DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

WEDNESDAY: 3 December 2025. Afternoon Paper. **Time Allowed: 3 hours.**

This paper has two (2) sections. SECTION I has twenty (20) short response questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Answer ALL questions. Marks allocated to each question are indicated in the question.

Required Resources:
- **R Studio**
- **Ms Excel 2016**
- **Hadoop software**

SECTION I (40 MARKS)

1. When a digital photograph contains details like the date taken, camera model and resolution, this information is known as _____. (2 marks)

2. The sequence of stages through which big data passes from creation to storage, processing and disposal is called the _____. (2 marks)

3. The type of analytics that recommends specific actions or strategies based on predictive models is referred to as _____. (2 marks)

4. When a machine-learning model performs well on training data but poorly on new data, it is said to suffer from _____. (2 marks)

5. Data that contains tags such as XML or JSON and lies between structured and unstructured formats is known as _____. (2 marks)

6. The practice of combining visuals and narratives to communicate insights clearly and persuasively is known as _____. (2 marks)

7. The Excel feature that automatically completes data entries based on detected patterns in adjacent cells is called _____. (2 marks)

8. The R data structure that can hold a combination of heterogeneous elements, such as numeric, character or logical values is known as _____. (2 marks)

9. _____ is a data structure in R that stores tabular data in rows and columns, where each column can be of a different type. (2 marks)

10. The Excel add-in used for performing regression analysis, correlation and descriptive statistics is known as the _____. (2 marks)

11. The Big Data characteristic that measures the usefulness or business relevance of information extracted from large datasets is known as _____. (2 marks)

12. A dataset that focuses on a single variable for analysis, such as the ages of students in a school is called a _____. (2 marks)

13. The probability distribution that describes the likelihood of a fixed number of successes in a fixed number of independent trials is known as _____. (2 marks)

14. _____ is a Big Data processing technique that achieves real-time analytics by keeping active datasets close to the processor rather than relying on disk storage. (2 marks)

15. The R environment window where users type commands and view immediate results is called the _____. (2 marks)

16. The cartographic technique that uses repeated symbols to represent the spatial distribution of a variable is known as _____. (2 marks)

17. A visualisation that plots values on two or three dimensions to show correlations between variables is called a _____. (2 marks)

18. Data that describe characteristics or attributes and are organised into distinct non-numeric groups or classes is known as _____. (2 marks)

19. _____ is the Hadoop ecosystem tool designed for querying and analysing large datasets using a SQL-like language. (2 marks)

20. The technique of executing multiple computational tasks at the same time to improve performance is called _____. (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (f).

Use the data in the patient table below to answer the questions that follow:

PatientNumber	PatientFname	WardNumber	Admissionfee_shs
D564	JOHN	149	40000
R768	ALICE	162	56000
R342	JOAN	175	34000
W435	VICTOR	149	45000
P786	TOM	201	34000
Y546	PETER	162	45000
R908	REEGAN	149	34000
D564	JOHN	162	40000
R768	ALICE	253	56000
R342	JOAN	149	34000
W435	VICTOR	162	45000
P786	TOM	292	34000
Y546	PETER	162	45000
R908	REEGAN	175	34000
D564	JOHN	175	40000

(a) Create a csv file using the data and save it as "mypatient.csv". (3 marks)

(b) Write the R Studio code to import the csv file into R and display the information above. (3 marks)

(c) Write the R Studio code to find the total admission fees paid for admission into different wards. (3 marks)

(d) Write the R Studio code to create a pie chart for average admission fees with reference to patient number. (3 marks)

(e) Using the patient table, create a Pivot table in Excel to analyse total Admission fees per Ward number. (4 marks)

(f) Using the patient data provided, create a bar chart in Excel that shows the average Admission fees for each Ward number. (4 marks)

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 21" document and upload. **(Total: 20 marks)**

22. Create a word processing document named "Question 22" and use the word processor document to save your answers to questions (a) to (e).

Use the data in the table below to answer the questions that follow:

CARNUMBER	CARMODEL	YEAROFMANUFACTURE	CARCOSTDOLLARS
KDP435N	TOYOTA	2014	3256
KDN546R	MAZDA	2010	4566
KDD564T	MAZDA	2011	4987
KCX453D	TOYOTA	2014	8765
KCT498K	NISSAN	2010	4567
KDC998R	TOYOTA	2015	7658
KDB321K	NISSAN	2015	4567
KDH765Y	MAZDA	2012	6989
KDC347K	BMW	2011	8765

(a) Write the R Studio code to find the mean, median and standard deviation of car cost in dollars. (4 marks)

(b) Write the R Studio code to find the average car cost in dollars for each car model. (3 marks)

(c) (i) Write the R Studio code to calculate the correlation between year of manufacture and car cost in dollars.

(ii) Write the code to plot a simple scatterplot to visualise if newer cars tend to cost more. Interpret the correlation. (8 marks)

(d) Write the R Studio code to summarise the data. (2 marks)

(e) Write the r studio code to find the number of cars that have a car cost in dollars value above the overall mean cost and list their car number and car model. (3 marks)

Capture screenshots to demonstrate how you have performed the above tasks.

Save "Question 22" document and upload. **(Total: 20 marks)**

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) to (c).

(a) Write the Hadoop hive statement to create and insert data into the "Loan" table. (8 marks)

LoanID	Age	Income	LoanAmount	LoanStatus
121	23	70000	2500	Yes
122	50	80000	5200	Yes
123	40	62000	5000	No
124	30	82000	1500	Yes
125	29	48000	2000	No
126	35	75000	4000	Yes
127	25	50000	3000	No
128	45	60000	2700	Yes
129	50	55000	4500	No

(b) Using the loans table, write the R Studio code to plot a line graph of income against loan amount. (6 marks)

(c) Write a Spark transformation to find the average LoanAmount for applicants whose LoanStatus is "Yes". (6 marks)

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 23" document and upload. **(Total: 20 marks)**

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

MONDAY: 18 August 2025. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two (2) sections. SECTION I has twenty (20) short response questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:
- **Hadoop software**
- **RStudio 2025.05.1 installed with the following libraries:**

(i) **rJava**
(ii) **rhdfs**
(iii) **rhbase**
(iv) **rmr2**

- **Ms Excel 2016**

SECTION I (40 MARKS)

1. _____ is the type of data visualisation where words or items are displayed in varying sizes based on their frequency or importance, often used to show keyword prominence. (2 marks)

2. JW Insurance Company constantly evaluates opportunities to improve efficiencies across all its operations, including commercial operations and charitable activities. What data mining technique would you use to categorise its customers? (2 marks)

3. The data format in which files store information in a human-readable format, with records separated by newline characters and additional characters like commas (for example: CS and TSV) is called _____ text files. (2 marks)

4. The Microsoft Excel chart feature which provides a description of what each colour or symbol in the chart represents, helping users interpret the data more easily is called _____. (2 marks)

5. The technological trend term used to refer to every observation is now being observed and stored to transform it into quantifiable data, making organisations more data-driven is _____. (2 marks)

6. _____ are visual representations of data that help stakeholders quickly understand patterns, trends and outliers in business intelligence dashboards. (2 marks)

7. Which essential element of big data is concerned with the trustworthiness, credibility and accuracy of the data used in data analytics to avoid flawed insight due to inaccurate data? (2 marks)

8. State the type of class in which the dataset below is categorised: (2 marks)

Company 2009	**Profits ($ millions)**
MCDonalds	10,555
Yum Brands	8,065
Starbucks	3,150
Darden Restaurants	445.70

9. In data analytics, _____ refers to the speed at which data is generated, collected, processed and analysed. (2 marks)

10. If a customer's monthly data usage is low (less than 2 GB), there is a very high probability (70%) that they will leave the company within the next three months. If their data usage is healthy, the next factor to consider is the customer's tenure with the company. If their tenure is low (less than 12 months), their likelihood of staying with the company is almost guaranteed (98%). If their tenure is higher, we then examine their average call duration. If their average call duration is healthy, their chances of staying are 89%, otherwise, the chance of retention drops to 50%.

