DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

MONDAY: 1 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:

- **A computer**
- **R**
- **SQL Server Management Studio**
- **Jupiter Notebook**
- **Python program**

SECTION I (40 MARKS)

1. In data warehouse implementations, what aspect of a phased delivery approach minimises disruptions and supports milestone tracking? (2 marks)

2. The data mining technique that partitions a set of data objects into groups of similar objects is called ____________. (2 marks)

3. Within a federated data warehouse framework that relies on multiple independent data marts, how do these decenrtralised structures increase the risk of inconsistent reporting and fragmented governance? (2 marks)

4. Which type of data is used to support prediction in data mining? (2 marks)

5. In a decision tree, the internal nodes represent a test on an attribute, what represents the leaf nodes? (2 marks)

6. Which technique in data mining is used to predict a continuous value such as a future sales amount? (2 marks)

7. When comparing enterprise-scale data warehouses with department-specific data marts, how do their metadata management requirements differ in terms of complexity, governance and overall scope? (2 marks)

8. When performing OLAP operations, a user wants to view the data from a different perspective by interchanging the rows and columns of a data cube. Which OLAP operation should the user perform? (2 marks)

9. In distributed data warehouse architectures that employ polyglot persistence and real-time processing, what is the primary technical challenge in maintaining synchronisation of extraction and transformation metadata? (2 marks)

10. Within OLAP operations, which process increases data granularity by navigating to progressively lower levels within a dimension hierarchy? (2 marks)

11. The SQL keyword used to sort the result set in ascending or descending order is known as ___________________. (2 marks)

12. Which data mining technique leverages multiple interconnected layers to model and capture complex, non-linear relationships in datasets? (2 marks)

13. Which enterprise data warehouse platform is best suited for analysing semi-structured and unstructured data due to its flexible architecture? (2 marks)

14. In modern data mining, there is a growing trend that focuses on analysing data collected from mobile devices and other personal gadgets, this trend is called ____________________. (2 marks)

15. A database schema that normalises dimension tables into multiple related tables to reduce redundancy is known as ______________. (2 marks)

16. Which dimensional schema uses a denormalised structure with a central fact table surrounded by dimensions to optimise query performance? (2 marks)

17. In the context of data warehousing, the process of extracting data from multiple sources, transforming it into a suitable format and loading it into a central repository is known as ____________________. (2 marks)

18. Within OLAP operations, which process reduces data granularity by aggregating values as the hierarchy is navigated upward? (2 marks)

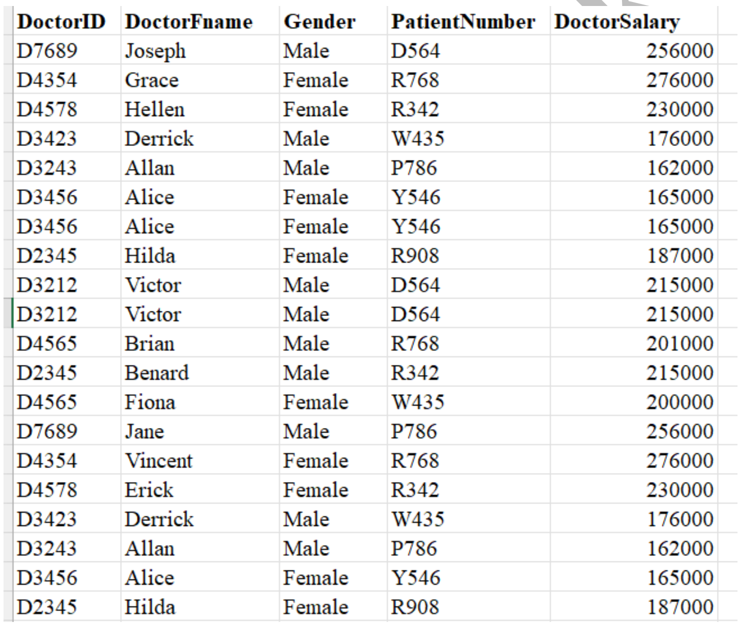
19. The data mining application that helps businesses identify groups of customers with similar buying habits for targeted marketing is known as ____________________. (2 marks)

20. In data mining, the challenge of efficiently handling and processing massive datasets is referred 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 (f).

