Hierarchical Clustering

Instructions:

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

Name: KAVALI RAKESH YADAV_____ Batch ID: _
DSWDMCON 180122_____

Topic: Hierarchical Clustering

Grading Guidelines:

1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.

2. Assignments submitted after the deadline will affect your grades.

Grading:

Ans	Date			Ans	Date
Correct	On time	A	100		
80% & above	On time	B	85	Correct	Late
50% & above	On time	C	75	80% & above	Late
50% & below	On time	D	65	50% & above	Late
		E	55	50% & below	
Copied/No Submission		F	45		

- **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
- **Grade B: (>= 80 and < 90):**
  - When assignments are submitted on time but less than 80% of problems are completed.

    (OR)
  - All assignments are submitted after the deadline.

- **Grade C: (>= 70 and < 80):**
  - When assignments are submitted on time but less than 50% of the problems are completed.

    (OR)
  - Less than 80% of problems in the assignments are submitted after the deadline.

- **Grade D: (>= 60 and < 70):**
  - Assignments submitted after the deadline and with 50% or less problems.

- **Grade E: (>= 50 and < 60):**
  - Less than 30% of problems in the assignments are submitted after the deadline.

    (OR)
  - Less than 30% of problems in the assignments are submitted before the deadline.

- **Grade F: (< 50):** No submission (or) malpractice.

Hints:

1. Business Problem

1.1. What is the business objective?

1.2. Are there any constraints?

2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

4. Exploratory Data Analysis (EDA):

4.1. Summary.

4.2. Univariate analysis.

4.3. Bivariate analysis.

5. Model Building

5.1 Build the model on the scaled data (try multiple options).

5.2 Perform the hierarchical clustering and visualize the clusters using dendrogram.

5.3 Validate the clusters (try with different number of clusters) – label the clusters and derive insights (compare the results from multiple approaches).

6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?

Problem Statements:

1. Perform clustering for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained. Refer to EastWestAirlines.xlsx dataset.

	ID.	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award.
1	1	28143	0	1	1	1	174	1	0	0	7000	0
2	2	19244	0	1	1	1	215	2	0	0	6968	0
3	3	41354	0	1	1	1	4123	4	0	0	7034	0
4	4	14776	0	1	1	1	500	1	0	0	6952	0
5	5	97752	0	4	1	1	43300	26	2077	4	6935	1
6	6	16420	0	1	1	1	0	0	0	0	6942	0
7	7	84914	0	3	1	1	27482	25	0	0	6994	0
8	8	20856	0	1	1	1	5250	4	250	1	6938	1
9	9	443003	0	3	2	1	1753	43	3850	12	6948	1
10	10	104860	0	3	1	1	28426	28	1150	3	6931	1
11	11	40091	0	2	1	1	7278	10	0	0	6959	0
12	12	96522	0	5	1	1	61105	19	0	0	6924	1
13	13	43382	0	2	1	1	11150	20	0	0	6924	0
14	14	43097	0	1	1	1	3258	6	0	0	6918	0
15	15	17648	0	1	1	1	0	0	0	0	6912	0

Ans: 1. **Business Objective**: It is to maximize the usage of offers provided by airlines to the customer.

Constraints: Lack of analyzing the previous data of the customer usage of offers provided by the airlines

Name of feature	Description	Types	Relevance
id	Unique ID	Numeric	Irrelevant since it is only id it is not useful for analyzing the data
Balance	Number of miles eligible for award travel	Numeric	Relevant
Qual_miles	Number of miles counted as qualifying for Topflight status	Numeric	Relevant
cc1_miles	Number of miles earned with freq. flyer credit card in the past 12 months:	Categorical	Relevant
cc2_miles	Number of miles earned with Rewards credit card in the past 12 months:	Categorical	Relevant
cc3_miles	Number of miles earned with Small Business credit card in the past 12 months:	Categorical	Relevant

Bonus_miles	Number of miles earned from non-flight bonus transactions in the past 12 months	Numeric	Relevant
Bonus_trans	Number of non-flight bonus transactions in the past 12 months	Numeric	Relevant
flight_miles_12mo	Number of flight miles in the past 12 months	Numeric	Relevant
Flight_trans_12mo	Number of flight transactions in the past 12 months	Numeric	Relevant
Days_since_enroll	Number of days since Enroll_date	Numeric	Relevant
Award?	variable for Last_award	Numeric	Relevant

1. Data is cleaned by removing the columns which are not required for the analysis.
2. Since the data contains different quantities it is unable to do the visualization properly, so Normalization of the data is done in order to cover the whole data in the range of (0,1) so that the data can be visualized easily.
3. The given data is clustered by using Hierarchical clustering in python and the Dendrogram is plotted and observed the complete data in categories and then decided to convert the data into 3 clusters to identify the customers easily belonging to the particular cluster.
4. After Clustering the data it is easy to identify the which customer belongs to which category.

