

SPRING END SEMESTER EXAMINATION-2023

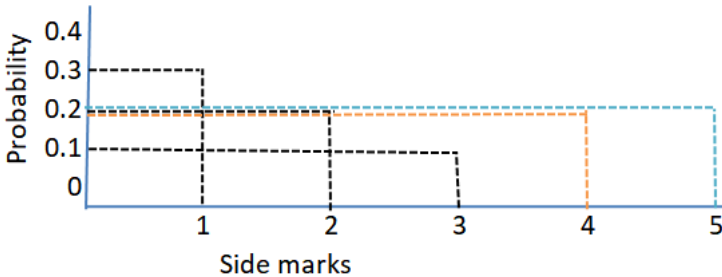
6th Semester, B.Tech

DATA ANALYTICS (IT-3006)

Evaluation Scheme and Solution

1.	Answer the following questions.																
(a)	<p>Explain the similarity and difference between JSON and BSON with suitable examples.</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. 0.5 mark for similarity and 0.5 for difference. No step-wise mark to be awarded.</p> <p><b>[Solution]</b></p> <p>Similarity: Both represent semi-structured format.</p> <p>Difference: BSON is not in a readable format wherein JSON is readable format.</p>																
(b)	<p>What is the difference between univariate, bivariate, and multivariate analysis?</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. Step-wise mark should be awarded based on the partial correctness of the solution.</p> <p><b>[Solution]</b></p> <p><u>Univariate</u> represents the type of data that consists of only one variable and its analysis involves central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.</p> <p><u>Bivariate</u> represents the type of data that consists of two variables and its analysis involves comparisons, relationships, causes and explanations.</p> <p><u>Multivariate</u> represents the type of data that consists of more than two variables and its analysis involves regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).</p>																
(c)	<p>Consider the below dataset that contains the number of hours of studies and the actual score received for 3 students in data analytics, and the predicted score was calculated with linear regression. Calculate <math>R^2</math>.</p> <table><tr><th>#</th><th>Number of hrs</th><th>Actual score</th><th>Predicted score</th></tr><tr><td>1</td><td>2</td><td>74</td><td>72</td></tr><tr><td>2</td><td>3</td><td>80</td><td>83</td></tr><tr><td>3</td><td>4</td><td>76</td><td>79</td></tr></table> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. No step-wise mark to be awarded.</p> <p><b>[Solution]</b></p> <p>Mean of actual score = 76.66 which is rounded to 77</p> <p><math>SSR = \text{Sum of squares regression} = (72 - 77)^2 + (83 - 77)^2 + (79 - 77)^2 = 25 + 36 + 4 = 65</math></p> <p><math>SSE = \text{Sum of squares error} = (72 - 74)^2 + (83 - 80)^2 + (79 - 76)^2 = 4 + 9 + 9 = 22</math></p>	#	Number of hrs	Actual score	Predicted score	1	2	74	72	2	3	80	83	3	4	76	79
#	Number of hrs	Actual score	Predicted score														
1	2	74	72														
2	3	80	83														
3	4	76	79														

		<p>SST = Sum of squares total = SSR + SSE = 65 + 22 = 87</p> <p><math>R^2 = SSR / SST = 65/87 = 0.747</math> and such value indicate moderately fit model.</p>																																													
	(d)	<p>A time series model is mathematically represented as <math>Y_t = f(T_t, S_t, C_t, I_t)</math> where <math>Y_t</math> is the time series value at time t. <math>T_t</math>, <math>S_t</math>, <math>C_t</math>, and <math>I_t</math> are the trend, seasonal, cyclic and irregular component value at time t respectively. Represents the model</p> <p>(1) When the amplitude of seasonal and irregular variations does not change as the level of trend rises or falls.</p> <p>(2) When the amplitude of both the seasonal and irregular variations increase as the level of trend rises.</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. 0.5 mark for 1<sup>st</sup> part and 0.5 for other part. No step-wise mark to be awarded.</p> <p><b>[Solution]</b></p> <p>(1) When the amplitude of seasonal and irregular variations does not change as the level of trend rises or falls, time series follows additive model and it is represented by <math>Y_t = T_t + S_t + C_t + I_t</math></p> <p>(2) When the amplitude of both the seasonal and irregular variations increase as the level of trend rises, time series follows multiplicative model and it is represented by <math>Y_t = T_t * S_t * C_t * I_t</math></p>																																													
	(e)	<p>Suppose a hierarchical clustering to be applied in segmenting the students and following sample has been collected. Create the proximity matrix for the below sample. The mark is out of 20 in the mid semester.</p> <table><tr><th>Roll No</th><th>Sex</th><th>Section</th><th>Mark</th></tr><tr><td>1</td><td>Male</td><td>CSE -1</td><td>10</td></tr><tr><td>2</td><td>Female</td><td>IT – 1</td><td>17</td></tr><tr><td>3</td><td>Male</td><td>CSSE – 1</td><td>18</td></tr><tr><td>4</td><td>Female</td><td>CSCE - 1</td><td>20</td></tr></table> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. No step-wise mark to be awarded.</p> <p><b>[Solution]</b></p> <p>Since the students dataset have 4 observations, so a 4 X 4 proximity matrix is to be created wherein the diagonal elements is 0 as the distance of a point with itself is always 0. Applying Euclidean distance formula, the matrix looks as follows:</p> <table><tr><th>Roll No</th><th>1</th><th>2</th><th>3</th><th>4</th></tr><tr><td>1</td><td>0</td><td><math>\sqrt{(10-17)^2} = 7</math></td><td><math>\sqrt{(10-18)^2} = 8</math></td><td><math>\sqrt{(10-20)^2} = 10</math></td></tr><tr><td>2</td><td><math>\sqrt{(17-10)^2} = 7</math></td><td>0</td><td><math>\sqrt{(17-18)^2} = 1</math></td><td><math>\sqrt{(17-20)^2} = 3</math></td></tr><tr><td>3</td><td><math>\sqrt{(18-10)^2} = 8</math></td><td><math>\sqrt{(18-17)^2} = 1</math></td><td>0</td><td><math>\sqrt{(18-20)^2} = 2</math></td></tr><tr><td>4</td><td><math>\sqrt{(20-10)^2} = 10</math></td><td><math>\sqrt{(20-17)^2} = 3</math></td><td><math>\sqrt{(20-18)^2} = 2</math></td><td>0</td></tr></table>	Roll No	Sex	Section	Mark	1	Male	CSE -1	10	2	Female	IT – 1	17	3	Male	CSSE – 1	18	4	Female	CSCE - 1	20	Roll No	1	2	3	4	1	0	$\sqrt{(10-17)^2} = 7$	$\sqrt{(10-18)^2} = 8$	$\sqrt{(10-20)^2} = 10$	2	$\sqrt{(17-10)^2} = 7$	0	$\sqrt{(17-18)^2} = 1$	$\sqrt{(17-20)^2} = 3$	3	$\sqrt{(18-10)^2} = 8$	$\sqrt{(18-17)^2} = 1$	0	$\sqrt{(18-20)^2} = 2$	4	$\sqrt{(20-10)^2} = 10$	$\sqrt{(20-17)^2} = 3$	$\sqrt{(20-18)^2} = 2$	0
Roll No	Sex	Section	Mark																																												
1	Male	CSE -1	10																																												
2	Female	IT – 1	17																																												
3	Male	CSSE – 1	18																																												
4	Female	CSCE - 1	20																																												
Roll No	1	2	3	4																																											
1	0	$\sqrt{(10-17)^2} = 7$	$\sqrt{(10-18)^2} = 8$	$\sqrt{(10-20)^2} = 10$																																											
2	$\sqrt{(17-10)^2} = 7$	0	$\sqrt{(17-18)^2} = 1$	$\sqrt{(17-20)^2} = 3$																																											
3	$\sqrt{(18-10)^2} = 8$	$\sqrt{(18-17)^2} = 1$	0	$\sqrt{(18-20)^2} = 2$																																											
4	$\sqrt{(20-10)^2} = 10$	$\sqrt{(20-17)^2} = 3$	$\sqrt{(20-18)^2} = 2$	0																																											
	(f)	<p>Consider the following dataset, wherein TID represents transaction ID and G to O represents individual products. In the dataset, 1 represents a</p>																																													

