CMR INSTITUTE OF TECHNOLOGY

USN | | | | | | | | | | |

Internal Assessment Test 1 – December 2022

Sub:	BIG DATA ANALYTICS							Sub Code:	20MCA352
Date:	28/12/2022	Duration:	90 min's	Max Marks:	50	Sem:	III	Branch:	MCA

Note : Answer FIVE FULL Questions, choosing ONE full question from each Module

		MARKS	OBE	
			CO	RBT
	PART I			
1	Write briefly about the following : a) Crowd-Sourcing b) Mobile Business Intelligence	[10]	CO2	L2
	OR			
2	List the various factors required for analytical model and explain.	[10]	CO1	L2
	PART II			
3	Examine and summarize the analytics application in marketing, risk management, government, web and logistics (2 each).	[10]	CO1	L2
	OR			
4	On the given dataset: 25, 32, 45, 68 54, 53 59, 63,76, 71, perform the following a) Min-Max Scaling. b) Decimal Scaling.	[10]	CO1	L4

CMR INSTITUTE OF TECHNOLOGY

USN | | | | | | | | | | |

Internal Assessment Test 1 – December 2022

Sub:	BIG DATA ANALYTICS							Sub Code:	20MCA352
Date:	28/12/2022	Duration:	90 min's	Max Marks:	50	Sem:	III	Branch:	MCA

Note : Answer FIVE FULL Questions, choosing ONE full question from each Module

		MARKS	OBE	
			CO	RBT
1	**PART I** Write briefly about the following : Crowd-Sourcing Mobile Business Intelligence	[10]	CO2	L2
	OR			
2	List the various factors required for analytical model and explain.	[10]	CO1	L2
	PART II			
3	Examine and summarize the analytics application in marketing, risk management, government, web and logistics(2 each).	[10]	CO1	L2
	OR			
4	On the given dataset: 25, 32, 45, 68 54, 53 59, 63,76, 71, perform the following a) Min-Max Scaling. b)Decimal Scaling.	[10]	CO1	L4

	PART III			
5	What is sampling? Explain in details the different types of sampling. OR	[10]	CO1	L2
6	Explain the following a) Data Discovery b) Hadoop Parallel world	[10]	CO2	L2
	PART IV			
7	Construct the box plot for 54, 60, 65, 66, 67, 69, 70, 72, 73, 75, 76 and identify lower and higher outliers OR Calculate the Z-Score and detect the outlier for the following data.	[10]	CO1	L3
8	Where mean = 40 Standard deviation = 10 and Data= 30 50 10 40 60 80	[10]	CO1	L3
	PART V			
9	Show that the Standard deviation of the z-scores of the following set of data is 1. 25, 15, 16, 23, 14, 29, 12, 9, 35, 40. OR	[10]	CO1	L4
10	Carl works at a computer store. He also recorded the number of sales he made each month. In the past 18 months, he sold the following numbers of computers: 51, 17, 25, 39, 7, 49, 72, 0, 95, 105, 56, 79, 67, 41, 20, 2, 43, 13. Construct the box plot for the above sales and identify outliers.	[10]	CO1	L4

	PART III			
5	What is sampling? Explain in details the different types of sampling. OR	[10]	CO1	L2
6	Explain the following a) Data Discovery b) Hadoop Parallel world	[10]	CO2	L2
	PART IV			
7	Construct the box plot for 54, 60, 65, 66, 67, 69, 70, 72, 73, 75, 76 and identify lower and higher outliers OR Calculate the Z-Score and detect the outlier for the following data.	[10]	CO1	L3
8	Where mean = 40 Standard deviation = 10 and Data= 30 50 10 40 60 80	[10]	CO1	L3
	PART V			
9	Show that the Standard deviation of the z-scores of the following set of data is 1. 25, 15, 16, 23, 14, 29, 12, 9, 35, 40. OR	[10]	CO1	L4
10	Carl works at a computer store. He also recorded the number of sales he made each month. In the past 18 months, he sold the following numbers of computers: 51, 17, 25, 39, 7, 49, 72, 0, 95, 105, 56, 79, 67, 41, 20, 2, 43, 13. Construct the box plot for the above sales and identify outliers.	[10]	CO1	L4

IAT 1 – 20MCA352 – BIG DATA ANALYTICS – SOLUTIONS

Q 1a. Crowd sourcing is a great way to capitalize on the resources that can build algorithms and predictive models. It is a disruptive business model whose roots are in technology but is extending beyond technology to other areas. There are various types of crowd sourcing, such as crowd voting, crowd creation, wisdom of crowds, crowd funding, and contests.