    Which data mining techniques can a mobile telecom company apply in data analytics to get insights about its customers using the information given in this scenario? (2 marks)

11. The processing engine in Hadoop that enables parallel computation of large data sets by dividing tasks into independent subtasks and combining their outputs is called _____. (2 marks)

12. The _____ is an indication of how different the numbers in the dataset are from one another. (2 marks)

13. The versatile data structures in R programming that support storing mixed types and complex data arrangements, playing a crucial role in data frames and object-oriented programming are the _____. (2 marks)

14. Data frames can combine information across datasets based on common variables using the _____ function. (2 marks)

15. _____ is used in R programming to represent missing data within a dataset which can affect statistical calculations unless explicitly handled, while NULL, on the other hand signifies that a value does not exist at all. (2 marks)

16. State the R command that would be used to set the current working directory to a directory named “cisse” on your local drive C _____. (2 marks)

17. After importing data, to make sure all the values were read in correctly, you need to inspect the first few lines of a dataset. In R programming, we use the _____command. (2 marks)

18. The component placed on top of Hadoop Distributed File System to provide capabilities like those of an operating system for Big Data analytics applications is called _____. (2 marks)

19. The programming paradigm that processes continuous data arising from telecommunications, financial and transportation domains is known as _____. (2 marks)

20. _____ is a machine learning technique commonly used in big data analytics where algorithms learn patterns from labelled data to make predictions or classifications. (2 marks)

SECTION II (60 MARKS)

21. Create a Word document named “Question 21” to capture and save the screenshots for answers to questions (a) to (h) below.

    Create a script called “Electronics.R” using an IDE of your choice.

Use the dataset provided below to answer the questions that follow:

salesman_id	item	price($)	sales_region	transaction_date
101	Laptop	1200	East	2025-06-03
102	Mouse	25	West	2025-06-03
101	Keyboard	75	East	2025-06-04
103	Monitor	300	North	2025-06-04
102	Webcam	50	West	2025-06-05
104	Laptop	1150	South	2025-06-05
102	Printer	200	North	2025-06-06
101	Mouse	28	East	2025-06-06
105	Headphones	80	West	2025-06-07
102	Keyboard	70	South	2025-06-07

(a) Create a folder named "cisse" on your C: drive. Use the table provided above to create a CSV dataset called “electronics” to store the data in a folder created. (4 marks)

(b) Install package “dplyr” and load the library for data manipulation. (2 marks)

(c) Load the CSV dataset created into an object named “electronics”. (2 marks)

(d) Convert the “transaction_date” into a Date format. (2 marks)

(e) Write the R code to display the summary statistics about the data stored in the data frame. (2 marks)

(f) Display the first **FOUR** rows of the dataset. (2 marks)

(g) Total sales by all employees. (3 marks)

(h) Display the individual Sales Records for Salesman_ID 102. (3 marks)

Save “Question 21” document and upload. **(Total: 20 marks)**

22. Create a word processing document named “Question 22” and use the word processor document to save your answers to questions (a) to (c).

(a) Write the Hadoop spark statements to create and insert data into the **vehicle** table below: (8 marks)

Vehicle type	Vehicle price ($)	Vehicle weight (Kgs)
Toyota car	654	760
Bedford lorry	867	1234
Scania bus	899	1354
Isuzu minibus	785	987
Nissan Matatu	877	1000
Isuzu lorry	909	1231
Nissan car	546	765
Toyota car	543	761
Nissan car	523	766
Scania bus	891	1350
Tata lorry	807	1323
Mazda car	511	765
Nissan minibus	827	988
Isuzu lorry	900	1300

(b) Write the Hadoop hive statements that will find the vehicle types where the number of records is more than 1. (4 marks)

(c) Use the data in the performance table below to answer the questions that follow:

Qualification type	Pass	Credit	Distinction	Fail	Number of students
CPA	8	7	5	2	22
DDMA	11	8	7	5	31
ATD	14	9	9	6	38
CAMS	34	23	8	9	74
DCSNA	23	15	10	1	49
CFFE	7	5	6	3	21
CIFA	16	12	9	6	43
CISSE	21	4	5	8	38

(i) Using R studio, create a matrix called qualification data with rows representing the qualification types and columns representing the grades. (4 marks)

(ii) Using R studio, write the code to print the qualification list where each element is named after a qualification (for example, "CPA" and "DDMA") and contains a vector of the 4 grade values; Pass, Credit, Distinction and Fail. (4 marks)

Capture screenshots to demonstrate how you have performed the above task.

Save "Question 22" document and upload. **(Total: 20 marks)**

23. Create a Word document named "Question 23" to capture and save the screenshots for answers to questions (a) to (h) below.

Create a script called "Stocks.R" using an IDE of your choice.

Use the following stock sales dataset for the "Nairobi Stock Exchange" for the four quarters over the last five years:

Year_Quarter	Sales_Sh._Millions
2020-Q1	150.5
2020-Q2	152.1
2020-Q3	155.8
2020-Q4	158.2
2021-Q1	160.7
2021-Q2	163.0
2021-Q3	165.5
2021-Q4	168.1
2022-Q1	170.8
2022-Q2	173.3
2022-Q3	176.0
2022-Q4	178.5
2023-Q1	181.2
2023-Q2	184.0
2023-Q3	186.7
2023-Q4	189.5
2024-Q1	192.3
2024-Q2	195.0
2024-Q3	197.8
2024-Q4	200.5

(a) Create a CSV dataset named "stocks" using the information provided in the above table and store it in the directory called "cisse", created in "Question 21". (3 marks)

(b) Load the libraries "dplyr" and "ggplot2" for data manipulation and visualisation, respectively. (1 mark)

(c) Define the path for your dataset and store it in an object named "file_path". (1 mark)

(d) Load the dataset and load it in an object named "stocks_df"and display the first records. (2 marks)

(e) Create a sequential time index by adding a new variable name "Time_Index" and display the new dataset after adding the time index. (2 marks)

(f) Construct a linear model to predict "Sales_Ksh_Millions" based on "Time_Index" and display the coefficients of the model summary. (3 marks)

(g) Use the trained model to predict sales values for the existing time points by showing the Actual versus Predicted Sales for the first six rows. (4 marks)

(h) Create a well-labelled visualisation using a line plot for the last five years. (4 marks)

Save "Question 23" document and upload. **(Total: 20 marks)**

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

TUESDAY: 22 April 2025. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two (2) sections. SECTION I has twenty (20) short response questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:
- **R Studio**
- **Ms Excel 2016**
- **Hadoop software**
- **Spreadsheet software**

SECTION I (40 MARKS)

1. Data such as social media posts, music, video and text are classified into a type of data known as _____ data. (2 marks)

2. In R programming, it is possible to add new functions using packages to expand the capabilities of R. Write the script that you would use to display the list of all the installed packages. (2 marks)

3. Write down the formula that is used to calculate the predictive accuracy of a model as used in predictive analytics. (2 marks)

4. Which characteristic of data is defined as "Data should be recorded and used in compliance with relevant requirements, including the correct application of any rules or definitions"? (2 marks)

5. In data analytics, data can be classified as either quantitative or qualitative. State an example of nominal qualitative data as used in data analytics. (2 marks)

6. The process of extraction of particular subsets of a series, matrix or vector for display on the console screen in R programming is called _____. (2 marks)

7. The server that stores information about file names and their attributes in an HDFS is called a _____ (2 marks)

8. The square root of the arithmetic mean of the squared values for a set of observation is known as_____ (2 marks)

9. State the R programming command that is used to show the information about the current working directory. (2 marks)

10. The type of machine learning where the model is trained using both labeled and unlabeled data is called _____. (2 marks)