	<p>transaction that includes the specific products. For instance, TID 1 includes all products and TID 3 includes M and O product. Calculate <math>\text{Confidence}(\{G, A\} \Rightarrow \{M\})</math>.</p> <table><tr><th>TID</th><th>G</th><th>A</th><th>M</th></tr><tr><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>2</td><td>1</td><td>0</td><td>1</td></tr><tr><td>3</td><td>0</td><td>0</td><td>1</td></tr><tr><td>4</td><td>0</td><td>1</td><td>0</td></tr><tr><td>5</td><td>1</td><td>1</td><td>1</td></tr><tr><td>6</td><td>1</td><td>1</td><td>0</td></tr></table> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. No step-wise mark to be awarded.</p> <p><b>[Solution]</b> <math>\text{Confidence}(\{G, A\} \Rightarrow \{M\}) = \text{Support}(G, A, M) / \text{Support}(G, A) = [2/6] / [3/6] = 0.667</math></p>	TID	G	A	M	1	1	1	1	2	1	0	1	3	0	0	1	4	0	1	0	5	1	1	1	6	1	1	0
TID	G	A	M																										
1	1	1	1																										
2	1	0	1																										
3	0	0	1																										
4	0	1	0																										
5	1	1	1																										
6	1	1	0																										
(g)	<p>Consider the decagon, which has 10 sides. Three sides are marked 1, two sides are marked 2, one side is marked 3, two sides are marked 4, and two sides are marked 5. Draw a graph representing occurrence of each mark verses its probability.</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. No step-wise mark to be awarded.</p> <p><b>[Solution]</b> Probability(side marked as 1) = <math>3 / 10 = 0.3</math> Probability(side marked as 2) = <math>2/10 = 0.2</math> Probability(side marked as 3) = <math>1 / 10 = 0.1</math> Probability(side marked as 4) = <math>2/10 = 0.2</math> Probability(side marked as 5) = <math>2/10 = 0.2</math> The graph looks as follows:</p> 																												
(h)	<p>Consider the following dataset. Consider support count is represented with SC. Calculate <math>(\text{SC}(\{E\}) + \text{SC}(\{A, B\}) + \text{SC}(\{C, D\})) / (\text{SC}(\{A, B, C, E\}) + \text{SC}(\{A, B, C, D, E\}))</math></p> <table><tr><th>Transaction</th><th>Itemset</th></tr><tr><td>T1</td><td>A, B</td></tr><tr><td>T2</td><td>B, D</td></tr><tr><td>T3</td><td>B, C</td></tr><tr><td>T4</td><td>A, B, D</td></tr><tr><td>T5</td><td>A, C</td></tr><tr><td>T6</td><td>B, C</td></tr><tr><td>T7</td><td>A, B, C, E</td></tr></table>	Transaction	Itemset	T1	A, B	T2	B, D	T3	B, C	T4	A, B, D	T5	A, C	T6	B, C	T7	A, B, C, E												
Transaction	Itemset																												
T1	A, B																												
T2	B, D																												
T3	B, C																												
T4	A, B, D																												
T5	A, C																												
T6	B, C																												
T7	A, B, C, E																												

		<p><b>[Evaluation Scheme]</b> Full mark for the correct answer. No step-wise mark to be awarded.</p> <p><b>[Solution]</b>  <math>SC(\{E\}) = 1</math>, <math>SC(\{A, B\}) = 3</math>, <math>SC(\{C, D\}) = 0</math>, <math>SC(\{A, B, C, E\}) = 1</math>, and <math>SC(\{A, B, C, D, E\}) = 0</math>  Numerator = <math>(SC(\{E\}) + SC(\{A, B\}) + SC(\{C, D\})) = 1 + 3 + 0 = 4</math>  Denominator = <math>SC(\{A, B, C, E\}) + SC(\{A, B, C, D, E\}) = 1 + 0 = 1</math>  Therefore, <math>(SC(\{E\}) + SC(\{A, B\}) + SC(\{C, D\})) / (SC(\{A, B, C, E\}) + SC(\{A, B, C, D, E\})) = 4/1 = 4</math></p>
	(i)	<p>A bloom filter with a size of 1000 slots is used to store the information of 100 data stream items using 4 hash functions. Calculate the false positive probability.</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. No step-wise mark to be awarded.</p> <p><b>[Solution]</b>  <math>n</math> = size of bloom filter = 1000  <math>m</math> = number of expected elements to be inserted = 100  <math>k</math> = number of hash functions = 4</p> $\text{False positive probability} = \left(1 - \left(\frac{1}{e}\right)^{\frac{km}{n}}\right)^k$ <p><math>(1/e)^{km/n} = (1/2.718)^{4*100/1000} = 0.670</math>. So, <math>(1-0.670)^4 = 0.329^4 = 0.011</math></p>
	(j)	<p>What is the probability that a slot is hashed in a bloom filter where <math>n</math> is the size and <math>k</math> is the number of hash functions?</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. No step-wise mark to be awarded.</p> <p><b>[Solution]</b>  Probability that a slot is hashed with one hash function = <math>1/n</math>, so with <math>k</math> hash functions it is <math>1/n^k</math></p>

2. (a) Consider the following dataset. Draw the MapReduce process to find the number of customers from each city followed by each state, both in the chronological order.