Crowd Voting : It's a type of crowdsourcing where customers are allowed to choose a winner. They can vote to decide which of the options is the best for them. This type can be applied to different situations. Consumers can choose one of the options provided by experts or products created by consumers. For instance, if a brand asks its consumers to create a new taste, package, or design of a product, other consumers vote to identify the best one.

Wisdom of the crowd : It's a collective opinion of different individuals gathered in a group. This type is used for decision-making since it allows one to find the best solution for problems. Many brands pay attention to the collective opinion of their customers because they help bring their businesses new ways of thinking, ideas, and strategies. As a result, the overall performance of a company improves.

Crowd creation : This type involves a company asking its customers to help with new products. This way, companies get brand new ideas and thoughts that help a business stand out. For instance, McDonald's is open to new ideas from its consumers. The famous fast food company asked customers to create their perfect burgers and submit their ideas to the brand. The company released winners' burgers each week, including the creator's short bio.

Crowdfunding: It's when people collect money and ask for investments for charities, projects, and startups without planning to return the money to the owners. People do it voluntarily. Often, companies gather money to help individuals and families suffering from natural disasters, poverty, social problems, etc.

Q 1b. **Mobile Business Intelligence :** Analytics on mobile devices is what some refer to as putting BI in your pocket. Mobile drives straight to the heart of simplicity and ease of use that has been a major barrier to BI adoption since day one.

Three elements that have impacted the viability of mobile BI:

1. Location—the GPS component and location . . . know where you are in time as well as the movement.

2. It's not just about pushing data; you can transact with your smart phone based on information you get.

3. Multimedia functionality allows the visualization pieces to really come into play.

Three challenges with mobile BI include:

1. Managing standards for rolling out these devices.

2. Managing security (always a big challenge).

3. Managing —bring your own device,‖ where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

Q2. A good analytical model should satisfy several requirements, depending on the application area.
• A first critical success factor is business relevance. The analytical model should actually solve the business problem for which it was developed.
• A second criterion is statistical performance. The model should have statistical significance and predictive power. Example: in a classification setting (churn or fraud) the model should have good discriminative power.
• Depending on the application, the model should also be interpretable and justifiable. Interoperability refers to understanding the pattern and justifiability refers to the degree to which model corresponds to prior business knowledge and institution
• Analytical models should also be operationally efficient. This refers to the efforts needed to collect the data, preprocess it, evaluate the model, and feed its outputs to the business application. Example: campaign management, capital calculation etc...
• Another key attention point is the economic cost needed to set up the analytical model. This includes the cost to gather and process the data, cost to analyze the data and cost to put the resulting analytical models into production.
• Finally, analytical models should also comply with both local and international regulation and legislation. Example: Credit risk setting.

Q3. Marketing : Response Modelling, Retention Modelling, Recommender System, Customer Segmentation, Market-based analysis

Risk Management : Credit Risk Modelling, Market Risk Modelling, Fraud Detection.

Government : Tax avoidance, social security fraud, money laundering, terrorism detection.

Web : Web analytics, Social media.

Logistics : Demand Forecasting, Supply chain analysis.

Q4.

DATA(X_{old})	X_{new}	
	Min-Max Scaling((X_{old}-X_{min})/X_{max}-X_{min})	Decimal Scaling(X_{old}/*10 raised to the power of max digits, here, 2*)
25(X_{min})	0.00	0.25
32	0.13	0.32
45	0.39	0.45
68	0.84	0.68
54	0.56	0.54
53	0.54	0.53
59	0.66	0.59
63	0.74	0.63
76(X_{max})	1.00	0.76
71	0.90	0.71

Q5. Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed which could either be probabilistic or non-probabilistic.

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Probability Sampling Methods:

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: Simple random sampling. You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: Systematic sampling. All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.
If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: Stratified sampling. The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which give you a representative sample of 100 people.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: Cluster sampling. The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Non Probability Sampling :

1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example: Convenience sampling. You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others, leading to self-selection bias.

Example: Voluntary response sampling. You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

3. Purposive sampling

This type of sampling, also known as judgment sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make

sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments.