11. The process of redistribution of data so that the pairs produced by map having the same keys are on the same machines in the MapReduce job is known as _____. (2 marks)

12. The technique used in Microsoft Excel to show a subset of the rows in the spreadsheet that pass a given condition as used in data analytics is called _____. (2 marks)

13. The type of correlation which is characterised by if the dependent variables go up when the independent variable also goes up is known as _____ (2 marks)

14. An observa tion that deviates so much from other observations in clustering arousing suspicion that it was generated by a different mechanism is known as an _____. (2 marks)

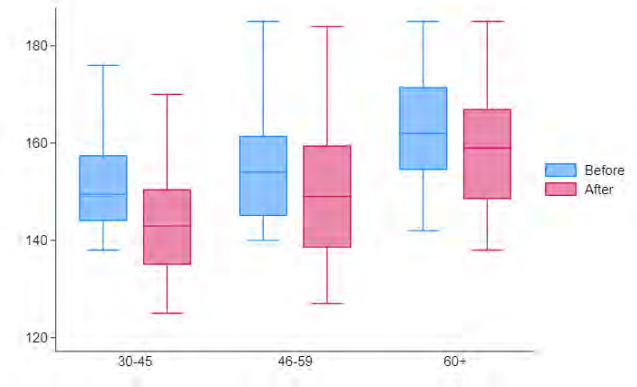
15. The data transformation technique done to scale the data values in a specified range such as -1.0 to 1.0 or 0.0 to 1.0 is known as _____. (2 marks)

16. The big data characteristic which is described as "data in doubt" due to data inconsistency, incompleteness and ambiguities is referred to as _____. (2 marks)

17. Illustrate how a variable named "varNums" can be assigned with a vector of four integer values using the leftward operator as used in R programming. (2 marks)

18. The type of clustering that organises a given collection of text documents according to their content similarities into clusters of related topics is known as _____. (2 marks)

19. Identify the type of data visualisation chart that is shown in the following figure. (2 marks)



20. In data visualisation a tiny chart displayed for a quick data overview in an Ms Excel worksheet cell, as shown in the figure below, is known as a _____. (2 marks)

	A	B	C	D	E	F	G
1	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**June**	
2	45	66	57	84	64	51	
3	45	66	57	84	64	51	
4	1	1	-1	1	-1	-1	

SECTION II (60 MARKS)

21. Create a Word document called "Question 21" to capture and save the screenshots for answers to questions (a) to (f) below. Create a script called "Analytics.R" using an IDE of your choice. Use the dataset provided below to answer the questions that follow.

```
emp_id emp_name salary start_date
   101     John  65000 2022-01-01
   102     Anna  80000 2023-09-23
   103     Mary  54000 2024-11-22
   104     Ryan  75000 2022-05-11
   105    Peter  95000 2019-03-27
   106     Ruth  56000 2020-03-02
```

(a) Using R programming language, create a data frame called "emp.data" to store and display the data in the table provided. (4 marks)

(b) Write R code to display the summary statistics about the data stored in the data frame. (2 marks)

(c) Write R code to display the data frame structure "emp.data". (2 marks)

(d) Add a new column called "gender" to include the gender for each person as either "male" or "female" and display the results. (4 marks)

(e) Write the code to display the employees whose salary is greater than or equal to 75000. (4 marks)

(f) Write the code to display all the records sorted in descending order by salary. (4 marks)

Save and upload "Question 21" document.

(Total: 20 marks)

22. Create a Word document called "Question 22" to capture and save the screenshots for answers to questions (a) to (e).

(a) Explain how you could perform the following configurations after Login into the Hadoop file, the Hadoop HDFS.

(i) Write an appropriate command that could allow you to switch to the Hadoop user for proper permissions. (2 marks)

(ii) Write a command to create a new directory for storing HDFS data called "data" in a current working directory. (2 marks)

(iii) Write a command that would display the text "Hello, Hadoop World!" and redirect it into a file called "sample.txt" inside the temporary file directory. (2 marks)

(iv) Write a command that could be used to upload the file created in (a) (iii) to the directory created in (a) (ii). 2 marks)

(v) Explain how you could display the contents of a directory called "data" to confirm if the file created in (a) (iii) exists. (2 marks)

(b) Use the data provided below to create Ms Excel workbook named "sales". (2 marks)

	A	B	C	D
1				
2		Product Category	Product Name	Sales Amount
3		Electronics	Laptop	1200
4		Electronics	Smartphone	800
5		Electronics	Tablet	600
6		Clothing	T-Shirt	300
7		Clothing	Jeans	450
8		Clothing	Dress	600
9		Home Appliances	Refrigerator	900
10		Home Appliances	Microwave	350
11		Home Appliances	Blender	200
12		Sports	Tennis Racket	150
13		Sports	Basketball	180
14		Sports	Yoga Mat	120
15		Beauty	Perfume	250
16		Beauty	Makeup Kit	320
17		Beauty	Skincare	200

(c) Create a pivot table to summarise the total sales for all products grouped by product category. (4 marks)

(d) Calculate the number of sales for the clothing product category and display the formula on the worksheet. (2 marks)

(e) Calculate the total sales for all clothing products sold and display the formula on the worksheet. (2 marks)

Save and upload "Question 22".

(Total: 20 marks)

23. Create a Word document called "Question 23" to capture and save the screenshots for answers to questions (a) to (f) below. Create a script called "patients.R" using an IDE of your choice.

Use the patients' health data provided in the table below to answer the questions (a) to (f):

```
     [,1] [,2] [,3]
[1,]   80  120 98.6
[2,]   75  110 99.1
[3,]   90  130 98.4
[4,]   70  115 98.7
[5,]   85  125 99.0
```

(a) Using R programming, create a script called "patients" and develop a matrix of 5 rows and 3 columns of the patients' data using the data provided. (4 marks)

(b) Write a script in R to display the matrix created in (a). (2 marks)

(c) Write a script in R to assign the names of the columns to represent the patients' attributes and the rows to represent the patients' numbers to produce the output as shown below. (4 marks)

	HeartRate	SystolicBP	Temperature
001	80	120	98.6
002	75	110	99.1
003	90	130	98.4
004	70	115	98.7
005	85	125	99.0

(d) Write a script to display the mean of the heart rates for all the patients. (4 marks)

(e) Write a script to filter and display all the patients whose heart rate is greater than 75. (4 marks)

(f) Write a script in R, to display the scatter plot for the "patient_data" dataset. (2 marks)

Save and upload "Question 23" document.

