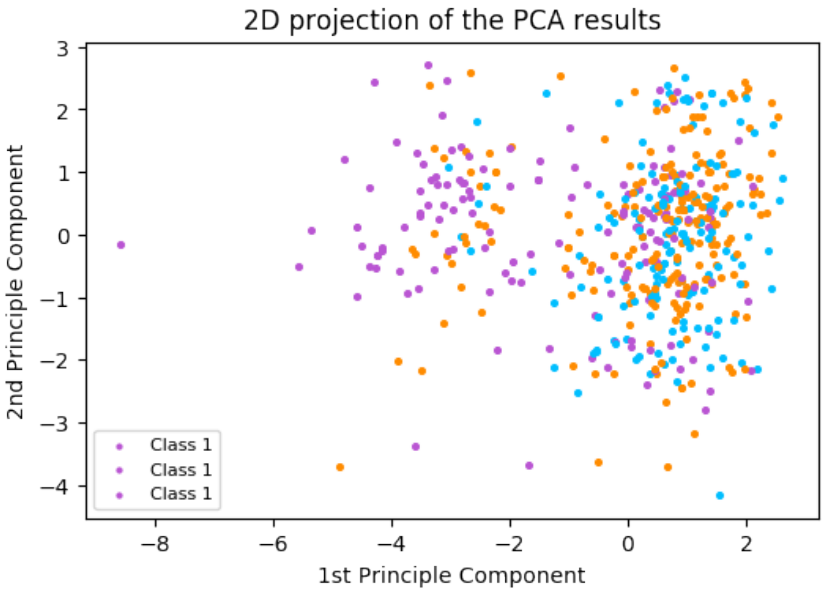

Figure 2.3.3: 2D PCA projection of ISI

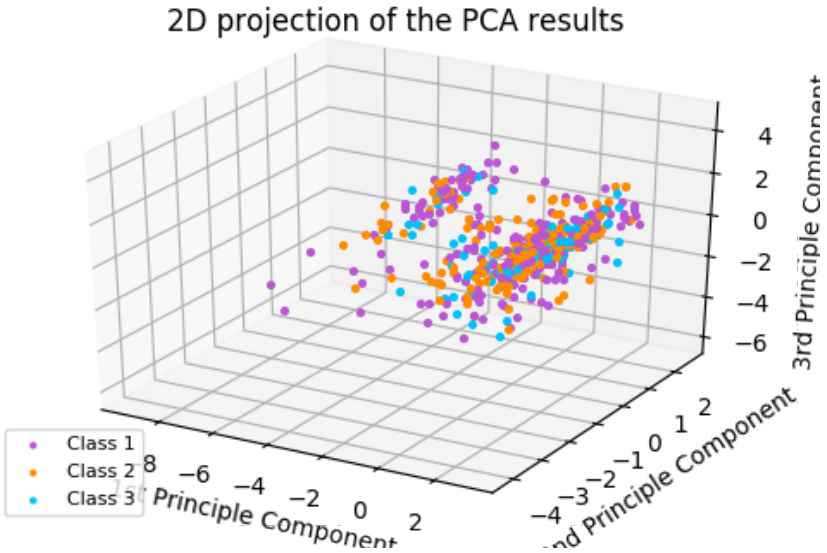

Figure 2.3.4: 3D PCA projection of ISI

2.4 Conclusion

As discussed above, ISI is most influenced by **Y** coordinates and **temp**, although the 2D and 3D PCA projection is not that obvious to observe.

Due to the limitation of working time and similarity of works, SOM and clustering were only written in **Chapter 1** (In **Chapter 1**, all requirements are covered!).

In future, I will continue working on the this chapter!

Thanks, Peter! You are really a perfect teacher!

We're looking forward to your next visit!