(a) Create the dataset below using Microsoft Excel program and save it as "mydoctor.csv" (5 marks)



DoctorID	DoctorFname	Gender	PatientNumber	DoctorSalary
D7689	Joseph	Male	D564	256000
D4354	Grace	Female	R768	276000
D4578	Hellen	Female	R342	230000
D3423	Derrick	Male	W435	176000
D3243	Allan	Male	P786	162000
D3456	Alice	Female	Y546	165000
D3456	Alice	Female	Y546	165000
D2345	Hilda	Female	R908	187000
D3212	Victor	Male	D564	215000
D3212	Victor	Male	D564	215000
D4565	Brian	Male	R768	201000
D2345	Benard	Male	R342	215000
D4565	Fiona	Female	W435	200000
D7689	Jane	Male	P786	256000
D4354	Vincent	Female	R768	276000
D4578	Erick	Female	R342	230000
D3423	Derrick	Male	W435	176000
D3243	Allan	Male	P786	162000
D3456	Alice	Female	Y546	165000
D2345	Hilda	Female	R908	187000

Required:
Write a python code to:

(b) Import the necessary libraries and load the data into the "**doctordata**" dataframe. (3 marks)

(c) Remove any duplicate rows and display the result. (4 marks)

(d) Calculate and print the median of the 'doctor salary' after cleaning. (3 marks)

(e) Classify doctor's salary by gender. (2 marks)

(f) Display the doctor IDs for all males whose patient number is W435. (3 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 (c).

(a) Write the SQL statements to create and insert data into the student and book tables: (10 marks)

STUDENT

StudentID	**StudentFname**	**Bookcode**
KK900	Joseph	BBK456
KK1024	Anne	CBA345
KK3456	Boaz	BBK765
KK8767	Hellen	BBF456

BOOK

Bookcode	**BookAuthor**	**Bookprice**
BBK456	Alice	$342
CBA345	Anita	$564
BBK765	James	$665
BBK878	Joseph	$786
BBF456	Victor	$657

(b) Write an SQL statement that will display the book code, book price and student first name for all students with “o” as a character in their first name. (5 marks)

(c) Write an SQL statement to find the average book price for each author and display only those authors whose average book price is greater than 600. The output should include Book Author and Average Price. (5 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 processing document named “Question 23” and use the word processor document to save your answers to questions (a) to (g) and save it in a folder “December_25”.

(a) Use the table below to create a CSV file and save it in “December_25” folder as “hospital_data.csv”. (5 marks)

ADMISSION_DATE	**DEPARTMENT**	**DOCTOR**	**PATIENT**	**DIAGNOSIS**	**DAYS_ADMITTED**	**COST_PER_DAY**	**TOTAL_COST**
2025-01-05	Cardiology	Dr. Kim	John Doe	Hypertension	5	2500	12500
2025-01-12	Neurology	Dr. Patel	Sarah Lee	Migraine	3	3000	9000
2025-02-01	Orthopedics	Dr. Wong	David Roy	Fracture	7	2000	14000
2025-02-10	Pediatrics	Dr. Green	Emily Clark	Pneumonia	6	1800	10800
2025-03-03	Cardiology	Dr. Kim	Michael Smith	Heart Attack	10	2700	27000
2025-03-18	Neurology	Dr. Patel	Olivia Brown	Stroke	8	3200	25600
2025-04-05	Orthopedics	Dr. Wong	Linda Johnson	Arthritis	4	2200	8800
2025-04-20	Pediatrics	Dr. Green	Daniel Wilson	Asthma	5	1500	7500

