VSIT

Business Intelligence
Practical #8

Name	Sanjeev Gupta	**Roll Number**	21302B0023
Class	TYBSc IT	**Division**	C
Subject/Course:	Business intelligence		
Topic	Clustering		

Q1. Write Overview of Clustering.

Clustering is a fundamental technique in data analysis and business intelligence (BI) that deals with grouping similar entities or data points together based on certain characteristics or attributes. It is an unsupervised machine learning approach, meaning that it does not rely on labeled data for training. The primary goal of clustering in BI is to discover hidden patterns, relationships, and structures within large datasets, enabling organizations to make informed decisions, optimize operations, and enhance overall performance.

Importance of Clustering in Business Intelligence

Clustering plays a crucial role in BI by providing a wide range of benefits, such as:

Data Exploration: Clustering helps organizations identify inherent patterns and relationships within their datasets, facilitating better understanding and interpretation of the underlying information.

Customer Segmentation: By segmenting customers based on shared characteristics, businesses can tailor their marketing strategies, product offerings, and customer support to meet specific customer needs and preferences.

Risk Management: Clustering can help identify groups of similar high-risk investments, enabling financial institutions to develop targeted risk mitigation strategies.

Anomaly Detection: Detecting outliers or anomalies within large datasets can alert organizations to potential issues or fraudulent activities, allowing them to take proactive measures to address these concerns.

Operational Optimization: By identifying patterns in operational processes and resource utilization, organizations can optimize resource allocation, reduce costs, and improve efficiency.

Predictive Analytics: Clustering results can be used as input for predictive models, enhancing the accuracy of forecasts and predictions.

Q2. Explain K-means Clustering method.

K-means clustering is a type of unsupervised machine learning algorithm used to classify unlabeled data, i.e., data without defined categories or groups. The goal of this algorithm is to find K groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of the K groups based on the features that are presented in the data.

The K-means Clustering Process

The K-means clustering process begins with randomly initializing K centroids, where K is the predefined number of clusters. The centroids are representative points of the clusters, and each data point is assigned to the cluster with the nearest centroid. The distance between a data point and a centroid is typically measured using Euclidean distance.

Once all data points have been assigned to a cluster, the centroids are recalculated as the mean of all the data points belonging to that cluster. This process is repeated iteratively, with data points being

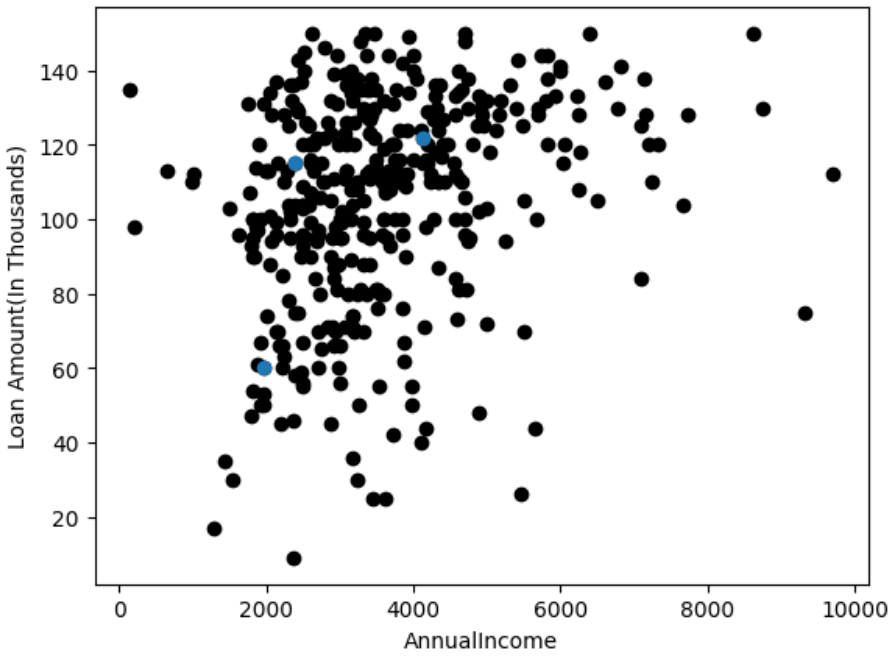
reassigned to clusters based on the new centroids, and the centroids being recalculated until the assignments no longer change or a maximum number of iterations has been reached.