ID	Name	City	State
1	Sujay Lila	Ambikapur	Chhattisgarh
2	Geetha Choudhary	Bhilai	Chhattisgarh
3	Anandi D'Cruz	Bilaspur	Chhattisgarh
4	Surendra Nagarkar	Cuttack	Odisha
5	Balwinder Nagarkar	Bangalore	Karnataka
6	Nitin Nibhanupudi	Mangalore	Karnataka
7	Dinesh Sharma	Cuttack	Odisha
8	Raj Chaudhri	Bilaspur	Chhattisgarh

		9	Govind Kumar	Mysore	Karnataka
		10	Jayanta Begam	Ambikapur	Chhattisgarh

**[Evaluation Scheme]** Full mark for the correct answer. 2 marks for city and 2 marks for state. Step-wise mark can be awarded based on the partial correctness of the solution.

**[Solution]**

MapReduce process to find the number of customers from each city:

```

graph LR
    subgraph Input
        A[Ambikapur]
        B[Bhilai]
        C[Bilaspur]
        D[Cuttack]
        E[Bangalore]
        F[Mangalore]
        G[Cuttack]
        H[Bilaspur]
        I[Mysore]
        J[Ambikapur]
    end
    subgraph K1_V1_Splitting
        A1[Ambikapur]
        B1[Bhilai]
        C1[Bilaspur]
        D1[Cuttack]
        E1[Bangalore]
        F1[Mangalore]
        G1[Cuttack]
        H1[Bilaspur]
        I1[Mysore]
        J1[Ambikapur]
    end
    subgraph List_K2_V2_Mapping
        A2[Ambikapur,1]
        B2[Bhilai,1]
        C2[Bilaspur,1]
        D2[Cuttack,1]
        E2[Bangalore,1]
        F2[Mangalore,1]
        G2[Cuttack,1]
        H2[Bilaspur,1]
        I2[Mysore,1]
        J2[Ambikapur,1]
    end
    subgraph K2_List_V2_Shuffling
        A3[Ambikapur, (1,1)]
        B3[Bangalore,1]
        C3[Bhilai,1]
        D3[Bilaspur, (1,1)]
        E3[Cuttack, (1,1)]
        F3[Mangalore,1]
        G3[Mysore,1]
    end
    subgraph List_K3_V3_Reducing
        A4[Ambikapur,2]
        B4[Bangalore,1]
        C4[Bhilai,1]
        D4[Bilaspur,2]
        E4[Cuttack,2]
        F4[Mangalore,1]
        G4[Mysore,1]
    end
    subgraph Customer_V4_Result
        R[Customer, 10]
    end
    A --> A1 --> A2 --> A3 --> A4 --> R
    B --> B1 --> B2 --> B3 --> B4 --> R
    C --> C1 --> C2 --> C3 --> D4 --> R
    D --> D1 --> D2 --> D3 --> E4 --> R
    E --> E1 --> E2 --> E3 --> B4 --> R
    F --> F1 --> F2 --> F3 --> F4 --> R
    G --> G1 --> G2 --> G3 --> E4 --> R
    H --> H1 --> H2 --> H3 --> D4 --> R
    I --> I1 --> I2 --> I3 --> G4 --> R
    J --> J1 --> J2 --> J3 --> A4 --> R
  
```

MapReduce process to find the number of customers from each state:

```

graph LR
    subgraph Input
        A[Chhattisgarh]
        B[Chhattisgarh]
        C[Chhattisgarh]
        D[Odisha]
        E[Karnataka]
        F[Karnataka]
        G[Odisha]
        H[Chhattisgarh]
        I[Karnataka]
        J[Chhattisgarh]
    end
    subgraph K1_V1_Splitting
        A1[Chhattisgarh]
        B1[Chhattisgarh]
        C1[Chhattisgarh]
        D1[Odisha]
        E1[Karnataka]
        F1[Karnataka]
        G1[Odisha]
        H1[Chhattisgarh]
        I1[Karnataka]
        J1[Chhattisgarh]
    end
    subgraph List_K2_V2_Mapping
        A2[Chhattisgarh,1]
        B2[Chhattisgarh,1]
        C2[Chhattisgarh,1]
        D2[Odisha,1]
        E2[Karnataka,1]
        F2[Karnataka,1]
        G2[Odisha,1]
        H2[Chhattisgarh,1]
        I2[Karnataka,1]
        J2[Chhattisgarh,1]
    end
    subgraph K2_List_V2_Shuffling
        A3[Chhattisgarh, (1,1,1,1,1)]
        B3[Karnataka,(1,1,1)]
        C3[Odisha, (1,1)]
    end
    subgraph List_K3_V3_Reducing
        A4[Chhattisgarh,5]
        B4[Karnataka,3]
        C4[Odisha,3]
    end
    subgraph Customer_V4_Result
        R[Customer, 10]
    end
    A --> A1 --> A2 --> A3 --> A4 --> R
    B --> B1 --> B2 --> B3 --> B4 --> R
    C --> C1 --> C2 --> C3 --> A4 --> R
    D --> D1 --> D2 --> D3 --> C4 --> R
    E --> E1 --> E2 --> E3 --> B4 --> R
    F --> F1 --> F2 --> F3 --> B4 --> R
    G --> G1 --> G2 --> G3 --> C4 --> R
    H --> H1 --> H2 --> H3 --> A4 --> R
    I --> I1 --> I2 --> I3 --> B4 --> R
    J --> J1 --> J2 --> J3 --> A4 --> R
  
```

(b) A retail company wants to enhance their customer experience by analysing the customer reviews for different products, so that they can inform the corresponding vendors and manufacturers about the product defects and shortcomings. You have been tasked to analyse the complaints filed under each product & the total number of complaints filed based on the geography, type of product, etc. You also have to figure out the complaints which have no timely response. Discuss and then model your views concerning descriptive, diagnostic and predictive analytics.