Required

(b) Write an R Script to:

(i) Calculate the total hospital cost for each DEPARTMENT and identify the department with the highest total cost. (3 marks)

(ii) Count the number of patients admitted by each DOCTOR and identify the doctor with the highest admissions. (3 marks)

(iii) Compute the average days admitted for each DIAGNOSIS and determine which diagnosis corresponds to the longest hospital stay. (3 marks)

(iv) Identify the patient with the highest hospital bill. (3 marks)

(v) Group hospital admissions by month, calculate the total hospital costs per month and determine which month recorded the highest cost. (3 marks)

Save “Question 23” and upload. **(Total: 20 marks)**

……………………………………………………………………

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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:

- **A computer**
- **Ms Access 2016**
- **SQL Server Management Studio**
- **Jupyter Notebook**
- **Python program**

SECTION I (40 MARKS)

1. Which schema type uses multiple fact tables sharing dimension tables to represent complex business processes? (2 marks)

2. The application of data mining in identifying fraudulent behavior in network systems is called _______________. (2 marks)

3. Which cloud-based tool by Amazon is commonly used to host enterprise-scale data warehouse? (2 marks)

4. A crucial aspect of database administration, ensuring that your databases are running efficiently, reliably and securely. Involving continuously tracking key metrics, analysing trends and proactively identifying and resolving issues before they impact users or business operations is known as? (2 marks)

5. Which industry provides spatial data mining so that geographical and location-based data can be analysed by those who need it? (2 marks)

6. The type of SQL operation used to perform OLAP functions such as aggregation and summarisation is called _________________. (2 marks)

7. Which normalised schema design in dimensional modeling prioritises data integrity by organising dimensions into hierarchical sub-dimensions? (2 marks)

8. A technology used to perform high speed, complex queries and multidimensional analysis on large volumes of data is known as _______________. (2 marks)

9. Which warehouse architecture combines subject-oriented data marts with a central enterprise data warehouse for flexible deployment? (2 marks)

10. A schema that supports multiple fact tables sharing dimension tables is called ____________. (2 marks)

11. Which preprocessing technique standardises numeric data to a shared scale, often required before applying gradient-based machine learning algorithms? (2 marks)

12. The method in Python that helps identify the correlation between numerical variables in a dataset is ____________. (2 marks)

13. Which mathematical transformation is commonly applied to right-skewed data to stabilise variance and approximate normality? (2 marks)

14. What is the fundamental building block of a neural network, representing a processing unit that receives inputs, performs a computation and produces an output? (2 marks)

15. Which property of a machine learning model refers to its capacity to be understood and trusted by human stakeholders? (2 marks)

16. A meteorological station in South Africa records daily temperatures, rainfall amounts and wind speeds. These measurements can take on any value within a given range. What type of data is characterised by such kind of recordings?
(2 marks)

17. The type of analytics that uses historical patterns to forecast future events or behaviors is known as ___________. (2 marks)

18. Which OLAP operation restricts a multidimensional cube to a specific value on a single axis, enabling focused analysis? (2 marks)

19. Which data mining functionality helps compare the general features of target class data objects with the general features of objects from one or a set of contrasting classes? (2 marks)

20. The process of identifying a numeric outcome based on historical patterns is known as _____________. (2 marks)

SECTION II (60 MARKS)

