# 案例三 各国幸福指数聚类

## 【1】案例背景

本题由 **顾翔 921127940122**负责

世界幸福报告是对世界各国幸福状态调查的一个具有里程碑意义的成果，第1份世界幸福报告出版于2012年,本案例提供了2016年全球157个国家和地区的幸福指标调查结果，其中包括国家经济指数，家庭指数健康指数和政府信息指数等信息，请将国家按照幸福状态进行聚类，并分析聚类结果与国家所在区域的关系

## 【2】方法陈述

首先使用pandas对数据进行处理,读取数据

对数据进行K-means++聚类，采用肘方法确定聚类数k，对结果进行分析

运行环境，python3安装有pandas、scikit-learn、numpy库，运行软件：Anaconda、jupyter notebook

## 【3】实验代码

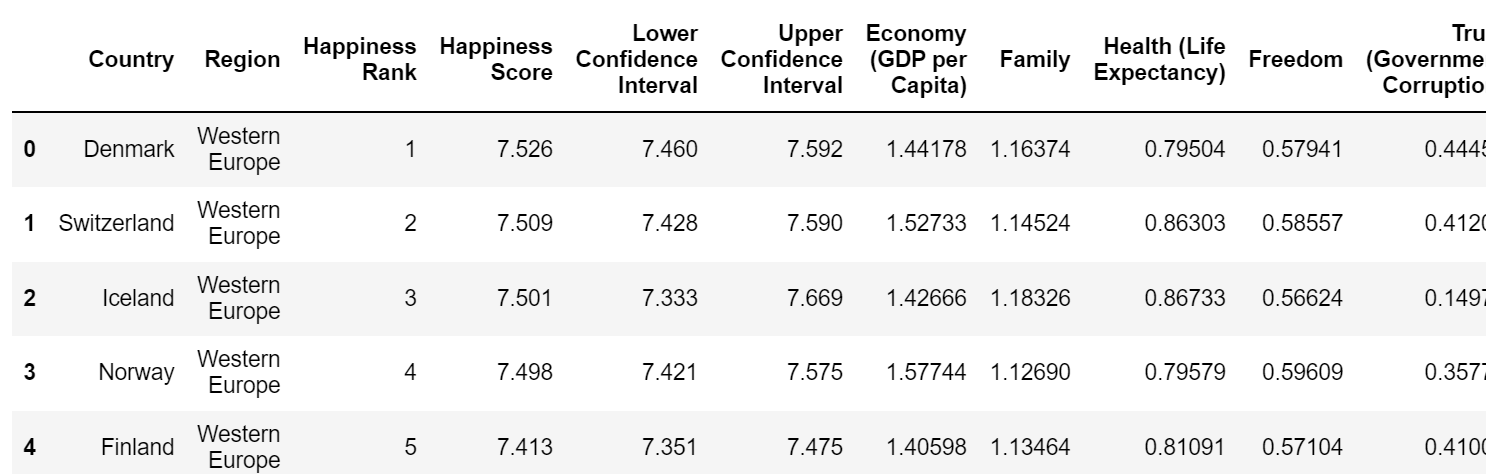
### 步骤一、数据处理

1.数据处理

import pandas as pd

df=pd.read\_csv('2016\_world\_happiness.csv')

df.head()



### 步骤二、聚类

**1. K-means++**

* 用K-means++聚类，在初始化质心的过程中，选择相互距离尽可能远的样本作为初始质心
* 采用肘方法确定聚类数k
* 将地区转化为了01向量
* 库参考文档：<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

import numpy as **np**

from sklearn.cluster import KMeans

import **matplotlib**.**pyplot** as **plt**

from sklearn import cluster

data=df.**drop**('Happiness Rank',axis=1)

data=data.**drop**('Happiness Score',axis=1)

data=data.**drop**('Country',axis=1)

data=**pd**.**get\_dummies**(data)

distortions = []

for i in **range**(1,11):

    km = KMeans(n\_clusters=i,init='k-means++',n\_init=10,max\_iter=300,random\_state=0)

    km.fit(data)

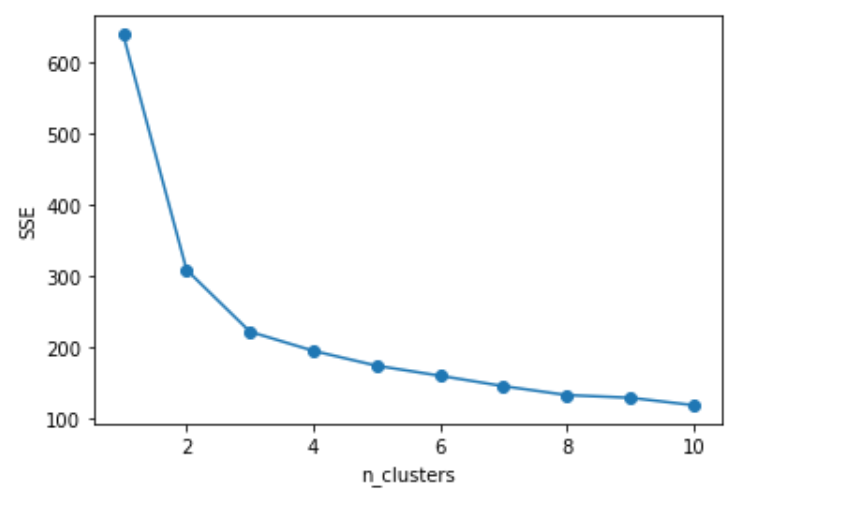
    distortions.**append**(km.inertia\_)

