11주차 과제

Clustering

담당교수님: 윤장혁 교수님

과목명: Data Analytics

이름: 박민성

학번: 201611145

전공: 산업공학과

제출일: 2020.05.29

Object

- 주어진 데이터는 2차원 벡터 값임
- 최적의 K 값을 고민 후 결정

최적의 K 값 산출방법 설명

- 최종적으로 결정된 K값으로 clustering을 진행한 후 그림으로 표현

Visualization을 위한 tool에는 제한을 두지 않음

최적의 K값 결정

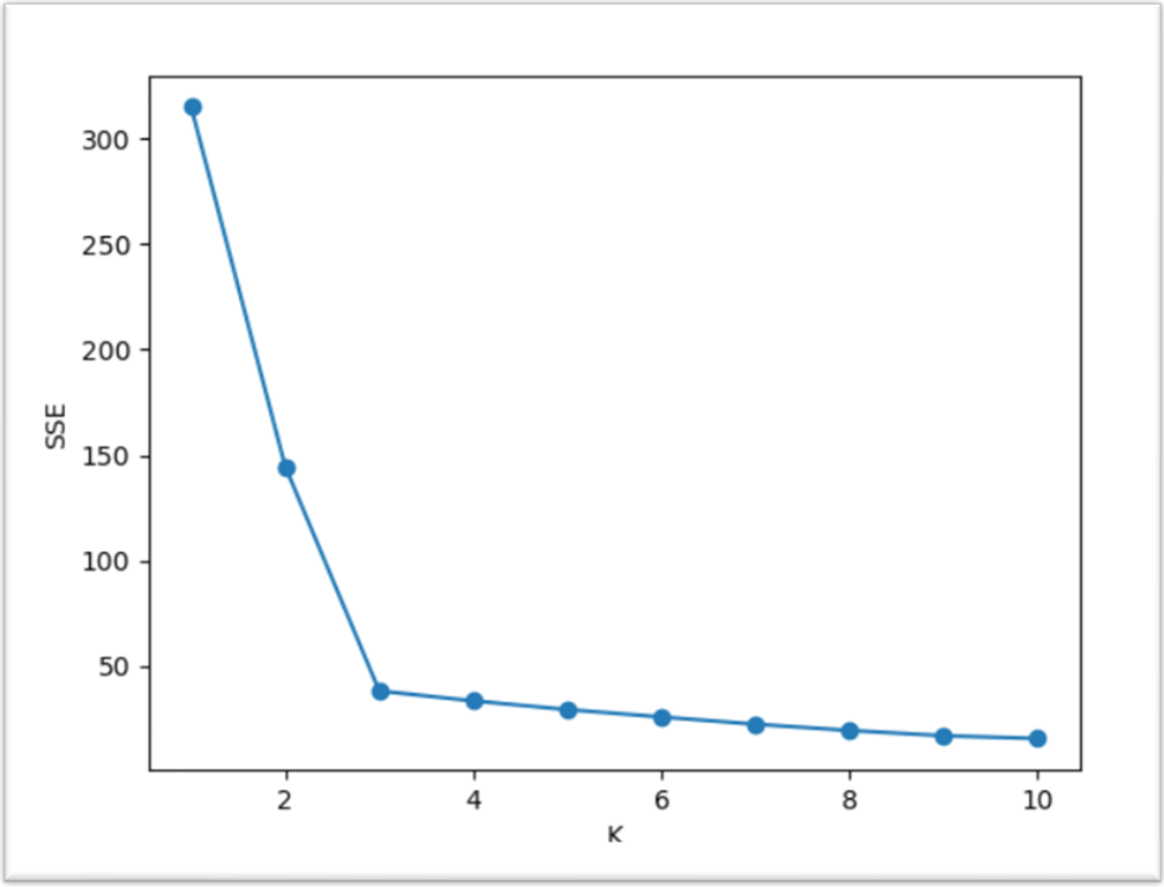
Elbow 기법을 통해서 K값을도출했고 분석툴로는 파이썬을 활용했다.

C: > Users > minisong > Desktop > 11주차.py > elbow

```python
import pandas as pd
import numpy as np
import csv
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
import seaborn as sns
import matplotlib
from sklearn.preprocessing import MinMaxScaler
data= []
f = open('./data_week11.csv','r')
rdr = csv.reader(f)
for i in rdr:
    data.append(i)
del data[0]
scaler = MinMaxScaler(feature_range=(0, 1))
scaler.fit_transform(data)
feature = scaler.fit_transform(data)
def elbow(X):
    sse = []
    for i in range(1,11):
        km = KMeans(n_clusters=i,algorithm='auto', random_state=42)
        km.fit(X)
        sse.append(km.inertia_)

    plt.plot(range(1,11), sse, marker='o')
    plt.xlabel('K')
    plt.ylabel('SSE')
    plt.show()
elbow(feature)
```

아래는 코드 29번에 대한 실행 결과이다.



