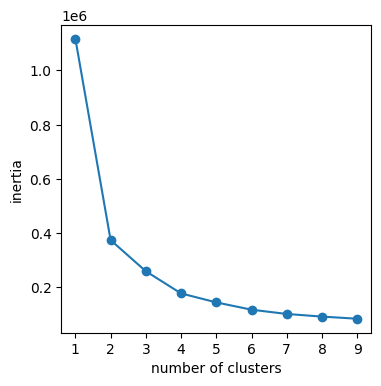
HW6\_20210815\_정재현

A. clustering

먼저 Clustering을 할 때, K-means 기법을 사용하기 위해서 sklearn.cluter의 KMeans를 import 해주었다. 우리에게 주어진 파일을 분석해보면 Health\_Condition\_Level.csv가 존재한다. 이때 csv 파일을 보면 SC,Smoker, Medical\_history,AC,HSM, Sex 등이 categorical로 되어 있는 것을 확인할 수 있었다. 그래서 이를 float64형태로 바꾸기 위해 인코딩을 진행하였다. 먼저 Smoker, Medical\_history, HSM 등은 no, yes의 binary이기 때문에 no의 문자를 가지는 값은 0으로 대치하고, yes 문자열은 1로 되도록 바꿔주었다. 그 다음, Sex의 경우 Female은 0으로, Male은 1로 대치 시켰다. 이어서 SC,AC는 빈도에 따라 no, Sometimes, Frequently, Always가 존재하여 순서대로 0, 0.33, 0.66, 1로 대치시켰다. 그 이후 astype을 이용하여 categorical한 데이터 타입을 가진 그룹에 대해 float64형으로 인코딩을 시켜주었다.

그 이후 inertias를 만들고 x\_list를 1부터 10개의 값을 가지도록 하였고, for문을 이용하여 KMeans의 n\_cluster의 값을 1부터 9까지 변화시키면서 분석을 해보았다. 그리고 저장한 값을 x axis = n\_cluster, y axis = inertias로 하여 분석하면 아래와 같은 사진이 나온다.



따라서 cluster의 값은 3에서 최적이라는 것을 알 수 있었다. 그 다음 model에 Kmeans를 실행하여 n\_cluster의 값이 3일때로 하고 fit을 진행하였다. 이후 값을 csv로 만들기 위해 dataframe을 구성하였는데, id라는 리스트에 range(0,1839)을 사용하여 0부터 1838까지의 값을 담고 dataframe에 id와 아까 추출한 predict를 행으로 하고 값을 각각 넣어준 후 to\_csv로 submission\_cluster.csv로 저장하는데, 이때 샘플 값으로 주어진 파일에서는 index가 없었으므로 index=False라는 서식을 부여하여 정상적으로 cluster 방법 중 KMeans 사용을 완료하였다.