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

(a) Using the data variables for a student below, write the python code that will produce a linear regression between the two variables and show the result generated. (6 marks)

height = 5.7,6.3,4.5,4.2,5.4,6.3,6.7,6.0,4.9,5.9,6.3,6.2,6.1

weight = 87,86,85,73,81,86,85,65,74,63,80,61,59

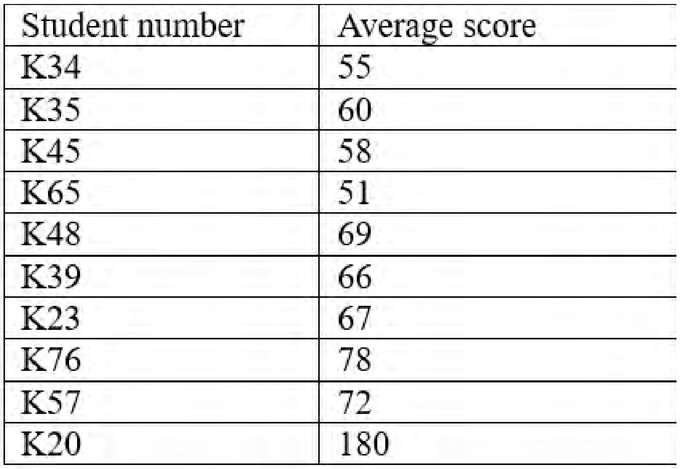
(b) Write the python functions that will perform the following tasks. Consider a dataframe called WDM.

(i) Remove cells with null entries in a dataset. (2 marks)

(ii) Discover duplicate entries in a dataset. (2 marks)

(iii) Split a dataset of 12 rows to have 8 rows for training and 4 rows for testing. (2 marks)

(iv) Write the python code to create the student dataset shown below, use z-score method to detect any outliers and visualise the result using a bar chart. (8 marks)



Student number	Average score
K34	55
K35	60
K45	58
K65	51
K48	69
K39	66
K23	67
K76	78
K57	72
K20	180

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 (f) and save it in the folder "KASNEB_08_25".

(a) Using a database application of your own choice, create a new database named "StudentDB" using SQL statements. (2 marks)

(b) Use SQL to create "DimStudent" and "FactEnrollment" tables shown below. (5 marks)

Tablename	**Column**	**Data Type**	**Description**
DimEmployee	StudentID	int	Primary Key
	FullName	varchar(100)	Employee full name
	Major	varchar(100)	Student major Field of study
	YearLevel	varchar(20)	Student's Year
FactEnrollment	EnrollmentID	int	Primary Key
	StudentID	int	Foreign Key
	CourseCode	varchar(20)	Unique Course Code
	Grade	varchar(2)	Final grade e.g. A,B,C,D)
	EnrollmentDate	date	Date of Course Enrollment

(c) Write SQL statement to populate the "DimStudent" with the following records. (2 marks)

StudentID	**Full Name**	**Major**	**Year**
1	John Silas	Data Science	Second
2	Sally Mary	Finance	Third
3	Jane Salome	Artificial Intelligence	First

(d) Write an SQL statement to populate the "FactEnrollment" with the following records. (3 marks)

EnrollmentID	**StudentID**	**Course code**	**Grade**	**Enrollmentdate**
01	1	DS201	A	2024-01-01
02	1	DS202	C	2024-01-01
03	2	FN300	B	2023-01-01
04	2	FN301	B	2023-01-01
05	3	AI001	C	2025-01-01
06	3	AI002	A	2025-01-01

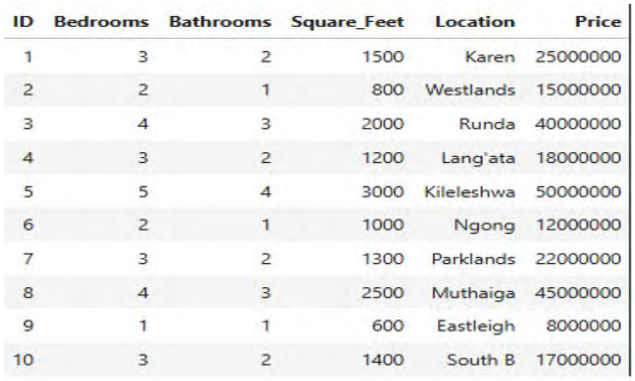
(e) Write SQL statement to display all the enrollment data. (2 marks)

(f) Write an SQL query to retrieve the top-performing student(s) per major, based on the average grade achieved across all their enrolled courses. Assume the grades are ranked A = 4, B = 3, C = 2, and D = 1. Display the FullName, Major, Average Grade Score and YearLevel for each top-performing student per major. (6 marks)

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

23. Create a word processing document named "Question 23" to save your answers to questions (a) to (g).

(a) Create a CSV file called "HouseSales.csv" containing the data below. Use the data to answer questions (b) to (g).