2. Perform clustering for the crime data and identify the number of clusters formed and draw inferences. Refer to crime_data.csv dataset.

	X	Murder	Assault	UrbanPop	Rape
1	Alabama	13.2	236	58	21.2
2	Alaska	10.0	263	48	44.5
3	Arizona	8.1	294	80	31.0
4	Arkansas	8.8	190	50	19.5
5	California	9.0	276	91	40.6
6	Colorado	7.9	204	78	38.7
7	Connecticut	3.3	110	77	11.1
8	Delaware	5.9	238	72	15.8
9	Florida	15.4	335	80	31.9
10	Georgia	17.4	211	60	25.8
11	Hawaii	5.3	46	83	20.2
12	Idaho	2.6	120	54	14.2
13	Illinois	10.4	249	83	24.0

Ans: (i) **Business Objective**: Identify the Murderer who has done more number of murders etc. to minimize the man work in identifying the murderer.
Constraints: Lack of Analyzing the persons previous data

(ii) Data type and its relevance

Name of feature	Description	Datatype	Relevance
x	Murdered name	Non numeric	Relevant
Murder	murder rate	Numeric	Relevant
Assault	assault rate	Numeric	Relevant
UrbanPop	urbanpop rate	Numeric	Relevant
Rape	rape rate	Numeric	Relevant

(iii) Since the complete data is required, there no requirement of cleaning the data
(iv) The quantities of the variable are different so in order to visualize it will be difficult so the given data is standardized in order to visualize.
(v) The given data is clustered by using Hierarchical clustering in R and python and the Dendrogram is plotted and observed the complete data in categories and then decided to convert the data into 4 clusters, so that it will be easy to identify the murderer easily.
(vi) After Clustering the data it is easy to identify the murderer who belongs to the particular category.

3. Perform clustering analysis on the telecom data set. The data is a mixture of both categorical and numerical data. It consists of the number of customers who churn out. Derive insights and get possible information on factors that may affect the churn decision. Refer to Telco_customer_churn.xlsx dataset.

Customer ID	Count	Quarter	Referred a Friend	Number of Referrals	Tenure in Months	Offer	Phone Service	Avg Monthly Long Distance Charges	Multiple Lines	Internet Service	Internet Type	Avg Monthly GB Download	Online Security	Online Backup	Device Protection	Premium Tech	Streaming TV	Streaming Movies	Streaming Music
8779-QRDMV	1	Q3	No	0	1	None	No	0	No	Yes	DSL	8	No	No	Yes	No	No	Yes	No
7495-OOKFY	1	Q3	Yes	1	8	Offer E	Yes	48.85	Yes	Yes	Fiber Optic	17	No	Yes	No	No	No	No	No
1658-BYGOY	1	Q3	No	0	18	Offer D	Yes	11.33	Yes	Yes	Fiber Optic	52	No	No	No	No	Yes	Yes	Yes
4598-XLKNJ	1	Q3	Yes	1	25	Offer C	Yes	19.76	No	Yes	Fiber Optic	12	No	Yes	Yes	No	Yes	Yes	No
4846-WHAFZ	1	Q3	Yes	1	37	Offer C	Yes	6.33	Yes	Yes	Fiber Optic	14	No	No	No	No	No	No	No
4412-YLTKF	1	Q3	No	0	27	Offer C	Yes	3.33	Yes	Yes	Fiber Optic	18	No	No	Yes	No	No	No	No
0390-DCFDQ	1	Q3	Yes	1	1	Offer E	Yes	15.28	No	Yes	Fiber Optic	30	No	No	No	No	No	No	No
3445-HXXGF	1	Q3	Yes	6	58	Offer B	No	0	No	Yes	DSL	24	No	Yes	Yes	No	No	Yes	No
2656-FMOKZ	1	Q3	No	0	15	Offer D	Yes	44.07	Yes	Yes	Fiber Optic	19	No	No	No	No	No	No	No
2070-FNEKE	1	Q3	No	0	7	Offer E	Yes	26.95	No	Yes	Fiber Optic	18	Yes	No	No	No	No	No	No

