**Business objective:** Minimize public’s Risk

**INFERENCE:**  The given data set contains crime data in different states in USA. So ,this data set contains the no.of murders,rapes,pop culture areas, assaults. After importing the data and normalizing all the columns except ‘states’ and after applying Euclidian distance and to find the distance ,I have experimented various combinations and single linkage and complete linkage are showing me some good results .Hence I have created 2 clusters cluster(1,2).one cluster refers to dangerous cities where there are more murders,rapes,assaults and in the other cluster there are cities are quite safe from cluster 2. I have also applying 3,4 clusters ,but when I applied 3,4 clusters the results were comparable ,there were no significant differences between all the clusters .Hence I have decided to create only 2 clusters. The below is the code for it in R language.

**Group.1 Murder Assault UrbanPop Rape**

**1 1 12.165 255.2500 68.40000 29.16500**

**2 2 4.870 114.4333 63.63333 15.9433**

**# R code for Agglomerative clustering**

library(readr)

library(readxl)

library(readr)

crime\_data <- read\_csv("E:/Assignments/Assignments week 2/m4/crime\_data.csv")

View(crime\_data)

summary(crime\_data)

normalize<-scale(crime\_data[,2:5])

View(normalize)

summary(normalize)

d<- dist(normalize,method='euclidean')

fit<-hclust(d,method="average")

plot(fit)

plot(fit,hang=-1)

groups<-cutree(fit,k=2)

rect.hclust(fit,k=2,border='red')

crimes<-as.matrix(groups)

final<-data.frame(crimes,crime\_data)

View(crime\_data)

View(final)

aggregate(crime\_data[,2:5],by=list(final$crimes),FUN = mean)

library(readr)

write\_csv(final,"crimess.csv")

getwd()

**INFERENCE:**  The given data set contains crime data in different states in USA. So ,this data set contains the no.of murders,rapes,pop culture areas, assaults. After importing the data and normalizing all the columns except ‘states’ and after applying Euclidian distance and to find the distance ,I have experimented various combinations and single linkage and complete linkage are showing me some good results .Hence I have created 2 clusters cluster(1,2).one cluster refers to dangerous cities where there are more murders,rapes,assaults and in the other cluster there are cities are quite safe from cluster 2. I have also applying 3,4 clusters ,but when I applied 3,4 clusters the results were comparable ,there were no significant differences between all the clusters .Hence I have decided to create only 2 clusters. The below is the code for it in python language.

Murder Assault Urban Pop Rape ranking

12.452 114.433333 63.633333 15.943333 0

4.987 255.250000 68.400000 29.165000 1

#python code

import pandas as pd

import matplotlib.pylab as plt

crime\_data=pd.read\_csv('E:/data/crime\_data.csv')

crime\_data.describe()

#Normalize function

def norm\_func(i):

x=(i-i.min()) / (i.max()-i.min())

return(x)

df\_norm=norm\_func(crime\_data.iloc[:,1:])

df\_norm.describe()

#for creating dendogram

from scipy.cluster.hierarchy import linkage

import scipy.cluster.hierarchy as sch

k=linkage(df\_norm,method="complete",metric="euclidean")

#dendogram

plt.figure(figsize=(15,8));plt.title('Agglomerative clustering');plt.xlabel('state Index');plt.ylabel('Crimes')

sch.dendrogram(k,leaf\_rotation=0,leaf\_font\_size=10)

plt.show()

#apply agglomerative clustering

from sklearn.cluster import AgglomerativeClustering

h\_complete=AgglomerativeClustering(n\_clusters=2,linkage='complete',affinity="euclidean").fit(df\_norm)

h\_complete.labels\_

cluster\_labels=pd.Series(h\_complete.labels\_)

crime\_data['ranking']=cluster\_labels

final\_data=crime\_data.iloc[:, [5,0,1,2,3,4]]

final\_data.head()

#aggregate mean of each cluster

crime\_data.iloc[:, 2:].groupby(final\_data.ranking).mean()

#creating a csv file

crime\_data.to\_csv("crimedatapython.csv",encoding="utf-8")

import os

os.getcwd().