**[Evaluation Scheme]** Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.

**[Solution]**

Descriptive analytics model – the model should use historical and current data to seek answer for the questions “what has been happened” using data analytics technique such as box plot. Few examples may be as follows:

- (1) Find the number of complaints by geography and type of product
- (2) Which geography contributed maximum number of negative review comments?
- (3) Which product type has maximum number of positive review

		<p>comments?</p> <p>Diagnostic analytics model – the model should use historical and current data to seek answer for the questions “why it has been happened” using data analytics technique such as drill-through, root cause analysis using fish bone, etc. Few examples may be as follows:</p> <ol style="list-style-type: none"> <li>(1) Why is the number of complaints by Asian geography and food and beverages product</li> <li>(2) Why male of Asian geography provided maximum number of negative review comments?</li> <li>(3) Why the suitable features of beauty care product type are has collected maximum number of positive review comments?</li> </ol> <p>Predictive analytics model – the model should use historical and current data to seek answer for the questions “what will happen in the” using data analytics technique such as regression, clustering, classifications etc. Few examples may be as follows:</p> <ol style="list-style-type: none"> <li>(1) What would be the total number of complaints by Asian geography and food and beverages product by the end of this quarter?</li> <li>(2) What is expected number of negative review comments from Asian geography by end of this month?</li> <li>(3) What would be the sale by end of this year?</li> </ol>
--	--	--

3.	(a)	<p>In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists want to test a new medication to see if it has either a positive or negative effect on intelligence or not effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Using hypothesis testing, find the answer to the question i.e., did the medication affect intelligence? The z value (i.e., critical value) from statistical table is found to be 1.96. The solution must mention the null (<math>H_0</math>) and alternative hypotheses (<math>H_a</math>).</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.</p> <p><b>[Solution]</b></p> <p><b>Step 1:</b> Set up the null and alternate hypothesis</p> <p><math>H_0</math>: medication affects intelligence  <math>H_a</math>: medication does not affect intelligence.</p> <p><b>Step 2:</b> Determine the type of test to use          Since the sample size is 30, the z-test is used.</p> <p><b>Step 3:</b> Calculate the tested statistic z using the formula</p>
----	-----	--

$$Z = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}$$

Where  $\bar{x}_n$  is the mean of the population,  $\mu_0$  is the null hypothesis (i.e., the mean) to be tested,  $\sigma$  is the standard deviation, and  $n$  is the sample size.

Using the data given in the equation we would have the following:

$\mu_0 = 100$ ,  $\sigma = 15$ ,  $n = 30$ ,  $\bar{x}_n = 140$

Plugging the values into the formula:  $((140 - 100) / 15) * \sqrt{30} = 14.606$

**Step 4:** In the question,  $z$  value is provided i.e., 1.96 and hence no need to look into  $z$  table.

**Step 5:** drawing conclusion

The tested statistic value of  $z$  calculated is more than the critical value obtained from statistical tables (i.e.,  $14.606 > 1.96$ ). Therefore the null hypothesis is rejected. This means that the medication administered does not affect intelligence.

- (b) Find the relationships of salary between millennials (between the ages of 18 and 34), gen X (between the ages of 35 and 50) and baby boomers (aged 51 and above) of below sample by plotting multiple boxplots in one graph.

Gender	Age	Salary
Male	20	81600
Female	55	61600
Male	38	64300
Female	25	71900
Male	58	76300
Male	45	68200
Female	30	60900
Female	49	78600
Male	60	81700

**[Evaluation Scheme]** Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.

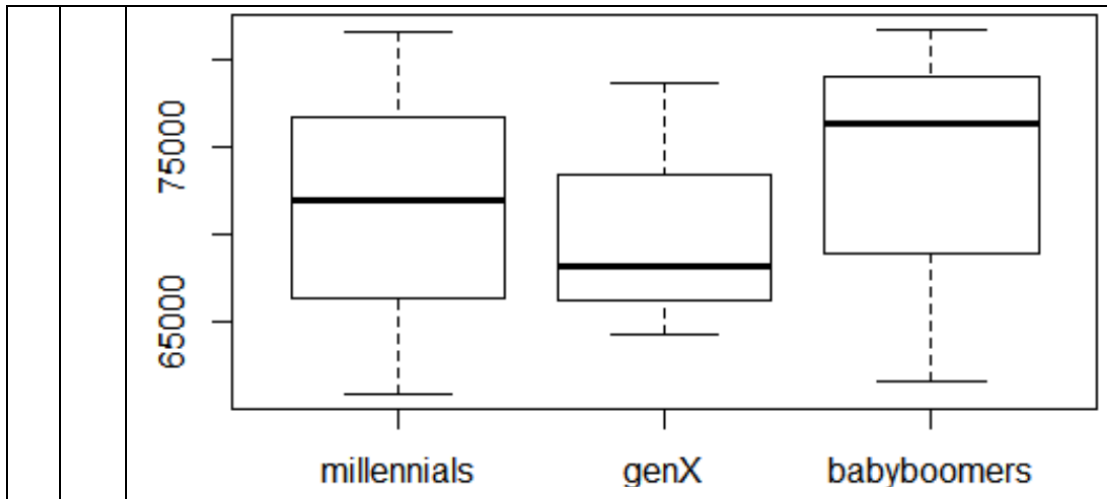
**[Solution]**

The data points (i.e., salary) for millennials = {81600, 71900, 60900}

The data points (i.e., salary) for gen X = {64300, 68200, 78600}

The data points (i.e., salary) for baby boomers = {61600, 76300, 81700}

Plotting multiple boxplots in one graph infers visualizing millennials, gen X and baby boomers boxplots side-by-side in the same graphic.



4. (a) A consumer electronics company has adopted an aggressive policy to increase sales of a newly launched product. The company has invested in advertisements as well as employed salesmen for increasing sales rapidly. Below dataset presents the sales, the number of employed salesmen, and advertisement expenditure for 4 randomly selected months. Develop a regression model to predict the impact of advertisement and the number of salesmen on sales.