(Total: 20 marks)

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

MONDAY: 2 December 2024. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two (2) sections. SECTION I has twenty (20) short response questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:

- **R Studio**
- **Ms Excel 2016**
- **Hadoop software**
- **R**

SECTION I (40 MARKS)

1. The depiction of data in visual form so that quality and relationships may be observed by a human analyst in data analytics is called_____. (2 marks)

2. Data structure in R refers to how the data is stored and organised to enable efficient data processing. Which non-primitive data structure is referred to as a vector with single data? (2 marks)

3. Write the R command that will assign the value 60 to the variable "age". (2 marks)

4. The correlation dataset with two variables move in the same direction is known as_____. (2 marks)

5. In R programming data analytics, data structures are essential for storing, manipulating, and analysing data. State the data structure that is best suited for elements of different types such as numeric, character, logical and other data structures. (2 marks)

6. Data is usually dirty and needs to be cleaned to ensure the quality of the analytical tasks. Write the R programming script that would be used to check the total missing values from the attribute name from a DataFrame named "dfmodel". (2 marks)

7. The data analysis approach that identifies general patterns in the data such as outliers and features of the data that might be unexpected is referred to as _____. (2 marks)

8. What tool is used in the Hadoop Framework for importing/exporting data between Hadoop and relational databases? (2 marks)

9. You would like to explore and visualise the possibility of an association between height and the weight of a group of adults. Which type of graph would be the best fit for the task? (2 marks)

10. The type of probability which gives the probability of an event happening k number of times within a given interval of time or space is known as _____. (2 marks)

11. The process of combining more than one predictive model to obtain more accurate results as used in data management and analytics is called _____. (2 marks)

12. A collection of members of a population that are similar enough that they can be regarded as "going together" in unsupervised machine learning is called _____. (2 marks)

13. You have downloaded the Hadoop file "Hadoop-3.1.0.tar.gz" on your local machine in a Linux environment. What command will you use to extract the downloaded file? (2 marks)

14. Data can be analysed for actionable insights, but with so much data of all types being analysed from across data sources, it is very difficult to ensure correctness and proof of accuracy. What name is given to this characteristic as used in big data? (2 marks)

15. A tool in data analytics that makes it easier for data analysts, decision makers and average users to understand their data, gain deeper insights and make better data-driven decisions is known as _____. (2 marks)

16. Name the R packages that allow for creating interactive dashboards where you can filter and zoom in on business Key Performance Indicators (KPIs). (2 marks)

17. In Data Analytics, understanding the properties of a dataset is critical for choosing the right algorithms and statistical methods. What name is given to when a dataset is imbalanced when one class significantly outnumbers another, common in classification problems? (2 marks)

18. What type of statistics is used to check data skewness and kurtosis and for techniques like Kernel Density Estimation (KDE) to model data distributions? (2 marks)

19. The third stage of the big data analytics life cycle is called _____. (2 marks)

20. Which machine learning algorithm fits a straight line to the data by minimising the distance between the actual and predicted values? (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use it to capture and save the screenshots for answers to questions (a) to (d) to demonstrate how you performed the tasks below.

(a) Give the Hadoop spark statement to create and insert data into the Product table below. (8 marks)

PRODUCT ID	PRODUCT DESCRIPTION	INVOICE NUMBER	QUANTITY	UNIT COST (DOLLARS)
R0098W	Laptop	B005678	453	2345
E5674W	Tablet	B0076986	657	1657
D54673W	Desktop	V09876Y5	234	860
R0098W	Laptop	R9878676	123	2345
R0098W	Laptop	K00768T6	256	2345
BS567R	Smartphone	D0975643	723	2001
E5674W	Tablet	V0980078	345	1657
D54673W	Desktop	D5645367	234	860
D54673W	Desktop	ER456545	267	860
BS567R	Smartphone	W8795R2	391	2001

(b) Write the Hadoop hive statement that will find the total cost of products by their description. Display your results table. (4 marks)

(c) Using Microsoft Excel pivot table function, summarise the average quantity of items by product description. (4 marks)

(d) Write the Microsoft Excel statement that will count the number of times the unit cost of 860 dollars appears in the worksheet. (4 marks)

Save and upload "Question 21".

(Total: 20 marks)

22. Create a word processing document called "Question 22" to capture and save the screenshots for answers to questions (a) to (f) below. Create a script called "employees.R" using an IDE of your choice.

Use the following employee data provided in the table below to answer the questions (a) to (f):

EmployeeID	Name	Age	Department	Salary
1	Trevor	30	HR	50000
2	Ombati	25	Finance	60000
3	Karli	28	IT	55000
4	Beryl	32	IT	62000
5	Sean	45	HR	70000

(a) Create the employee dataset as a CSV file called "employee" saving it in the appropriate location on your machine. (3 marks)

(b) Write a script in R to load the dataset in an object called "dfEmps", store and display the salary column as a vector

(c) Arrange the employees' records from the highest to the least paid employee using the R script. (4 marks)

(d) Write a script in R to calculate and display the average salary by department. (3 marks)

(e) Write a script that will convert the DataFrame created in (b) as a matrix called "empMatrix" and display the matrix created. (3 marks)

(f) Using the employee dataset provided, use the first two records to create and display a list data structure in R called "empList".

Save and upload "Question 22".

(Total: 20 marks)

23. Create a word processing document called "Question 23" to capture and save the screenshots for answers to questions (a) to (f) below. Create a script called "prediction.R" using an IDE of your choice.

You have been contracted as a data scientist in a human resources company to build a machine learning model to predict the attrition rate (the rate at which employees leave the companies) of employees in companies. The following staff data is provided in the table below.

Name	Age	SalaryPaid	Department	Attrition
John	25	45000	HR	0
Mary	30	50000	Finance	1
Chris	28	60000	IT	0
Lisa	35	70000	Marketing	1
Paul	40	55000	HR	0
Nina	45	72000	Finance	1
Mike	50	80000	IT	1
Sophia	38	65000	Marketing	0

Use the data provided to answer the questions (a) to (f):

(a) Use the information provided to create a CSV dataset named "staffInfo". (4 marks)

(b) Write an R script to load the CSV file into an R DataFrame named "staffData" and display the results loading the appropriate library for the task. (4 marks)

(c) Use R code to covert the department and attrition attributes as a factor. (2 marks)

(d) Write a code to display the structure of your dataset (1 mark)

(e) Using R script, build the logistic regression model, using attrition as the dependent variable and independent variable Age, SalaryPaid, Department and display the summary of the model. (5 marks)

(f) Write a script to predict probabilities of attrition, display predicted probabilities, set a threshold to classify as attrition (1) or no attrition (0), and display the predicted class. (4 marks)

Save and upload "Question 23".

(Total: 20 marks)

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

MONDAY: 19 August 2024. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two (2) sections. SECTION I has twenty (20) short response questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are indicated in the question.

Required Resources:
- **R Studio**
- **Spreadsheet program**
- **Hadoop software**

SECTION I (40 MARKS)

1. A workflow scheduler system to manage Apache Hadoop jobs that is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs as well as system specific jobs is known as? (2 marks)

2. The machine learning algorithms are better suited for more complex processing tasks, such as organising large datasets into clusters and are useful for identifying previously undetected patterns in data and can help identify features useful for categorising data is known as? (2 marks)

3. The statistical method that is used to test the difference between two or more mean formulas in a spreadsheet application is known as? (2 marks)

4. A table that allows a data scientist to analyse the relationship between each pair of numeric variables of a dataset is known as? (2 marks)

5. The use of statistics and modeling techniques to forecast future outcomes is referred to as? (2 marks)

6. The module on top of Spark Core that provides machine learning primitives as Application Programming Interfaces (APIs) is known as? (2 marks)

7. The technique in the big data analytics life cycle that involves examining, analysing and creating useful summaries of data which aids in the discovery of data quality issues, risks and overall trends is known as? (2 marks)

8. The process of collecting unstructured data from online platforms to derive valuable customer and behavioral insights that support critical business decision-making is known as? (2 marks)

9. A thematic map in which a mapping variable such as travel time, population, is substituted for land area or distance is called? (2 marks)

10. When data points on a bell curve are not distributed symmetrically to the left and right sides of the median is known as? (2 marks)

11. The information that doesn't reside in a traditional row-column database and it's usually text-heavy but may include data such as dates, numbers and facts is called? (2 marks)

12. Which function in R-programming is used to know all the variables that are available in the workspace? (2 marks)

13. The R objects in which the elements of the same atomic type are arranged in a two-dimensional rectangular layout are known as? (2 marks)

14. The data visualisation tool that empowers users to create captivating visuals with ease, with an intuitive drag-and-drop interface and makes it easy to design captivating and interactive charts, maps and dashboards is referred to as? (2 marks)

15. When deploying a model into a big data platform, the process that involves collecting data from different sources like social media platforms and business applications is referred to as? (2 marks)

16. A two-dimensional representation of data in which values are represented by colours is known as? (2 marks)

17. The standard deviation, variance, minimum variable, maximum variable, kurtosis and skewness are collectively referred to as? (2 marks)

18. A component on Hadoop that offers centralised service for maintaining configuration information, naming, providing distributed synchronisation and providing group services is called? (2 marks)

19. The big data technology that focuses on the real-time processing of continuous streams of data in motion is called? (2 marks)

20. The statistical study of experiments in which multiple measurements are made on each experimental unit is referred to as? (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Question 21" and use the word processor document to save your answers to questions (a) to (d). Capture screenshots to demonstrate how you have performed the tasks below.