4. Perform clustering on mixed data. Convert the categorical variables to numeric by using dummies or label encoding and perform normalization techniques. The data set consists of details of customers related to their auto insurance. Refer to Autoinsurance.csv dataset.

Customer	State	Customer	Response	Coverage	Education	Effective To Date	Employme	Gender	Income	Location C	Marital Sta	Monthly Pr	Months Sir	Months Sir	Number o	Number o	Policy Typ	Policy	Renew Off	Sales Char	Total Clair	Vehicle Cla	Vehicle Siz
BU79786	Washingto	2763.519	No	Basic	Bachelor	2/24/2011	Employed	F	56274	Suburban	Married	69	32	5	0	1	Corporate	Corporate	Offer1	Agent	384.8111	Two-Door	Medsize
QZ44356	Arizona	6979.536	No	Extended	Bachelor	1/31/2011	Unemploy	F	0	Suburban	Single	94	13	42	0	8	Personal A	Personal L	Offer3	Agent	1131.465	Four-Door	Medsize
AI49188	Nevada	12887.43	No	Premium	Bachelor	2/19/2011	Employed	F	48767	Suburban	Married	108	18	38	0	2	Personal A	Personal L	Offer1	Agent	566.4722	Two-Door	Medsize
WW63253	California	7645.862	No	Basic	Bachelor	1/20/2011	Unemploy	M	0	Suburban	Married	106	18	65	0	7	Corporate	Corporate	Offer1	Call Cente	529.8813	SUV	Medsize
HB64268	Washingto	2813.693	No	Basic	Bachelor	3/2/2011	Employed	M	43836	Rural	Single	73	12	44	0	1	Personal A	Personal L	Offer1	Agent	138.1309	Four-Door	Medsize
OC83172	Oregon	8256.298	Yes	Basic	Bachelor	1/25/2011	Employed	F	62902	Rural	Married	69	14	94	0	2	Personal A	Personal L	Offer2	Web	159.383	Two-Door	Medsize
XZ87318	Oregon	5380.899	Yes	Basic	College	2/24/2011	Employed	F	55350	Suburban	Married	67	0	13	0	9	Corporate	Corporate	Offer1	Agent	321.6	Four-Door	Medsize
CF85061	Arizona	7216.1	No	Premium	Master	1/18/2011	Unemploy	M	0	Urban	Single	101	0	68	0	4	Corporate	Corporate	Offer1	Agent	363.0297	Four-Door	Medsize
DY87989	Oregon	24127.5	Yes	Basic	Bachelor	1/26/2011	Medical Le	M	14072	Suburban	Divorced	71	13	3	0	2	Corporate	Corporate	Offer1	Agent	511.2	Four-Door	Medsize
BQ94931	Oregon	7388.178	No	Extended	College	2/17/2011	Employed	F	28812	Urban	Married	93	17	7	0	8	Special Au	Special L2	Offer2	Branch	425.5278	Four-Door	Medsize
SX51350	California	4738.992	No	Basic	College	2/21/2011	Unemploy	M	0	Suburban	Single	67	23	5	0	3	Personal A	Personal L	Offer1	Agent	482.4	Four-Door	Small
VQ65197	California	8197.197	No	Basic	College	6/1/2011	Unemploy	F	0	Suburban	Married	110	27	87	0	3	Personal A	Personal L	Offer2	Agent	528	SUV	Medsize
DP39365	California	8798.797	No	Premium	Master	6/2/2011	Employed	M	77026	Urban	Married	110	9	82	2	3	Corporate	Corporate	Offer2	Agent	472.0297	Four-Door	Medsize
SJ95423	Arizona	8819.019	Yes	Basic	High Schoo	10/1/2011	Employed	M	99845	Suburban	Married	110	23	25	1	8	Corporate	Corporate	Offer2	Branch	528	SUV	Medsize