**CS5542 Big Data Apps and Analytics**

**In Class Programming –7 Report**

**(Jongkook Son)**

# Project Overview:

**Use a different data and use the model provided in ICP7 to perform clustering. You must try 5 different number of clusters (for example n\_clusters= 5 or n\_clusters=6,7,8,or 9 etc) based on elbow curve and for each cluster visualize the clustering results and report your findings in detail.**

# Requirements/Task(s):

1) Successfully executing the code and trying 5 different number of clusters based on elbow curve and visualizing those clusters using python plotting libraries (75 points)

2) Using a new and good dataset (5 points)

3) overall code quality (10 points)

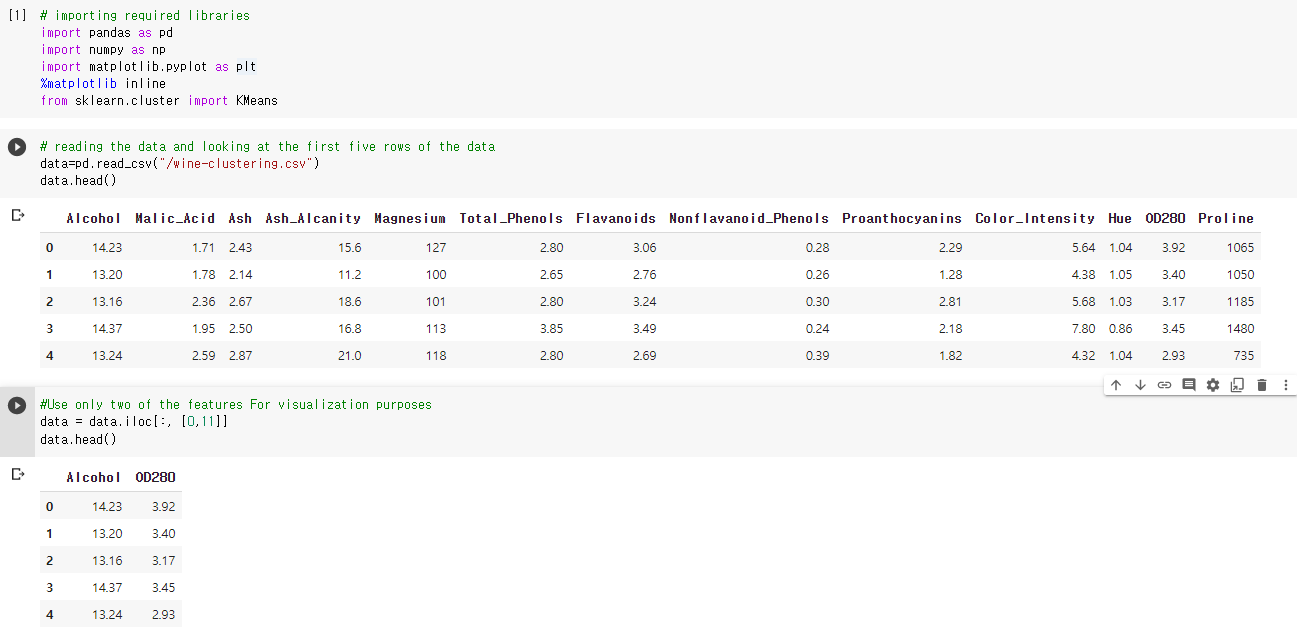
4) Pdf Report quality, video explanation (10 points)

**What I learned in ICP:**

I could have learned how to cluster data with K means algorithm. It was a little different kind of task because it was the unsupervised learning. All of the task before were supervised learning which have labels for the data. However it seems more easier than before. Thanks to this ICP7 I could get some insight how to get the optimized number of clustering group by using elbow method. It was basically draw a elbow graph adjusting cluster group number then find the point that is leveled off. And the start point of the leveled off is likely to be an optimized point. By doing this process the result of the clustering could get better.