(a) Write R studio functions that will perform the following tasks:

(i) Plot a boxplot. (2 marks)

(ii) Specify the range of values allowed in X axis of a bar chart. (2 marks)

(iii) Subtract two vectors. (2 marks)

(iv) Compare two line plots. (2 marks)

(b) Write R studio code that will create the table below: (5 marks)

Patient Number	Weight	Patient fees
PBH2345	80	45
PBH4563	67	50
PBH5456	85	45
PBH678	76	56
PBH599	78	43

(c) Using R studio, create a bar chart called 'Patient' for the weight against patient fees. (4 marks)

(d) Write R programming statement to calculate the mode value of the patient fees. (3 marks)

Save "Question 21" and Upload. **(Total: 20 Marks)**

22. Create a word processing document named "Question 22" and use the word processor document to save your answers to questions (a) to (c). Capture screenshots to demonstrate how you have performed the tasks below.

(a) Give the Hadoop Spark statement to create and insert data into the table below. (10 marks)

AGE	INCOME	EDUCATION	EMPLOYMENT	MARITALSTATUS	GENDER	TARGET
25	25000	High	Full-time	Single	Male	Yes
32	34000	High	Part-time	Married	Female	Yes
45	65000	Medium	Full-time	Married	Male	Yes
22	18000	Low	Unemployed	Single	Female	No
23	55000	High	Full-time	Married	Male	Yes
28	32000	Medium	Part-time	Single	Male	No
40	60000	High	Full-time	Married	Female	Yes
30	28000	Low	Part-time	Single	Male	No
48	70000	High	Full-time	Married	Female	Yes
29	40000	Medium	Full-time	Single	Male	Yes

(b) Write the Hadoop Hive query that will display the Income, age, employment and marital status for all full time employees whose income is above 40000. (5 marks)

(c) Write the Hadoop hive statement that will count the number of male employees who are married and are aged 25 years and above. (5 marks)

Save and upload Question 22.

(Total: 20 marks)

23. Create a word processing document named "Question 23" and use the word processor document to save your answers to questions (a) to (d). Capture screenshots to demonstrate how you have performed the tasks below.

(a) Create an Excel document shown below, save it as a comma separated version (CSV) file named "**salesdata**". (4 marks)

	Lastname	Year	Country	Sales
2	Mwaura	2021	Kenya	$23,000
3	Anyango	2021	Kenya	$28,000
4	Ochieng	2022	Uganda	$25,000
5	Odhiambc	2023	Tanzania	$23,000
6	Makhulo	2021	Sudan	$29,000
7	Naisho	2022	Kenya	$35,000
8	Rioka	2023	Uganda	$41,000
9	Moraa	2020	Sudan	$47,000
10	Mwende	2020	Tanzania	$53,000
11	Wambui	2019	Tanzania	$59,000
12	Kyalo	2020	Kenya	$65,000
13	Njeri	2023	Ethiopia	$71,000
14				

(b) Type R Studio code that will import and display the data. (5 marks)

(c) Write R studio statement to create a scatter diagram between the year and sales. (5 marks)

(d) Using a spreadsheet application, find the summary of total sales by country. (6 marks)

Save "Question 23" and upload.

(Total: 20 marks)

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

MONDAY: 22 April 2024. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two (2) sections. SECTION I has twenty (20) Short Response Questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:
- **R Studio**
- **Spreadsheet program**
- **Hadoop software**

SECTION I (40 MARKS)

1. The characteristic of big data that deals with the trustworthiness and reliability of the data is known as_____. (2 marks)

2. The type of unsupervised machine learning technique which groups unlabeled data points based on their similarity and differences is referred to as_____. (2 marks)

3. The type of data analysis that helps businesses to understand complex relationships between linked entity data in a network is known as_____. (2 marks)

4. The type of social media analytics platform capability that involves the use of natural language processing technologies to help understand entities and relationships in order to reveal positive or negative attributes is referred to as_____. (2 marks)

5. The Hadoop component where the data containing code is used to process the entire data is referred to as_____. (2 marks)

6. The Hadoop component which performs job scheduling to make sure that the jobs are scheduled in the right place is known as_____. (2 marks)

7. The statistical term that **BEST** describes measures of central tendency, measures of variability and frequency distribution is referred to as_____. (2 marks)

8. A popular R package used in data analytics and machine learning to provide a unified interface and a set of functions for training and evaluating a wide range of machine learning models is known as_____. (2 marks)

9. Which term **BEST** describes R programming language as a free software that is easy to integrate with different applications and processes? (2 marks)

10. The type of probability distribution that gives the probability of an event happening a certain number of times within a given interval of time or space is referred to as_____. (2 marks)

11. A data visualisation library used in data analytics and data science to create interactive and visually appealing charts, graphs and dashboards is known as_____. (2 marks)

12. The type of data that is tabular with rows and columns that clearly define data attributes is called_____. (2 marks)

13. A data visualisation chart that displays multiple circles in a two dimensional plot is called_____. (2 marks)

14. Which spreadsheet function is used for counting cells in a range that satisfies a single condition? (2 marks)

15. _____ is a characteristic of infographics where images, icons, charts, graphs and other visual components are used to represent information. (2 marks)

16. The dataset that comprises of individual measurements that are acquired as a function of three or more than three variables is referred to as_____. (2 marks)

17. Which is the most important "V" of big data which comes from insight discovery and pattern recognition that lead to more effective operations and stronger customer relationships? (2 marks)

18. _____ intelligence refers to the combined knowledge, skills and problem-solving abilities of a group or community of individuals who collaborate and share information to achieve a common goal or make decisions. (2 marks)

19. Which term defines the roadmap of how data is generated, collected, processed, used and analysed to achieve business goals? (2 marks)

20. _____ databases are well-suited for the storage and retrieval of semi-structured data due to their flexible and schema-less nature. (2 marks)

SECTION II (60 MARKS)

21. Create a word document called "Question 21" to capture and save the screenshots for answers to questions (a) to (d) below.

Use the CSV data provided below to answer the question that follows:

Product	Category	Date	Sales	Profit
Product A	Electronics	2023-01-01	1500	300
Product B	Home & Garden	2023-01-02	1200	240
Product C	Electronics	2023-01-03	1800	360
Product D	Apparel	2023-01-04	800	160
Product E	Home & Garden	2023-01-05	900	180
Product F	Apparel	2023-01-06	700	140
Product G	Electronics	2023-01-07	2100	420
Product H	Home & Garden	2023-01-08	1000	200
Product I	Electronics	2023-01-09	1400	280
Product J	Apparel	2023-01-10	950	190

(a) Use a spreadsheet to create a CSV dataset and save it as "salesdata.csv". (5 marks)

(b) Using the dataset created in question 21 (a), write an R programming statement to load "salesdata.csv" data into a data frame, calculate and display the total profit. The code should be in a script named "Analytics.R". (5 marks)

(c) Write an R programming statement in the "Analytics.R" script to filter the data to include only products with sales greater than 1000 and a profit margin above 20%. (4 marks)

(d) Write an R programming statement in the "Analytics.R" script to create a bar chart to visualise the top 10 products with the highest sales from the dataset using R programming. (6 marks)

Save "Question 21" and upload.