Month No	1	2	3	4
Sales	5000	5200	5700	6300
Salesmen	25	35	15	27
Advertisement	180	250	150	240

**[Evaluation Scheme]** Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.

**[Solution]**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Where Y= Sales, x1=Salesmen , and x2=Advertisement. From the data in table we calculate,

Month No	Y	X1	X2
1	5000	25	180
2	5200	35	250
3	5700	15	150
4	6300	27	240

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2$$

$$\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Month No	Y	X1	X2		X1*X1	X2*X2	X1X2	X1X2*X1X2	X1Y	X2Y
1	5000	25	180		625	32400	4500	20250000	125000	900000
2	5200	35	250		1225	62500	8750	76562500	182000	1300000
3	5700	15	150		225	22500	2250	5062500	85500	855000
4	6300	27	240		729	57600	6480	41990400	170100	1512000
Average	5550	25.5	205	SUM	2804	175000	21980	143865400	562600	4567000

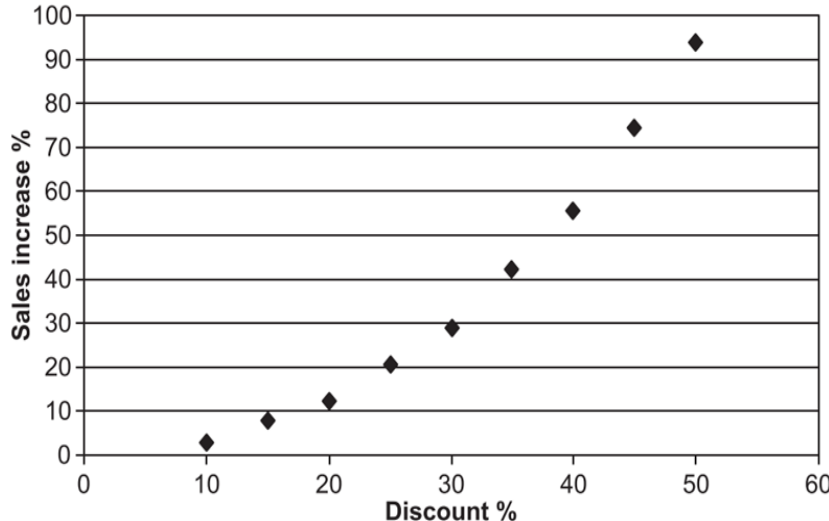
$$\beta_1 = -5.558$$

$$\beta_2 = 1.268$$

$$\beta_0 = 5431.789$$

$$Y = 5431.789 - 5.558 * X_1 + 1.268 * X_2$$



	(b)	<p>Explain non-linear regression with a suitable example. Subsequently, establish narrate second degree (quadratic), third degree (cubic) and n degree polynomial mathematical model. In general, what techniques applied to determine the right degree of the model?</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.</p> <p><b>[Solution]</b></p> <p>In the case of linear and multiple linear regression, the dependent variable is linearly dependent on the independent variable(s). But, in several situations, the situation is no simple where the two variables might be related in a non-linear way. This may be the case where the results from the correlation analysis show no linear relationship but these variables might still be closely related. If the result of the data analysis shows that there is a non-linear (also known as curvilinear) association between the two variables, then the need is to develop a non-linear regression model. Imagine a dataset whose scatter plot looks as follows:</p> <div><table><caption>Data points from the scatter plot</caption><tr><th>Discount %</th><th>Sales increase %</th></tr><tr><td>10</td><td>5</td></tr><tr><td>15</td><td>8</td></tr><tr><td>20</td><td>12</td></tr><tr><td>25</td><td>20</td></tr><tr><td>30</td><td>28</td></tr><tr><td>35</td><td>42</td></tr><tr><td>40</td><td>55</td></tr><tr><td>45</td><td>75</td></tr><tr><td>50</td><td>95</td></tr></table></div> <p>The non-linear data can be handled in 2 ways:</p> <ul style="list-style-type: none"><li>• Use of polynomial rather than linear regression model</li><li>• Transform the data and then use linear regression model.</li></ul> <p>The polynomial mathematical model are represented below:</p> <p>Second degree: <math>y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e</math></p> <p>Third degree: <math>y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e</math></p> <p>n degree: <math>y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots \dots + \beta_n x_n + e</math></p> <p>To determine the right degree of the model, 2 approaches are followed:</p> <p><b>Forward Selection:</b> This method increases the degree until it is significant</p>	Discount %	Sales increase %	10	5	15	8	20	12	25	20	30	28	35	42	40	55	45	75	50	95
Discount %	Sales increase %																					
10	5																					
15	8																					
20	12																					
25	20																					
30	28																					
35	42																					
40	55																					
45	75																					
50	95																					

	<p>enough to define the best possible model.</p> <p><b>Backward Elimination:</b> This method decreases the degree until it is significant enough to define the best possible model.</p>
--	---

5.	<p>(a) Consider the following dataset consisting of 6 observations that depicts automobile battery sales. Using Simple Exponential Smoothing, calculate the forecasted value of month 7 by calculating smooth observation (<math>S_t</math>) for each month and mean of the squared errors. The smoothing constant is 0.5 and <math>S_1</math> value is 20.</p> <table><tr><th>Month No</th><th>Actual</th></tr><tr><td>1</td><td>20</td></tr><tr><td>2</td><td>22</td></tr><tr><td>3</td><td>21</td></tr><tr><td>4</td><td>18</td></tr><tr><td>5</td><td>17</td></tr><tr><td>6</td><td>23</td></tr></table> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.</p> <p><b>[Solution]</b></p> <p><math>S_t = \alpha * Y_{t-1} + (1-\alpha) * S_{t-1}</math> where <math>\alpha = 0.5</math> and <math>S_1 = 20</math></p> <table><tr><th>Month</th><th>Actual (<math>Y_t</math>)</th><th>Forecast (<math>S_t</math>)</th><th>Err</th><th>Sq-Err</th></tr><tr><td>1</td><td>20</td><td>20</td><td>0</td><td>0</td></tr><tr><td>2</td><td>22</td><td>20</td><td>2</td><td>4</td></tr><tr><td>3</td><td>21</td><td>21</td><td>0</td><td>0</td></tr><tr><td>4</td><td>18</td><td>21</td><td>-3</td><td>9</td></tr><tr><td>5</td><td>17</td><td>19.5</td><td>-2.5</td><td>6.25</td></tr><tr><td>6</td><td>23</td><td>17.26</td><td>5.74</td><td>32.94</td></tr><tr><td>7</td><td></td><td>19.14</td><td></td><td></td></tr></table> <p>Sum of Square errors = 52.19</p> <p>Mean Square error = <math>52.19/6 = 8.698</math></p> <p><math>S_2 = 0.5 * 20 + 0.5 * 20 = 10 + 10 = 20</math></p> <p><math>S_3 = 0.5 * 22 + 0.5 * 20 = 11 + 10 = 21</math></p> <p><math>S_4 = 0.5 * 21 + 0.5 * 21 = 10.5 + 10.5 = 21</math></p> <p><math>S_5 = 0.5 * 18 + 0.5 * 21 = 9 + 10.5 = 19.5</math></p> <p><math>S_6 = 0.5 * 17 + 0.5 * 19.5 = 8.5 + 9.75 = 17.26</math></p> <p><math>S_7 = 0.5 * 23 + 0.5 * 17.26 = 11.5 + 8.63 = 19.14</math></p>	Month No	Actual	1	20	2	22	3	21	4	18	5	17	6	23	Month	Actual ( $Y_t$ )	Forecast ( $S_t$ )	Err	Sq-Err	1	20	20	0	0	2	22	20	2	4	3	21	21	0	0	4	18	21	-3	9	5	17	19.5	-2.5	6.25	6	23	17.26	5.74	32.94	7		19.14		
Month No	Actual																																																						
1	20																																																						
2	22																																																						
3	21																																																						
4	18																																																						
5	17																																																						
6	23																																																						
Month	Actual ( $Y_t$ )	Forecast ( $S_t$ )	Err	Sq-Err																																																			
1	20	20	0	0																																																			
2	22	20	2	4																																																			
3	21	21	0	0																																																			
4	18	21	-3	9																																																			
5	17	19.5	-2.5	6.25																																																			
6	23	17.26	5.74	32.94																																																			
7		19.14																																																					
	<p>(b) Consider the following dataset capturing monthly sales of actual vs. predicted of an Indian B2C (business to customer) firm. The sales figures are in lakh and presented in INR.</p>																																																						