**plt**.**plot**(**range**(1,11), distortions, marker = 'o')

**plt**.**xlabel**('n\_clusters')

**plt**.**ylabel**('SSE')

**plt**.**show**()



## 结果分析

1. 根据肘方法，选择k=6较为合适
2. 查看整体情况

km = KMeans(n\_clusters=6,init='k-means++',n\_init=10,max\_iter=300,random\_state=5)

km.fit(data)

**print**('误差为:',km.inertia\_)

labels = km.labels\_

subDataDF1=df[["Country"]]

**pd**.**DataFrame**(labels, index=subDataDF1.Country,columns=['Country ID'])

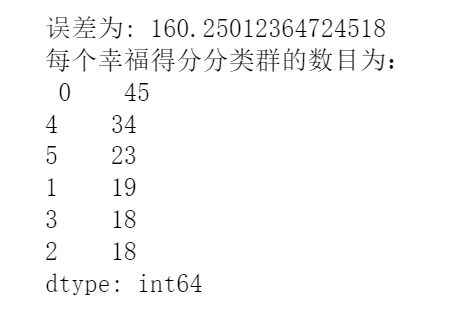
num=**pd**.**Series**(labels).**value\_counts**()

**print**('每个幸福得分分类群的数目为：\n',num)

for i in **range**(0,6):

    temp=**pd**.**DataFrame**()

    temp=temp.append(df[df['Country']=='Norway'])



1. 接下来对六个聚类，进行数据统计分析幸福状态与区位条件关系

subDataDF1=df[["Country"]]

res=**pd**.**DataFrame**(labels, index=subDataDF1.Country,columns=['Country ID'])

temp=res[res['Country ID']==0]

temp.**head**()

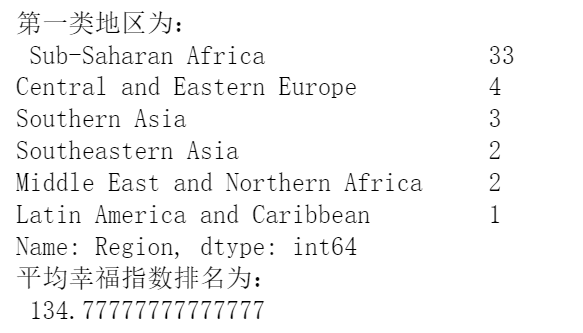
ans=**pd**.**DataFrame**()

for i in **range**(temp.shape[0]):

    ans=ans.append(df[df['Country']==temp.index.values[i]])

**print**('第一类地区为：\n',ans['Region'].**value\_counts**())

**print**('平均幸福指数排名为：\n',ans['Happiness Rank'].**mean**())



temp=res[res['Country ID']==1]

temp.**head**()

ans=**pd**.**DataFrame**()

for i in **range**(temp.shape[0]):

    ans=ans.append(df[df['Country']==temp.index.values[i]])

**print**('第二类地区为：\n',ans['Region'].**value\_counts**())

**print**('平均幸福指数排名为：\n',ans['Happiness Rank'].**mean**())



temp=res[res['Country ID']==2]

temp.**head**()

ans=**pd**.**DataFrame**()

for i in **range**(temp.shape[0]):

    ans=ans.append(df[df['Country']==temp.index.values[i]])

**print**('第三类地区为：\n',ans['Region'].**value\_counts**())

**print**('平均幸福指数排名为：\n',ans['Happiness Rank'].**mean**())



temp=res[res['Country ID']==3]

temp.**head**()

ans=**pd**.**DataFrame**()

for i in **range**(temp.shape[0]):

    ans=ans.append(df[df['Country']==temp.index.values[i]])

**print**('第四类地区为：\n',ans['Region'].**value\_counts**())

**print**('平均幸福指数排名为：\n',ans['Happiness Rank'].**mean**())



temp=res[res['Country ID']==4]

temp.**head**()

ans=**pd**.**DataFrame**()

for i in **range**(temp.shape[0]):

    ans=ans.append(df[df['Country']==temp.index.values[i]])

**print**('第五类地区为：\n',ans['Region'].**value\_counts**())

**print**('平均幸福指数排名为：\n',ans['Happiness Rank'].**mean**())



temp=res[res['Country ID']==5]

temp.**head**()

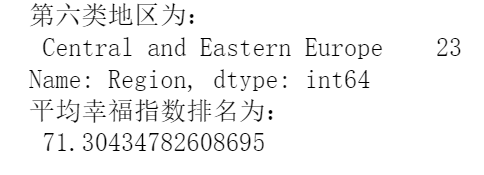
ans=**pd**.**DataFrame**()

for i in **range**(temp.shape[0]):

    ans=ans.append(df[df['Country']==temp.index.values[i]])

**print**('第六类地区为：\n',ans['Region'].**value\_counts**())

**print**('平均幸福指数排名为：\n',ans['Happiness Rank'].**mean**())



1. 从结果中不难发现，区位因素与幸福状态是强相关的，撒哈拉以南非洲资源匮乏，环境恶劣，政局动荡，经济落后，最不发达的国家集中于此；西欧北美经济发达，资源雄厚，政治民主，医疗优越，因此生活幸福指数最高国家集中于此。这在逻辑上也十分合理，发达国家/不发达国家有很大程度受到所在区位的气候地质环境水文生态的影响，当地国家的资源、地理位置，直接影响经济政治发展程度，从而影响人们生活幸福指数。