(Total: 20 marks)

22. Create a word processing document named "Question 22" and use the word processor document to save your answers to questions (a) to (c).

(a) Write the Hadoop HIVE commands to create the Supplier table whose content is shown below. (5 marks)

Supplier table

	SUPPLIER NAME	COUNTRY	SALES AMOUNT
3	GREAT WALL DISTRIBUTORS	NIGERIA	$224,000
4	ORUBA OIL	GHANA	$201,000
5	BOYA METAL DEALERS	NIGERIA	$193,000
6	BIDII LAPTOPS	NIGERIA	$184,000
7	KELLER RETAILERS	GHANA	$173,000
8	BOTTOM RETAILERS	KENYA	$124,000
9	JOHSON AND JAY	UGANDA	$124,000
10	ZEAL HONEY	UGANDA	$123,000

(b) Type the Hadoop HIVE command to insert the data in the table created in (a) above. (5 marks)

(c) Using a spreadsheet application, prepare a worksheet using the data in the table below and use the data to find the total sales amount and units sold by each salesman. (10 marks)

Year	Month	Type	Salesman	Region	Sales	Units
2009	January	Chocolate	Joel	West	$2,395.50	1597
2009	January	Chocolate	Joel	West	$11,761.50	7841
2009	January	Ilara Yoghurt	Joel	West	$8,943.00	5962
2009	January	Chocolate	Joel	West	$2,395.50	1597
2009	January	Chocolate	Joel	West	$11,761.50	7841
2008	February	Ilara Yoghurt	Joel	West	$8,943.00	5962
2008	February	Ilara Yoghurt	Alice	Central	$14,596.50	9731
2008	February	Bacon	Alice	Central	$8,793.00	5862
2008	February	Ilara Yoghurt	Alice	Central	$14,596.50	9731
2008	February	Bacon	Alice	Central	$8,793.00	5862
2008	January	Chocolate	Pauline	North	$4,666.00	5623
2008	January	Chocolate	Pauline	North	$7,318.50	4879
2008	January	Chocolate	Pauline	North	$3,553.50	5623
2008	April	Chocolate	Pauline	North	$7,318.50	4879
2008	April	Popsicles	Victor	South	$3,553.50	2369
2008	January	Popsicles	Victor	South	$14,596.50	2369
2008	May	Ilara Yoghurt	Walter	Central	$8,793.00	9731
2008	May	Bacon	Walter	Central	$14,596.50	5862

Save "Question 22" and upload. **(Total: 20 marks)**

23. Create a word document called "Question 23" and use it to save solution to question (a) to (f).

Use the table given below to answer questions (a) to (f).

Name	Age	Score
John Kioko	30	85
Alice Maina	002	92
Bob Mokola	003	78
Eve Kuvuna	004	95

(a) Use a spreadsheet program to create a CSV dataset using the table provided above. Save your file as "exam.csv". (3 marks)

(b) Create an R programming script called "examanalysis.R". Type the statements to load data in (a) above into R and display its structure and the first row. (4 marks)

(c) Use "examanalysis.R" script to write R programming statement to calculate the mean and median score in the dataset. (4 marks)

(d) Use "examanalysis.R" script to write R programming statement to create a scatter plot to visualise the relationship between "Age" and "Score". (3 marks)

(e) Use "examanalysis.R" script to write R programming statement to calculate and display the total score for all individuals in the dataset. (2 marks)

(f) Write a statement to find the name and the age of the student with the highest score. (4 marks)

Save "Question 23" and upload.

(Total: 20 marks)

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

MONDAY: 21 August 2023. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two (2) sections. SECTION I has twenty (20) Short Response Questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:

- **R Studio**
- **Spreadsheet program**

SECTION I (40 MARKS)

1. Packages are libraries containing functions, data, and documentation that extend the capabilities of R. Which R package is commonly used for data manipulation and transformation? (2 marks)

2. The type of data analytics that looks at past data to give an account of what has happened before is known as _____. (2 marks)

3. What is the name given to data processing steps and transformations that move data from its source to its destination, typically involving extraction, transformation, and loading (ETL) process? (2 marks)

4. _____ is an open source framework from Apache and is used to store, process and anlyse data which are very huge in volume. (2 marks)

5. What is the type of clustering approach used in machine learning where the output provided is a probability likelihood of a data point belonging to each of the pre-defined number of cluster? (2 marks)

6. A suitable term that describes the activity of applying statistical and machine learning techniques to be able to infer any information from a text-mined data is known as _____. (2 marks)

7. The tools and technologies used to collect, anlyse, and present data to support business decision-making are referred to as _____. (2 marks)

8. In descriptive statistics, the collective term that include standard deviation, variance, minimum and maximum variables is called _____. (2 marks)

9. The shared intelligence, knowledge, and problem-solving abilities that emerge from the collaboration and collective efforts of groups or communities is known as _____. (2 marks)

10. What is the name given to data that does not follow the tabular structure associated with relational database or other forms of data tables but contains tags and metadata? (2 marks)

11. The process of combining multiple datasets from different sources to create a unified view is known as _____. (2 marks)

12. A table that summarises the performance of a classification model, showing the counts of true positives, true negatives, false positives, and false negatives is called _____. (2 marks)

13. The data analytics activity that is used to anlyse and display the geographically related data and present it in the form of maps is called _____. (2 marks)

14. Which is the interactive analysis tool used by businesses to track and monitor the performance of their strategies with quality Key Performance Indicators (KPIs)? (2 marks)

15. What is the name given to a tiny chart in a worksheet to show trends in a series of values, such as seasonal increases or decreases, economic cycles, or highlight maximum and minimum values? (2 marks)

16. A powerful data summarization tool in a spreadsheet application that allows you analyse, summarise, and present large amounts of data in a flexible and customisable manner is called _____. (2 marks)

17. Consider the R Studio syntax of a barplot as (H, xlab,ylab,main,names.arg,col). What does H represent? (2 marks)

18. The data visualization method that is used to display the distribution of continuous or interval data by grouping the data into bins or intervals and representing the frequency or count of data points within each bin is known as _____. (2 marks)

19. A measure of the "peakedness" or "flatness" of a distribution as used in statistical measures is known as _____. (2 marks)

20. The type of data mining where a transaction and the relationship between its items are used to identify a pattern is referred to as _____. (2 marks)

SECTION II (60 MARKS)

21 Create a word document called "Question 21" and use it to save solutions to questions (i) to (vii).

(i) Create a data frame to store three of your favorite friends' names in an object called "mypals" with their first names, age and counties. Display the result on a console screen. (4 marks)

(ii) Define a new variable called "gender" within the data frame "mypals", to include your friends gender and display the update data frame (3 marks)

(iii) Using indexing, display the second row from "mypals" data frame. (3 marks)

(iv) Create a subset for all the friends whose age is greater or equal to 20 years and display the results on the console screen. (3 marks)

(v) Display the data set mypals, sorted in ascending order by age. (3 marks)

(vi) Display the summary statistics for the data frame "mypals". (1 mark)

(vii) Calculate the mean and totals of the column Age and display the output. (3 marks)

Save "Question 21" and upload. **(Total: 20 marks)**

22 Create a word document called "Question 22" and use it to save solutions to questions (i) to (vii).

(i) Using R programming data frame, create the data object called "Employees", using the data provided below and display the results. (8 marks)

Employee ID	Name	Department	Age	Salary	Years Of Experience
1	Ted	Sales	32	50000	5
2	James	HR	28	45000	3
3	Ann	Finance	35	60000	8
4	Peter	Marketing	30	55000	4
5	Alex	IT	38	650000	10
6	Emily	Sales	29	48000	2
7	Lisa	HR	31	58000	6
8	Emma	Finance	33	62000	7
9	Lee	Marketing	36	43000	9
10	Mark	IT	27	52000	1

(ii) Write R code to display the number of employees in the dataset. (2 marks)

(iii) Write R code to display the average salary paid to all the employees. (2 marks)

(iv) Write R code to display the highest paid salary among all the employees. (2 marks)

(v) Write R code to display the number of employees working in each department. (2 marks)

(vi) Write R code to return the total years of experience across all employees. (2 marks)

(vii) Write R code to display the most paid employee by their name. (2 marks)

Save "Question 22" document and upload.