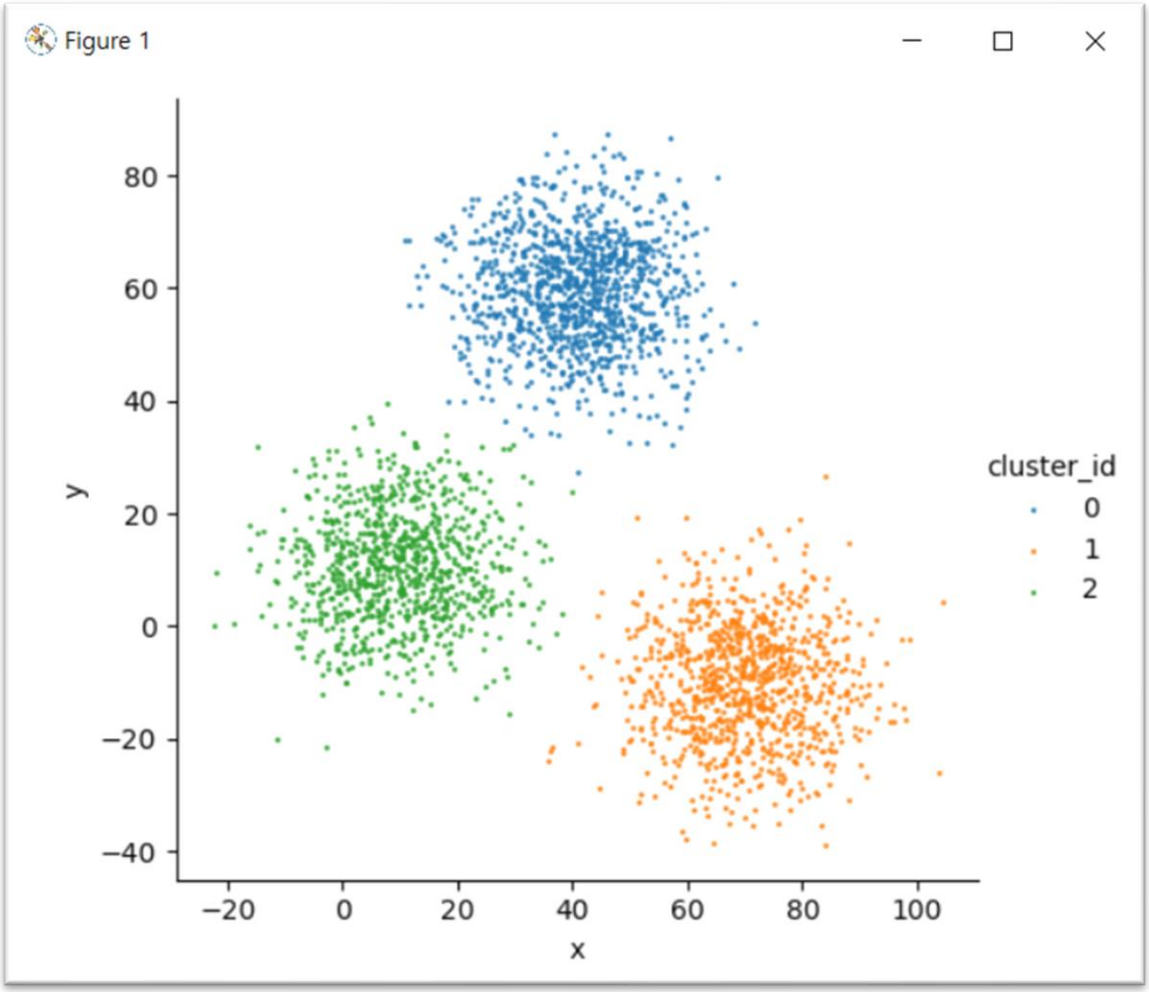
그래프가 flat 지점에 진입한 것이 K=3일 때라고 판단하고 K값을 3으로 정했다.

Clustering and Visualization

시각화툴은 파이썬의 matplotlib을 활용하였다.

```
30  df = pd.DataFrame(columns =('x','y'))
31  for i in range(0,len(data)):
32      df.loc[i] = [float(data[i][0]),float(data[i][1])]
33  data_points = df.values
34  kmeans = KMeans(n_clusters=3).fit(data_points)
35  kmeans.labels_
36  df['cluster_id'] = kmeans.labels_
37  sns.lmplot('x','y',data=df,fit_reg = False, scatter_kws ={"s":1},hue="cluster_id")
38  plt.show()
```

아래는 위의 코드의 최종 실행 결과이다.



K=3일 때 data_week11.csv를 Clustering한 결과를 시각화한 자료이다.