**ICP description what was the task you were performing and Screen shots that shows the successful execution of each required step of your code**

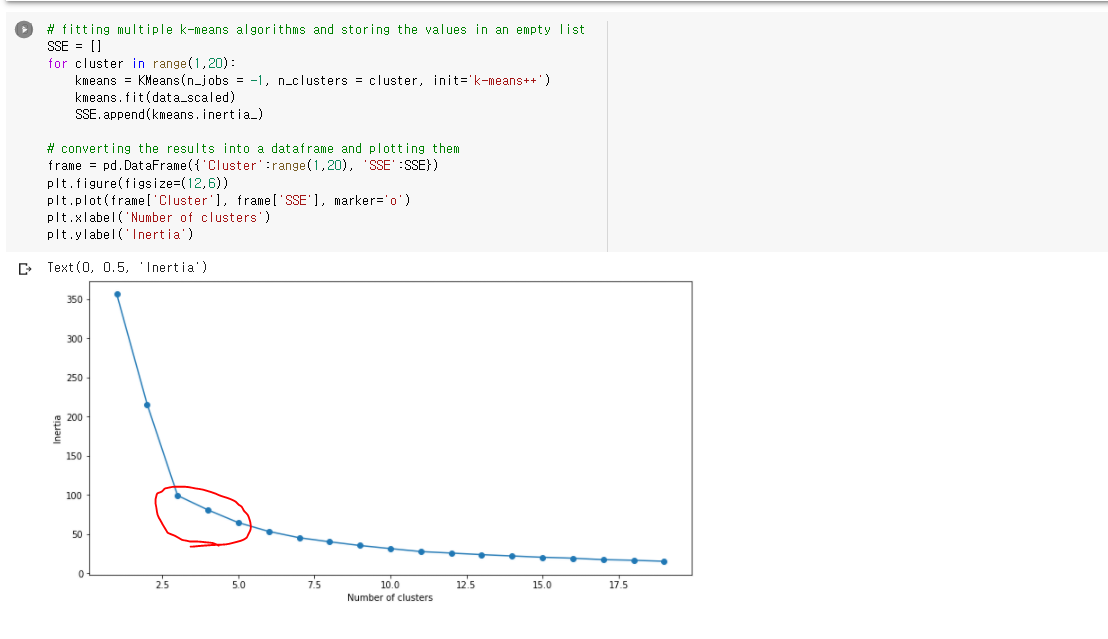
**Use a new data set which contains wine informations.**

****

* **For visualization purposes, I chose to include only two of the features. Two features that I decided to include is Alcohol and OD280.**

**Trying 5 different number of clusters based on elbow curve and visualizing those clusters using python plotting libraries**