(Total: 20 marks)

23 Create a word document called "Question 23" and use it to save solutions to questions (i) to (iii).

A Retail shop sale for different products at different dates is captured as shown below.

- **Product Category** = "Electronics", "Clothing", "Home & Kitchen", "Books", "Electronics", "Clothing", "Home & Kitchen", "Books"
- **Date** = "2023-01-01", "2023-01-02", "2023-01-03", "2023-01-04", "2023-01-05", "2023-01-06", "2023-01-07", "2023-01-08"
- **Sales** = 5000, 3000, 7000, 2000, 6000, 4000, 8000, 2500
- **Price** = 100, 50, 80, 30, 120, 60, 90, 25

Required:

(i) Create a data frame called "Sales" to store the above information and display the output. (8 marks)

(ii) Import the ggplot2 library and write R programming code to visualise the sales by product category using a bar chart. (6 marks)

(iii) Write R programming code to visualise the sales trends over time using a line chart. (6 marks)

Save "Question 23" document and upload. **(Total: 20 marks)**

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

MONDAY: 24 April 2023. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) Short Response Questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:

- **R Studio**
- **Spreadsheet program**

SECTION I (40 MARKS)

1. Which R programming function provides a method for displaying the structure of a data frame to give a wealth of information about the dataset during the pre-processing stage? (2 marks)

2. The table and storage management layer for Hadoop that supports different components available in Hadoop ecosystems to easily read and write data from the cluster is referred to as? (2 marks)

3. In data analytics, the process analysing entire dataset can be time consuming and therefore a subset of observations from a dataset should be selected. What is the name given to the process of choosing a subset of observation from the entire data set for analytics purpose called? (2 marks)

4. The machine learning technique that trains a model to identify hidden models or intrinsic structure in the input data is referred to as? (2 marks)

5. The component of Spark ecosystem that is used for implementing graphs and graph-parallel computation is called? (2 marks)

6. Machine learning is a subset of artificial intelligence concerned with the development of models that can learn from data without being explicitly programmed. The process of dividing a dataset into the training and the test subsets is called? (2 marks)

7. The data analytics and data management characteristics that refers to the quality of data as being uncertain, incomplete, or inconsistent is known as? (2 marks)

8. The type of data analytics that uses statistical and machine learning techniques to forecast what might happen in the future is referred to as? (2 marks)

9. The Microsoft Excel add-in program that is helpful when performing what if analysis is referred to as? (2 marks)

10. The government, individuals and organisations, may provide publicly available data to anyone for access, use and share usually stored in formats such as CSV and JSON, such data is referred to as? (2 marks)

11. The component of HDFS (Hadoop Distributed File System) responsible for managing the namespace and access to files is called? (2 marks)

12. Which characteristic of a dataset is referred by the statement “degree to which a dataset comprises of all the necessary information and is free of missing value”? (2 marks)

13. What is the method of data analysis that involves organising and summarising data in order to identify trends and patterns? (2 marks)

14. The dataset that comprises of individual measurements that are acquired as a function of three or more than three variables is referred to as? (2 marks)

15. Type an R Programming code, to generate a line chart for a dataset object called “employees” for the feature “salary” (2 marks)

16. For the data given by D <- c (8,13,27,3,45), give the R-Studio command that will be used to plot a bar chart. (2 marks)

17. R programming support data storytelling though data visualisation techniques. Cite the R programming library that is used to create interactive plots such as scatter plots, line plots and bar charts. (2 marks)

18. In big data, data that is collected and organised over time, such as stock prices, weather data, and website logs. State the name given to such a datasets? (2 marks)

19. State the name given to Apache Spark partitions that split the data in MapReduce into smaller, logical divisions? (2 marks)

20. Which is the R-studio logical operator that combines each element of the first vector with the corresponding element of the second vector and gives a true output if both the elements are true? (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named “Question 21”, use the word processor document to save your answers to questions (a) to (d).

(a) Give the screenshot showing how to enable the “Solver Add In” in Excel to support data analysis. (2 marks)

(b) Create the following data set in Microsoft Excel and save it as “Stores”: (3 marks)

STORE NUMBER	SALES
B234S34	40000
R657D45	60000
X757R34	57000
R656E45	45000
W657V94	43000
B657C48	54000
V677R87	56000
F667U49	32000
A786C71	41000

(c) Use the data set to find the descriptive statistics values of mean, standard error, median, range, sum and count using Microsoft Excel. Capture and display the screen shots. (6 marks)

(d) (i) Create the following data set in Microsoft Excel and save it as "Enrolment". (3 marks)

YEAR	NO. OF STUDENTS ENROLLED	NO. OF STUDENTS GRADUATED
2001	323	239
2002	454	324
2003	521	564
2004	450	356
2005	324	287
2006	376	300
2007	436	400
2008	476	450
2009	368	350
2010	329	300
2011	453	401
2012	398	356
2013	350	342
2014	447	436

(ii) Using regression analysis, estimate the relationships between the number of students enrolled and the number of students who graduated. (6 marks)

Save "Question 21" document and upload.

(Total: 20 marks)

22. Create a word processing document named "Question 22", use the word processor document to save your answers to questions (a) to (g).

Create an R script called "Question22.R" and use it to answer the questions that follow:

(a) Using R, create a 4 by 4 matrix of integer elements between 20 and 35 called "num_mat" and print the result on the console screen. (4 marks)

(b) Create an identity 4 by 4 matrix called "identity mat" and print it on the console screen. (2 marks)

(c) Create a 4 by 4 matrix of ones for all elements called "ones_mat" and print it on the console screen. (2 marks)

(d) Using R, create another 4 by 4 matrix of integer elements between 10 and 25 called "num_mat2" and print it on the consoles. You should then display the product of multiplying the two matrices "num_mat2" and "num_mat2". (4 marks)

(e) Write an R script to display the transpose of the matrix "num_mat2" and display it on the console screen. (2 marks)

(f) Perform the Principal Components Analysis (PCA) on the matrix "num_mat" and display the result on the console screen. (2 marks)

(g) Create the data visualisation for matrix "num_mat2" using a "Heat map" and an "image plot" and capture their displayed output. (4 marks)

Save "Question 22" document and upload.

(Total: 20 marks)

23. Create a word processing document named "Question 23" use the word processor document to save your answers to questions (a) to (b).

(a) Write an R-Studio code to draw a scatter plot to display the relationship between two variables namely unit code and unit score, and to plot one dot for each observation. The program should compare the two sets of variables given below using a function. Capture and display the code and output.

Set 1

Unit code = 6,8,7,5,3,3,10,4,13,12,10,6,2

Unit score = 79,82,78,98,80,93,77,64,88,75,85,96,91

Set 2
Unit code = 6,3,9,2,16,9,8,9,11,4,10,3,7,14,13,6
Unit score = 76,82,83,106,80,99,90,95,94,99,89,92,75,90,85,78

(9 marks)

(b) Create an R script called "Question23.R" and use it to answer the questions that follow:

(i) Use a relevant function to display all the In-built datasets that comes with RStudio. (1 mark)

(ii) Use a relevant function to load and display the dataset called "mtcars". (1 mark)

(iii) Display the various statistical measures of the dataset "mtcars". (1 mark)

(iv) Display the structure of the dataset "mtcars". (1 mark)

(v) Display the scatterplot for the dataset "mtcars". (1 mark)

(vi) Generate a pairwise correlation matrix for the variables in above dataset and display on the console screen. (2 marks)

(vii) Fit a linear regression model between "mpg" and "wt" variables and display the model results. (2 marks)

(viii) Convert and display the mpg variable to its natural logarithm. (1 mark)

(ix) Create a scatterplot between "mpg" and "wt" variables and display the result. (1 mark)

Save "Question 23" document and upload.