Example: Purposive sampling. You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

4. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

Example: Snowball sampling. You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

Q6a. **Data discovery** involves the collection and evaluation of data from various sources and is often used to understand trends and patterns in the data. It requires a progression of steps that organizations can use as a framework to understand their data. Data discovery, usually associated with business intelligence (BI), helps inform business decisions by bringing together disparate, siloed data sources to be analyzed. Having mounds of data is useless unless you find a way to extract insights from it. The data discovery process includes connecting multiple data sources, cleansing and preparing the data, sharing the data throughout the organization, and performing analysis to gain insights into business processes.

Data discovery offers businesses a way to make their data clean, easily understandable, and user-friendly. A comprehensive solution should be able to be used by all members of the business. The main benefit of data discovery is the actionable insights that are uncovered in the data. These insights help users spot valuable opportunities before competitors without having to consult the IT organization. Visual data discovery can enhance this value, allowing line of business workers to find answers faster.

Q6b. Hadoop Parallel World :

- There are many Big Data technologies that have been making an impact on the new technology stacks for handling Big Data, but Apache Hadoop is one technology that has been the darling of Big Data talk. Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive the complete value from all their data.

- Hadoop gives organizations the flexibility to ask questions across their structured and unstructured data that were previously impossible to ask or solve: The scale and variety of data have permanently overwhelmed the ability to cost-effectively value using traditional platforms.

- The scalability and elasticity of free, open-source Hadoop running on standard hardware allow organizations to hold onto more data than ever before. Hadoop excels at supporting complex analyses across large collection of data.

- Hadoop handle variety of workloads, including search, log processing, recommendation system, data warehousing and video/image analysis. Today's explosion of data types and volumes means that Big Data equals big opportunities and Apache empowers organizations to work on the most modern scale out architectures.

- Apache Hadoop is an open source project administered by the Apache Software Foundation. The software was originally developed by the world's largest internet companies to capture and analyze the data that they generate.

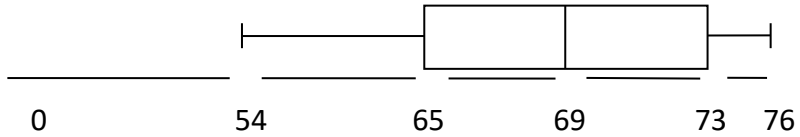
Hadoop runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system

The two critical components of Hadoop are:

- The Hadoop Distributed File System (HDFS)
- MapReduce

Q 7. **54, 60, 65, 66, 67, 69, 70, 72, 73, 75, 76**

There are 11(odd) data points so the median is (n+1)/2 th observation which is **69.**
Min = 54 and Max = 76
Q1 = 65 and Q3 = 73. IQR = Q3 – Q1 = 8.
Range = [Q1 – 1.5*IQR, Q3 + 1.5*IQR] = [65-12, 73+12] = [53, 85]



Q 8.

X	$X - X_{mean}$	Z score $((X - X_{mean})/SD)$
10	-30	-3
30	-10	-1
40	0	0
50	10	1
60	20	2
80	40	4

Z-score which is greater than +3 or are less than -3 is considered as an outlier, thus the point **80** is an outlier.

Q9. **25, 15, 16, 23, 14, 29, 12, 9, 35, 40.**

X_{Mean} = 21.8, Standard Deviation(SD) = 9.84

X	**X - X_{Mean}**	Z score ((X - X_{mean})/SD)	
25	**3.2**	**0.32**	
15	**-6.8**	**-0.69**	
16	**-5.8**	**-0.58**	
23	**1.2**	**0.12**	
14	**-7.8**	**-0.79**	
29	**7.2**	**0.73**	
12	**-9.8**	**-0.99**	
9	**-12.8**	**-1.30**	
35	**13.2**	**1.34**	
40	**18.2**	**1.84**	

$Zscore_{mean}$ = 0, $Zscore_{SD}$ = 0.997(almost 1)

Q10. **51, 17, 25, 39, 7, 49, 72, 0, 95, 105, 56, 79, 67, 41, 20, 2, 43, 13.**

Median(Q2) = 42
Min = 0 and Max = 105
Q1 = 17 and Q3 = 67. IQR = Q3 – Q1 = 50.
Range = [Q1 – 1.5*IQR, Q3 + 1.5*IQR] = [17-75, 67+75] = [-58, 142]
Outliers : None