Month No	1	2	3	4
Actual	112	113	122	120
Predicted	113	115	121	119

As a data consultant, the B2C firm hires you for the following and you need to justify your response.

(1) Determine the hybrid error and a hybrid error is determined by  $0.3 * \text{MSE} + 0.25 * \text{RMSE}$ .

(2) Determine MAPE.

**[Evaluation Scheme]** Full mark for the correct answer. 2 marks for hybrid error and rest 2 marks for MAPE calculation. Step-wise mark can be awarded based on the partial correctness of the solution.

**[Solution]**

The hybrid error calculation is as follows:

Month No	1	2	3	4
Actual	112	113	122	120
Predicted	113	115	121	119
Error	-1	-2	1	1
Squared Error	1	4	1	1

Sum of Square Error =  $1 + 4 + 1 + 1 = 7$   
Mean Square Error (MSE) =  $7 / 4 = 1.75$   
Root Mean Square Error (RMSE) =  $\sqrt{\text{MSE}} = \sqrt{1.75} = 1.322$   
So, hybrid error =  $0.3 * 1.75 + 0.25 * 1.322 = 0.855$

The MAPE calculation is as follows:

Month No	1	2	3	4
Actual	112	113	122	120
Predicted	113	115	121	119
Predicted – Actual	1	2	1	1
Predicted – Actual   / Actual	$1/112 = 0.008$	$2/113 = 0.017$	$1/122 = 0.008$	$1/120 = 0.008$

SUM(| Predicted – Actual | / Actual) =  $0.008 + 0.017 + 0.008 + 0.008 = 0.041$   
MAPE =  $(100/4) * 0.041 = 1.025$

6.	(a)	<p>Consider the following transactional data in which minimum support is 2 and minimum confidence is 50%. Find frequent itemsets and generate association rules for them by illustrating it with step-by-step process.</p> <table><tr><th>Transactions</th><th>List of items</th></tr><tr><td>T1</td><td>I1, I2, I3</td></tr><tr><td>T2</td><td>I2, I3, I4</td></tr><tr><td>T3</td><td>I4, I5</td></tr><tr><td>T4</td><td>I1, I2, I4</td></tr><tr><td>T5</td><td>I1, I2, I3, I5</td></tr><tr><td>T6</td><td>I1, I2, I3, I4</td></tr></table> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. Step-wise mark</p>	Transactions	List of items	T1	I1, I2, I3	T2	I2, I3, I4	T3	I4, I5	T4	I1, I2, I4	T5	I1, I2, I3, I5	T6	I1, I2, I3, I4
Transactions	List of items															
T1	I1, I2, I3															
T2	I2, I3, I4															
T3	I4, I5															
T4	I1, I2, I4															
T5	I1, I2, I3, I5															
T6	I1, I2, I3, I4															

can be awarded based on the partial correctness of the solution.

**[Solution]**

Table-1 :-

Transaction	List of items
T1	I1, I2, I3
T2	I2, I3, I4
T3	I4, I5
T4	I1, I2, I4
T5	I1, I2, I3, I5
T6	I1, I2, I3, I4

minimum confidence = 50%, minimum support = 2.

Step-1:- Count of each item.

Table-2 :-

Item	Count
I1	4
I2	5
I3	4
I4	4
I5	2

Step-2 (Prune step):-

Since all the items meet the min-sup count, thus no item is deleted

Step 3:- Join step :- 2-itemset are formed from Table-2

Step-4:-

From table-1 the count of 2-itemset occurrences is taken.

Table-3

Item	count
$I_1, I_2$	4
$I_1, I_3$	3
$I_1, I_4$	2
$I_2, I_3$	4
$I_2, I_4$	3
$I_1, I_5$	1
$I_2, I_5$	1
$I_3, I_4$	2
$I_3, I_5$	1

Table-3 shows.

$(I_1, I_5), (I_2, I_5), \&(I_3, I_5)$   
does not meet the  
min-sup, thus is deleted.

Table-4

Item	count
$I_1, I_2$	4
$I_1, I_3$	3
$I_1, I_4$	2
$I_2, I_3$	4
$I_2, I_4$	3
$I_3, I_4$	2

Step-5 3-itemset is formed and  
from table-1 the occurrences  
of the 3-itemset is found out.

Table-5

Item	Count
$I_1, I_2, I_3$	3
$I_1, I_2, I_4$	2
$I_1, I_3, I_4$	2
$I_1, I_3, I_4$	1

here since  $(I_1, I_2, I_4)$  does not satisfy the min-sup hence it is removed.

Step 6:- 4-itemset is formed from table-1 the occurrences is taken.