(Total: 20 marks)

……

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL III

DATA MANAGEMENT AND ANALYTICS

MONDAY: 5 December 2022. Afternoon Paper. **Time Allowed: 3 hours.**

Answer ALL questions. This paper has two sections. SECTION I has twenty (20) short response questions of two (2) marks each. SECTION II has three practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:

- **R Studio**
- **Spreadsheet program**

SECTION I (40 MARKS)

1. The table and storage management layer for Hadoop that supports different components available in Hadoop ecosystems to easily read and write data from the cluster is referred to as? (2 marks)

2. Name any one core component of Hadoop. (2 marks)

3. Which element stores any function, object or value that is created during an R studio session? (2 marks)

4. The term that best describes to how quickly data is generated and how quickly that data moves is referred to as? (2 marks)

5. When relating the dependent and independent variables in a bivariate dataset, the relationship which states that if the independent variable increases then the dependent variable would also increase and vice versa is referred to as? (2 marks)

6. The type of metadata that relates to the technical source of a digital asset is referred to as? (2 marks)

7. Which term best describes the set of techniques that analyse current and historical data to determine what is most likely to happen or not to happen is referred to as? (2 marks)

8. The measure of dispersion that measures the spread of data about the mean value is referred to as? (2 marks)

9. Which is the third stage of the big data analytics life cycle? (2 marks)

10. The big data visualisation tool which is a representation of complex datasets and frequency of numerical data displayed through bars is referred to as? (2 marks)

11. The incremental copy of data for establishing a point in time in which the system can be rolled back in case of system failure in a HDFS is called? (2 marks)

12. When working with a dataset in R programming. It is advisable to have the dataset in the folder with the project. Which built-in R function is used set manually the path where the dataset is located? (2 marks)

13. In Data analytics, we can use scatter plots for crime and detection in a banking stem by looking at the data points distribution on the X-Y plane. What name is given to the stray data points from the general distribution of data points? (2 marks)

14. Illustrate how to create a list in R programming for three of your favorite colors. (2 marks)

15. The name given to the statistical technique, which is used to quantify the relationship between predictor and response variable is? (2 marks)

16. Which type of Distribution is depicted by the figure 1 shown below? (2 marks)

$$P(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

Figure 1

17. In machine Learning, the difference between the actual data and predicted values, predicted by the regression line of best fit is known as? (2 marks)

18. You have been approached by the owner of an e-commerce based business as a data analyst to develop a machine learning model to suggest to customers, other items that they can buy, based on their previous purchases and their spending habit. Which machine learning technique is best based to do the task? (2 marks)

19. The data visualisation method that comprises of nodes and edges and allows users to easily understand relationships in data is referred to as? (2 marks)

20. State how you would load a dataset called "**students.csv**" in R programming, using the relevant function if the dataset is stored directly on drive C. (2 marks)

SECTION II (60 MARKS)

21. Create a word processing document named "Employees" and use the word processor document to save your answers to questions (a) to (f).

Use the data table shown in table 1 below to answer the questions that follow:

Name	gender	age	department	salary
John Peter	M	42	ICT	130000
Ann Lukas	F	34	Sales	45000
Annette Jones	F	34	ICT	135000
John Ford	M	35	Sales	60000
Mark Robert	M	34	Accounts	138000

Table 1

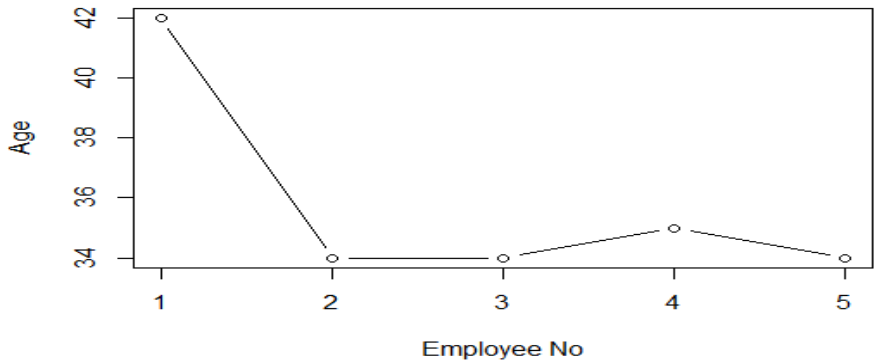
(a) Write an R program, to create a dataset called "Staff", to store a list of employees details using the data provided in table 1. (8 marks)

(b) Use an appropriate R built-in function to display all the data frame created as a table. (3 marks)

(c) Write a script to examine, how the data in the data frame "Staff" is organised. (3 marks)

(d) Write a script to display the last two records from the dataset created. (3 marks)

(e) Create the line graph shown below using the dataset created in question (a) above. (5 marks)



(f) Write a script to calculate the total salary paid to all the employees. (3 marks)

Capture screenshots to demonstrate how you have performed the above task.

Upload "Employees" document.

(Total: 25 Marks)

22. Create a word processing document named "Supplier" and use the word processor document to save your answers to questions (a) and (b).

(a) Create the dataset shown in table 2 using Excel and save it as "Student" in "**Data**" folder on the desktop. **(**3 marks**)**

SUPPLIER NAME	COUNTRY	SALES AMOUNT
XYZ WHOLESALERS	KENYA	$40,000
ZEAL HONEY	UGANDA	$123,000
BOTTOM RETAILERS	KENYA	$124,000
COCK AND HEN	KENYA	$123,000
BIDII LAPTOPS	NIGERIA	$184,000
KELLER RETAILERS	GHANA	$173,000
BOX AND BOX	NIGERIA	$64,000
MULUMBU AGROVET	GHANA	$105,000
JOHSON AND JAY	UGANDA	$124,000
BOYA METAL DEALERS	NIGERIA	$193,000
ORUBA OIL	GHANA	$201,000
GREAT WALL DISTRIBUTORS	NIGERIA	$224,000

Table 2

(b) Use the data set in (a) above to perform the following data analyis tasks in Excel:

(i) Write the excel function to give an incentive of 10% if the Sales amount is greater than $100,000 otherwise no incentive is given. (6 marks)

(ii) Create a pivot chart for the data in the table and filter the pivot chart by sales amount. (6 marks)

(iii) Write an excel function that will sum the sales values for the country called Nigeria. (5 marks)

Capture screenshots to demonstrate how you have performed the above task.

Upload Supplier document.

(Total: 20 Marks)

23. Create a word processing document named "Watering" and use the word processor document to save your answers to questions (a) to (e).

The table below shows the amount of water (in liters) required for a particular number of plants.

Study the table and answer the questions that follows:8

A	B
liters	**plants**
0	0
1	3
2	6
3	9
4	12
5	15
6	18
7	21
8	24
9	27
10	30

(a) Using the data provided in the table above create a comma separated value dataset called "water.csv". (3 marks)

(b) Create a script in R called "water.R" and use it to load the dataset created using the relevant R function into an object called "waterdata". (3 marks)

(c) Create a staircase graph using the dataset "water.csv". (2 marks)

(d) Create a boxplot of the two attributes in the dataset "water.csv". (3 marks)

(e) Visualise the correlation between the two attributes in the dataset. (4 marks)

Capture a screenshot to demonstrate how you have performed the above task.

Upload Watering document.

Total: 15 Marks)

……