ID	Bedrooms	Bathrooms	Square_Feet	Location	Price
1	3	2	1500	Karen	25000000
2	2	1	800	Westlands	15000000
3	4	3	2000	Runda	40000000
4	3	2	1200	Lang'ata	18000000
5	5	4	3000	Kileleshwa	50000000
6	2	1	1000	Ngong	12000000
7	3	2	1300	Parklands	22000000
8	4	3	2500	Muthaiga	45000000
9	1	1	600	Eastleigh	8000000
10	3	2	1400	South B	17000000

(4 marks)

(b) Import the Python libraries and load data from the "HouseSales.csv" dataset into a pandas DataFrame object and display the data. (2 marks)

(c) Write a python code to create objects X and y, for feature(Bedroom) and target (Price) variables respectively, to be used in a prediction model. (2 marks)

(d) Write python code to split the data into training and testing sets in the ratio of 80/20 and display the training set for the features. (4 marks)

(e) Write a Python code to train your model using linear regression. (3 marks)

(f) Write a python code to make the prediction model developed. (2 marks)

(g) Write the code to evaluate your model using the Mean Squared Error and display the result. (3 marks)

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

……………………………………………………………………

kasneb

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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:

- **A computer**
- **Ms Access 2016**
- **SQL Server Management Studio**
- **Jupyter Notebook**
- **Python program**

SECTION I (40 MARKS)

1. A dataset with a large number of variables can present a number of issues within the data mining techniques. In data mining, the process of selecting relevant features and reducing dimensionality is known as ______________. (2 marks)

2. What name is given to the OLAP operation in multidimensional data that selects two or more dimensions from a given cube and provides a new sub-cube? (2 marks)

3. Data warehouse design used multidimensional modeling, unlike the operational databases that use ER modeling. The conceptual, and abstract design for organising data in a data warehouse is called _______________. (2 marks)

4. The descriptive statistics data mining task that divides a dataset into multiple groups that share some common characteristics such as a partitioning of the market for a product based on customer profiles is known as __________________. (2 marks)

5. Which characteristic of data warehouse requires only two operations of data loading and access to ensure the historical data remains intact? (2 marks)

6. The type of data mining approach used in the data mining process to predict equipment failures before they occur in manufacturing process control is referred to as ___________________. (2 marks)

7. A file contains dates as integers 20100919211037 instead of the expected format 2010-09-19 21:10:37. What name is given to the process that ensures that data is in the correct format as used in warehousing and data mining before the modeling step? (2 marks)

8. Profiles of customers discontinuing a particular product or service can be analysed and prediction models generated for customers likely to switch to other competitors. This data mining technique is known as _________________. (2 marks)

9. Which type of data mart combines elements of dependent and independent data marts? (2 marks)

10. Which is the mathematical model that is used to predict a continuous response variable such as the sales volume in the next twenty months? (2 marks)

11. The method used for grouping individuals in a population to discover structure in data in such a way that the individuals within a group are close to each other but dissimilar from individuals in the other groups is known as _________________. (2 marks)

12. Which is the Structured Query Language (SQL) clause that is used to filter specific information from a vast repository of data as used in data aggregation in the data warehouse? (2 marks)

13. In many situations, the data to be used in a data mining project may not be represented as a table such as a collection of documents or a sequence of page clicks for a particular website. This type of data is classified as __________________. (2 marks)

14. The process of reading and understanding the source data, and copying the data needed into the staging layer of the data warehouse system for further manipulation is known as _______________. (2 marks)