Since by join only one 4-itemset  $(I_1, I_2, I_3, I_4)$  is formed and it does not satisfy the min-sup  $\therefore$  the frequent 2-item are taken from table-5 itself they are

$(I_1, I_2, I_3)$ ,  $(I_1, I_2, I_4)$  and  $(I_1, I_3, I_4)$

Generation of Association Rules:-

From the above frequent item set we can obtain association as follows:-

$$\begin{aligned}
 1) \{I_1, I_2\} &\Rightarrow \{I_3\} \quad \text{conf} = \frac{\text{sup}\{I_1, I_2, I_3\}}{\text{sup}\{I_1, I_2\}} = \frac{3}{4} \times 100 = 75\% \\
 \{I_1, I_3\} &\Rightarrow \{I_2\} \quad \text{conf} = \frac{3}{3} \times 100 = 100\% \\
 \{I_2, I_3\} &\Rightarrow \{I_1\} \quad \text{conf} = \frac{3}{4} \times 100 = 75\% \\
 \{I_1\} &\Rightarrow \{I_2, I_3\} \quad \text{conf} = \frac{3}{4} \times 100 = 75\% \\
 \{I_2\} &\Rightarrow \{I_1, I_3\} \quad \text{conf} = \frac{3}{5} \times 100 = 60\% \\
 \{I_3\} &\Rightarrow \{I_1, I_2\} \quad \text{conf} = \frac{3}{4} \times 100 = 75\%
 \end{aligned}$$

from  $\{I_1, I_2, I_4\}$

$\{I_1, I_2\} \rightarrow \{I_4\}$  conf =  $\frac{2}{4} \times 100 = 50\%$   
 $\{I_1, I_4\} \rightarrow \{I_2\}$  conf =  $\frac{2}{2} \times 100 = 100\%$   
 $\{I_2, I_4\} \rightarrow \{I_1\}$  conf =  $\frac{2}{3} \times 100 = 66.66\%$   
 $\{I_1\} \rightarrow \{I_2, I_4\}$  conf =  $\frac{2}{4} \times 100 = 50\%$   
 $\{I_2\} \rightarrow \{I_1, I_4\}$  conf =  $\frac{2}{5} \times 100 = 40\%$   
 $\{I_4\} \rightarrow \{I_1, I_2\}$  conf =  $\frac{2}{4} \times 100 = 50\%$

from  $\{I_1, I_3, I_4\}$

$\{I_1, I_3\} \rightarrow \{I_4\}$  conf =  $\frac{2}{3} \times 100 = 66.66\%$   
 $\{I_1, I_4\} \rightarrow \{I_3\}$  conf =  $\frac{2}{2} \times 100 = 100\%$   
 $\{I_3, I_4\} \rightarrow \{I_1\}$  conf =  $\frac{2}{2} \times 100 = 100\%$   
 $\{I_1\} \rightarrow \{I_3, I_4\}$  conf =  $\frac{2}{2} \times 100 = 100\%$   
 $\{I_3\} \rightarrow \{I_1, I_4\}$  conf =  $\frac{2}{4} \times 100 = 50\%$   
 $\{I_4\} \rightarrow \{I_1, I_3\}$  conf =  $\frac{2}{4} \times 100 = 50\%$

This shows that except  $\{I_2\} \rightarrow \{I_1, I_4\}$  all other association rules are strong.

(b) Consider the following dataset.

Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apple	Cheese
3	Apple	Banana	
4	Milk	Cheese	
5	Apple	Banana	
6	Milk	Cheese	Banana

Calculate Support, Confidence and Lift for the followings:

- (1) Apple, Milk
- (2) (Apple, Milk)  $\Rightarrow$  Cheese
- (3) Milk  $\Rightarrow$  Cheese
- (4) (Apple, Cheese)  $\Rightarrow$  Milk

**[Evaluation Scheme]** Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.

**[Solution]**



The support formula written out would look something like:

$$\text{Support} = \frac{(A + B)}{\text{Total}}$$

$$\text{Support for Basket 1} = \frac{(\text{Milk} + \text{Cheese})}{\text{Total}} = \frac{6}{9} = .6666667$$

The confidence formula written out would like something like:

$$\text{Confidence} = \frac{(A + B)}{A}$$

$$\text{Confidence for Basket 1} = \frac{(\text{Milk} + \text{Cheese})}{\text{Milk}} = \frac{6}{6} = 1.000$$

The lift formula written out would look something like:

$$\text{Lift} = \left( \frac{\left( \frac{(A + B)}{A} \right)}{\left( \frac{B}{\text{Total}} \right)} \right)$$

$$\text{Lift for Basket 1} = \left( \frac{\left( \frac{(\text{Milk} + \text{Cheese})}{\text{Milk}} \right)}{\left( \frac{(\text{Cheese})}{\text{Total}} \right)} \right) = \left( \frac{\left( \frac{6}{6} \right)}{\left( \frac{3}{9} \right)} \right) = \left( \frac{1}{.3333333} \right) = 1.2857$$

	Support	Confidence	Lift
Apple, Milk	1/6	NA	NA
(Apple, Milk) => Cheese	1/6	1	(1/1)/(4/6)=1.5
Milk => Cheese	4/6	(4/4)=1	(4/4)/(4/6)=1.5
(Apple, Cheese) => Milk	1/6	(1/6)/1 = 1/6	(1/6)/(4/10)=0.417



7. (a) Consider the following hypothetical dataset concerning student characteristics whether or not each student should be hired. Use Naive Bayes Classifier to determine whether or not someone with poor GPA and lots of effort should be hired.

Name	GPA	Effort	Hirable?
Sarah	Poor	Lots	Yes
Dana	Average	Some	No
Alex	Average	Some	No
Annie	Average	Some	Yes
Emily	Excellent	Lots	Yes
Pete	Excellent	Lots	No
John	Excellent	Lots	No
Kathy	Poor	Some	No

**[Evaluation Scheme]** Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.

**[Solution]**

Sal (7)

(a) Naive Bayes Classifier.

$$P(\text{Hirable} = \text{yes}) = \frac{3}{8} = 0.375$$

$$P(\text{Hirable} = \text{No}) = \frac{5}{8} = 0.625$$

Conditional probability of each attributes

GPA	yes	No	Effort	yes	No
Poor	$\frac{1}{3}$	$\frac{1}{5}$	Lots	$\frac{2}{3}$	$\frac{2}{5}$
Avg	$\frac{1}{3}$	$\frac{2}{5}$	Some	0	$\frac{3}{5}$
Excellent	$\frac{1}{3}$	$\frac{2}{5}$			

New instance = GPA = (Poor, Effort = Lots)

$$V_{NB}(\text{yes}) = P(\text{yes}) \cdot P(\text{Poor}|\text{yes}) \cdot P(\text{Lots}|\text{yes})$$

$$V_{NB}(\text{yes}) = \cancel{P(\text{yes})} 0.375 \times 0.333 \times 1$$

$$V_{NB}(\text{yes}) = 0.1248$$

$$V_{NB}(\text{No}) = P(\text{No}) \cdot P(\text{Poor}|\text{No}) \cdot P(\text{Lots}|\text{No})$$

$$V_{NB}(\text{No}) = 0.625 \times 0.2 \times 0.4$$

$$V_{NB}(\text{No}) = 0.05$$

$$V_{NB}(\text{yes}) = \frac{V_{NB}(\text{yes})}{V_{NB}(\text{yes}) + V_{NB}(\text{No})} = \frac{0.1248}{0.1248 + 0.05}$$

$$= \frac{0.1248}{0.1748} = 0.7139$$

$$VNB(NO) = \frac{VNB(NO)}{VNB(YES) + VNB(NO)}$$

$$= \frac{0.05}{0.1748} = 0.2860$$

$\therefore$  The probability of getting hired with poor CGPA and lots of efforts are high.  
 so, There is a chance of getting hired with poor CGPA and lots of effort.