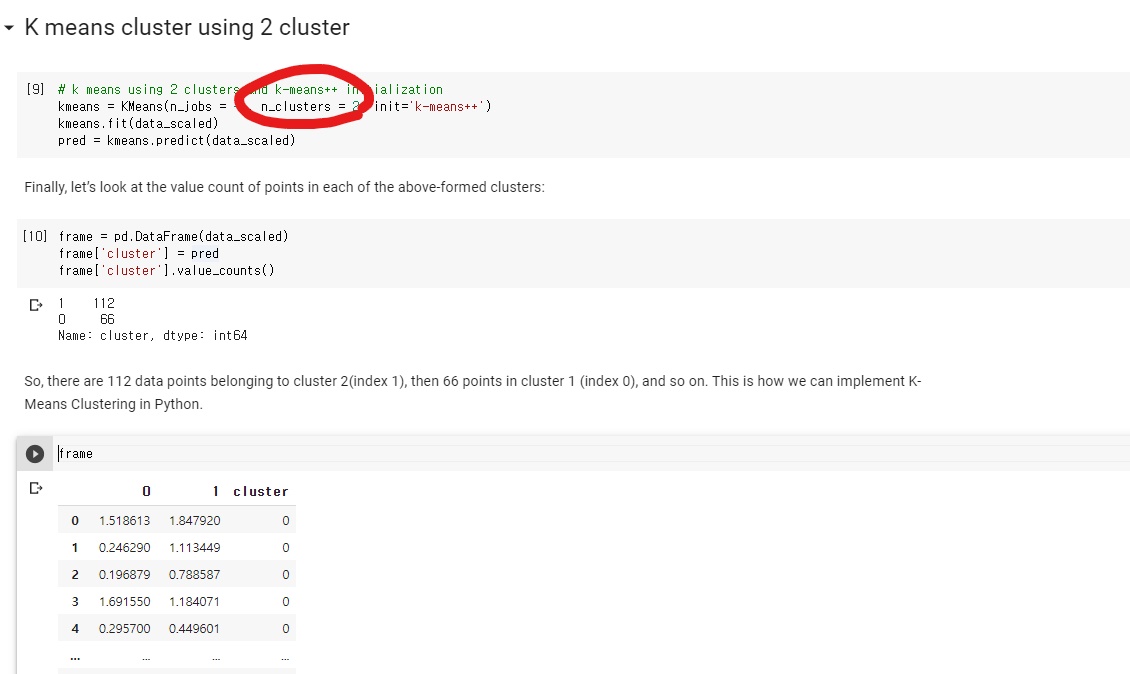
1. **Graph a Elbow Curve**

****

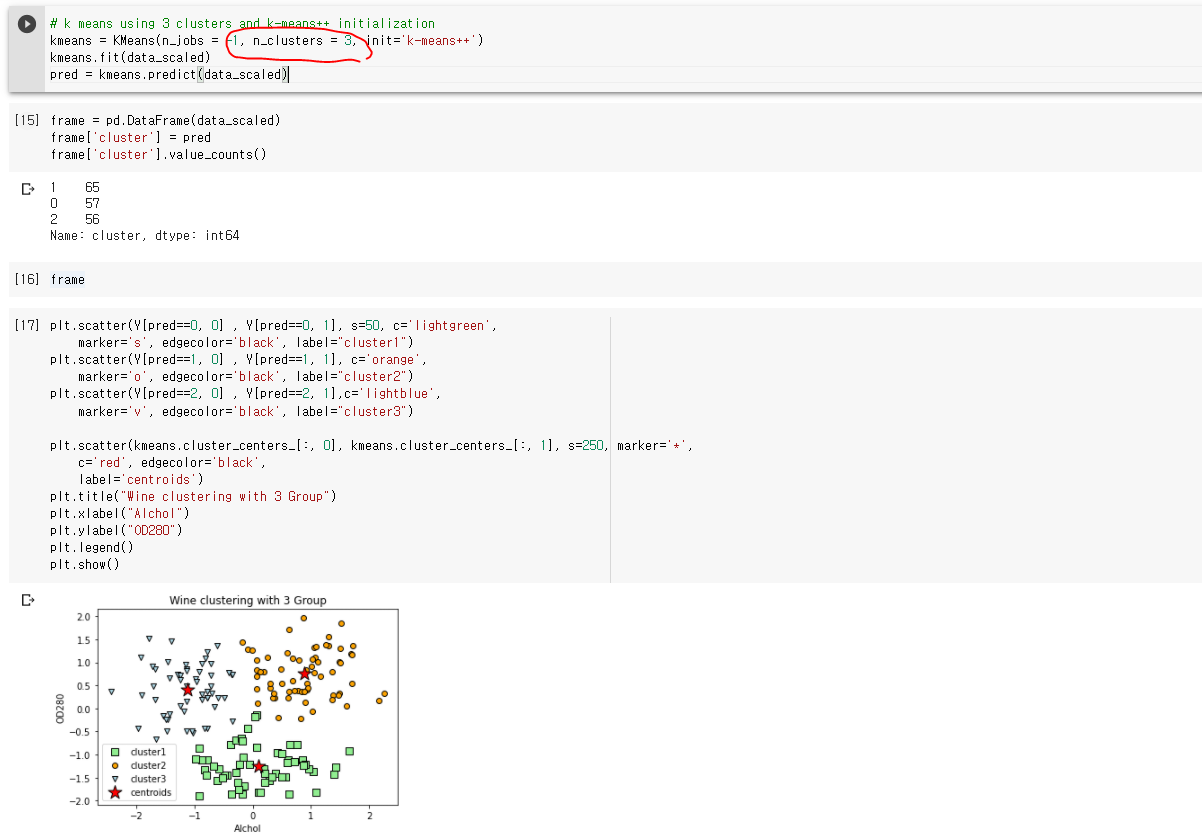
* **We graph the relationship between the number of clusters and Inertia then we select the number of clusters where the change in Inertia begins to level off (elbow method). In this case, Inertial value get leveled off after 2.5. So we can guess we need to start K with 2 or 3.**