Q3. Perform Clustering using K-means method.

```
import pandas as pd
import numpy as np
import random as rd
import matplotlib.pyplot as plt
data=pd.read_csv('clustering.csv')
data.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
1	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
2	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
3	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y
4	LP001013	Male	Yes	0	Not Graduate	No	2333	1516.0	95.0	360.0	1.0	Urban	Y

```
X=data[["LoanAmount","ApplicantIncome"]]
plt.scatter(X["ApplicantIncome"],X["LoanAmount"],c='black')
plt.xlabel("AnnualIncome")
plt.ylabel("Loan Amount (In Thousands)")
plt.show()
K=3
K=3
Centroids=(X.sample(n=K))
plt.scatter(X["ApplicantIncome"],X["LoanAmount"],c='black')
plt.scatter(Centroids["ApplicantIncome"],Centroids["LoanAmount"])
plt.xlabel('AnnualIncome')
plt.ylabel('Loan Amount(In Thousands)')
plt.show()
```

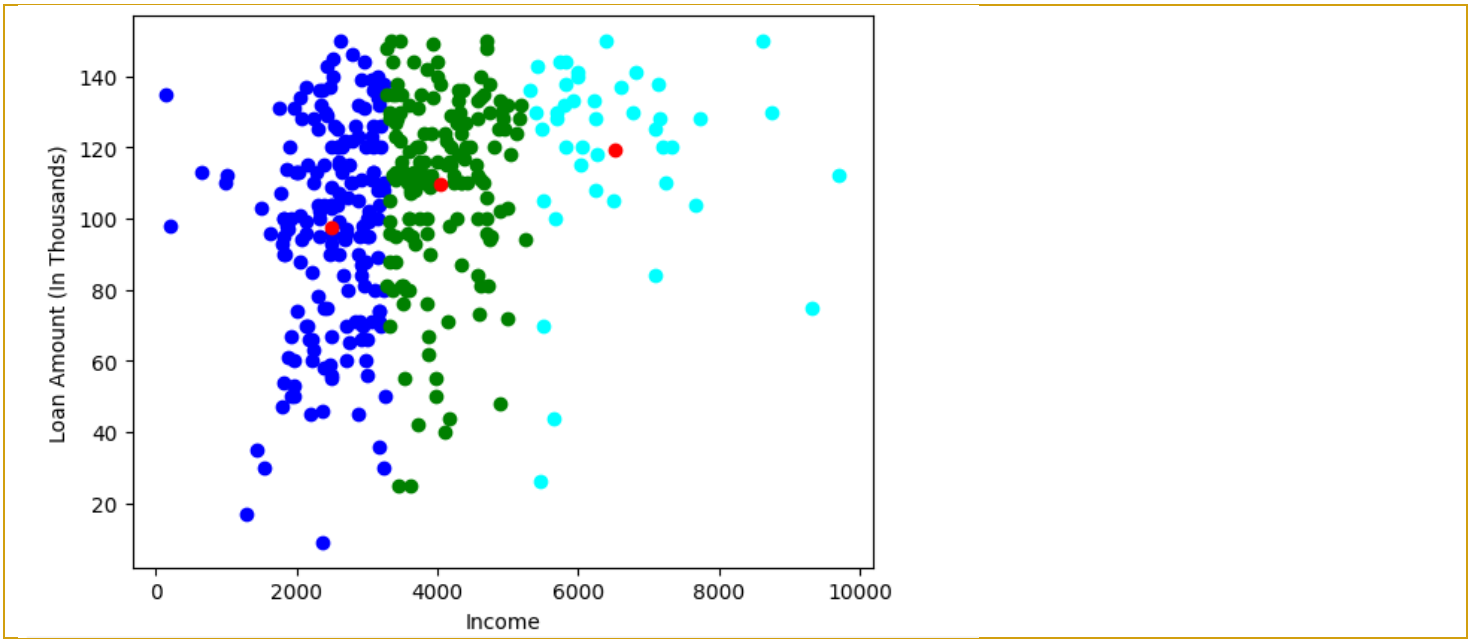


```
diff = 1
j=0

while(diff!=0):
    XD=X
    i=1
    for index1,row_c in Centroids.iterrows():
        ED=[]
        for index2,row_d in XD.iterrows():
            d1=(row_c["ApplicantIncome"]-row_d["ApplicantIncome"])**2
            d2=(row_c["LoanAmount"]-row_d["LoanAmount"])**2
            d=np.sqrt(d1+d2)
            ED.append(d)
        X[i]=ED
        i=i+1

    C=[]
    for index,row in X.iterrows():
        min_dist=row[1]
        pos=1
        for i in range(K):
            if row[i+1] < min_dist:
                min_dist = row[i+1]
                pos=i+1
        C.append(pos)
    X["Cluster"]=C
    Centroids_new = X.groupby(["Cluster"]).mean()[["LoanAmount","ApplicantIncome"]]
    if j == 0:
        diff=1
        j=j+1
    else:
        diff = (Centroids_new['LoanAmount'] - Centroids['LoanAmount']).sum() +
(Centroids_new['ApplicantIncome'] - Centroids['ApplicantIncome']).sum()
        print(diff.sum())
    Centroids = X.groupby(["Cluster"]).mean()[["LoanAmount","ApplicantIncome"]]

# Step#6:
#Visualize the clusters
color=['blue','green','cyan']
for k in range(K):
    data=X[X["Cluster"]==k+1]
    plt.scatter(data["ApplicantIncome"],data["LoanAmount"],c=color[k])
plt.scatter(Centroids["ApplicantIncome"],Centroids["LoanAmount"],c='red')
plt.xlabel('Income')
plt.ylabel('Loan Amount (In Thousands)')
plt.show()
```


Loan Amount (In Thousands)
140
120
100
80
60
40
20
0
2000
4000
6000
8000
10000
Income