(b) Demonstrate a step-by-step process of Agglomerative hierarchical clustering with the following dataset. In addition, illustrate the merge with Dendrogram (keep the threshold as 5). Use Manhattan distance for the construction of matrix.

Roll	Mark
1	80
2	90
3	65
4	75
5	95
6	55

**[Evaluation Scheme]** Full mark for the correct answer. Step-wise mark

can be awarded based on the partial correctness of the solution.

**[Solution]**

7b step-1 calculate distance  
Matrix

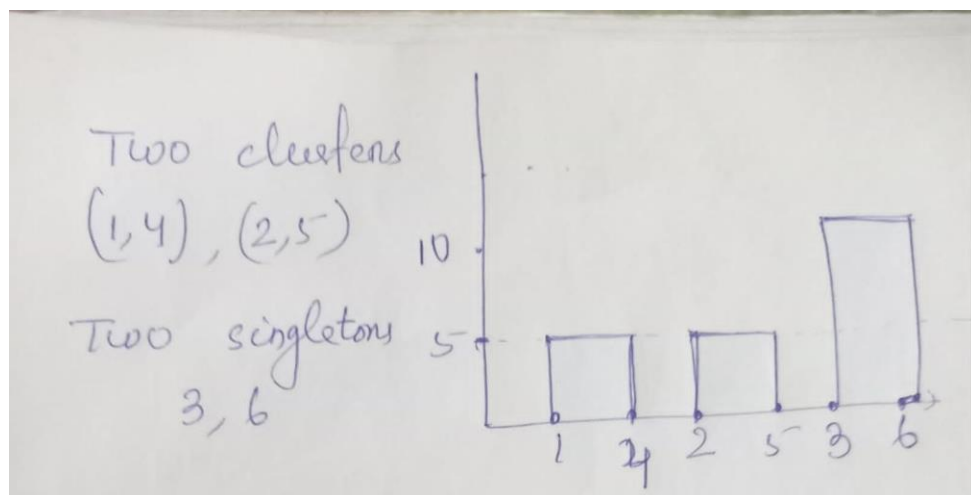
	1	2	3	4	5	6
1	0					
2	<u>10</u>	0				
3	15	25	0			
4	(5)	<u>15</u>	10	0		
5	15	5	30	20	0	
6	25	35	10	20	40	0

step-2 choose minimum distance to  
form cluster. (1,4)

	(1,4)	2	3	5	6
(1,4)	0				
2	10	0			
3	10	25	0		
5	15	(5)	30	0	
6	20	35	10	40	0

	(1,4)	3	(2,5)	6
(1,4)	0			
3	10	0		
(2,5)	10	10	0	
6	20	10	35	0



8.	<p>(a) Design an optimised algorithm for the updation of an element in a Bloom filter.</p> <p><b>[Evaluation Scheme]</b> Full mark for the correct answer. Step-wise mark can be awarded based on the partial correctness of the solution.</p> <p><b>[Solution]</b></p> <p>The steps of optimised algorithm is as follows</p> <ol style="list-style-type: none"> <li>1. Clear the bloom filter</li> <li>2. Insert all the elements into the bloom filter except the element to be updated.</li> <li>3. Insert the updated value into the bloom filter.</li> </ol> <p>The insert function code is as follows.</p> <pre> insert(e) begin /* Loop all hash functions k */ for j : 1 ... k do     m ← hj(e) //apply the hash function on e     Bm ← bf[m] //retrieve val at mth pos from Bloom filter bf     if Bm == 0 then         /* Bloom filter had zero bit at index m */         Bm ← 1;     end if end for end </pre> <p>The clear function code is as follows.</p> <pre> clear() begin     for i : 1 ... n // n is the size of the bloom filter         bf[i] = 0     end for end </pre>
	<p>(b) Consider a Bloom Filter of size 11, with integers as stream elements and two hash functions as follows:</p> <ul style="list-style-type: none"> <li>– <math>H1(x)</math> = take odd number of bits from right in the binary representation of X. Subsequently, treat it as an integer i, and result is i modulo 11.</li> <li>– <math>H2(x)</math> = same, but take even numbered bits.</li> </ul> <p>(1) Find the filter after the insertion of elements 25, 15 and 35.</p> <p>(2) Check whether the element <math>y=18</math> exists in the bloom filter or not. Is it</p>

the case of False Positive or False Negative? Explain.

**[Evaluation Scheme]** Full mark for the correct answer. Step-wise mark should be awarded based on the partial correctness of the solution.

**[Solution]**

Step 1: Initialization of bloom filter

0	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---

Step 2:

- Insertion of 25:

$$(25)_{10} = (11001)_2$$

Considering odd number of bits from right in  $(11001)_2 = 101$ , So

$$H1(101) = 101 \bmod 11 = 2$$

Considering even number of bits from right in  $(11001)_2 = 1$ , So

$$H2(1) = 1 \bmod 11 = 1$$

The revised bloom filter is as follows:

0	1	1	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---

- Insertion of 15

$$(15)_{10} = (1111)_2$$

Considering odd number of bits from right in  $(1111)_2 = 11$ , So

$$H1(11) = 11 \bmod 11 = 0$$

Considering even number of bits from right in  $(1111)_2 = 11$ , So

$$H2(11) = 11 \bmod 11 = 0$$

The revised bloom filter is as follows:

1	1	1	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---

- Insertion of 35

$$(35)_{10} = (100011)_2$$

Considering odd number of bits from right in  $(100011)_2 = 100$ , So

$$H1(100) = 100 \bmod 11 = 1$$

Considering even number of bits from right in  $(100011)_2 = 101$ , So

$$H2(101) = 101 \bmod 11 = 2$$

The revised bloom filter is as follows:

1	1	1	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---

Step 3:

Membership test of 18

$$(18)_{10} = (10010)_2$$

Considering odd number of bits from right in  $(10010)_2 = 1$ , So  $H1(1) = 1 \bmod 11 = 1$

		<p>Considering even number of bits from right in <math>(100011)_2 = 10</math>, So <math>H2(10) = 10 \bmod 11 = 10</math></p> <p>Since 10<sup>th</sup> slot of bloom filter is 0, it is concluded that 18 is definitely does not exist in bloom filter.</p>
--	--	--