**2.** **Executing the code and trying 5 different number of clusters based on elbow curve and visualizing those clusters**

1. **k means using 2 clusters and k-means++ initialization**

****

1. **k means using 3 clusters and k-means++ initialization**

****

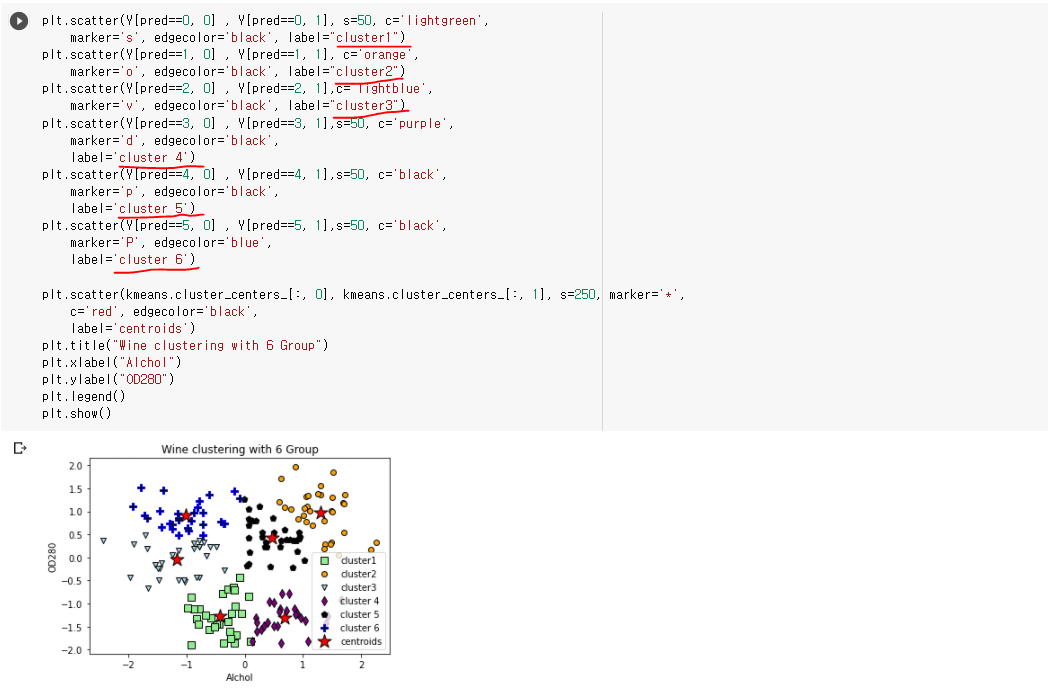
1. **k means using 4 clusters and k-means++ initialization**

****

1. **k means using 5 clusters and k-means++ initialization**

****

1. **k means using 6 clusters and k-means++ initialization**

****

# Conclusion:

# Based on elbow curve, I guessed k number for clustering from 2 to 6. Actually it is hard to accurately select k number in k means clustering. However, When You see Elbow curve, inertia value is leveled off after 2.5. So we can guess the best number for k is 3. As You can see when k=3, it is clustered well visually. So one can decide 3 is the appropriate number for k.

# Challenges that I faced:

The most difficult challenge that I faced was that how to find optimized number for k means clustering. I want prove it with some number but It was hard to do that. I could have guessed the best number for the clustering number based on elbow curve finally. It would be better if I would use different kind of cluster algorithm like hierarchical clustering.

# Video link

# <https://www.youtube.com/watch?v=teqxvmg1gKQ>