15. In data mining, summary statistics should be calculated for the whole dataset and each class individually. The method used in Python programming to calculate the summary statistics is known as _________________. (2 marks)

16. The physical data warehouse installed and maintained on a company's server is referred to as ______________. (2 marks)

17. The technique used in warehousing and data mining to create new variables from existing columns of data such as creating a new column age from an existing column month is referred to as __________________. (2 marks)

18. The logical design of a data warehouse is known as a schema. In a star schema model which tables surround the central fact table? (2 marks)

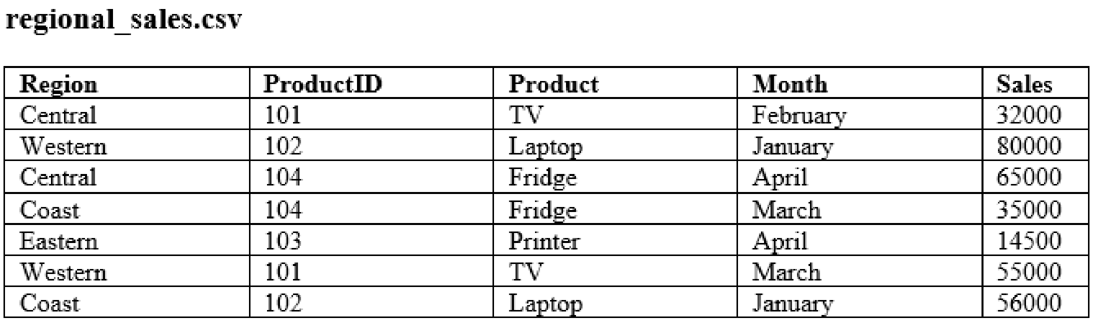
19. An information-processing unit that is fundamental to the operation of a neural network motivated by how the human brain computes is known as _________________. (2 marks)

20. The quantitative data points being analysed, representing the metrics or key performance indicators (KPIs) of interest such as total revenue, units sold, and average selling price in a data warehouse are called____________. (2 marks)

SECTION II (60 MARKS)

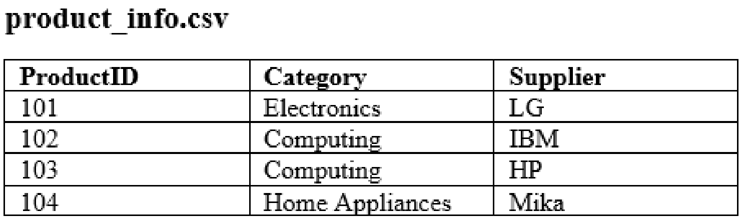
21. Create a word processing document named "Question 21" to save your answers to questions (a) to (e).

(a) Create the two named datasets provided below and save them in the appropriate location as CSV files. (6 marks)



regional_sales.csv

Region	ProductID	Product	Month	Sales
Central	101	TV	February	32000
Western	102	Laptop	January	80000
Central	104	Fridge	April	65000
Coast	104	Fridge	March	35000
Eastern	103	Printer	April	14500
Western	101	TV	March	55000
Coast	102	Laptop	January	56000



product_info.csv

ProductID	Category	Supplier
101	Electronics	LG
102	Computing	IBM
103	Computing	HP
104	Home Appliances	Mika

(b) Using pandas, load the datasets into data Frames called "regional_data" and "product_data" and display the results. (4 marks)

(c) Merge the two DataFrame created in 21 (b) using pandas to transform them into a single DataFrame called "merged_data" based on the "ProductID" column of both DataFrame. (4 marks)

(d) Using the merged DataFrame, filter the details of all the products, whose sales are greater than 60000 from the "merged_data" into a DataFrame called "filtered_data" and display the results. (3 marks)

(e) Load the filtered data to a CSV file called "warehouse.csv" which is not indexed. (3 marks)

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

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

(a) Use the table below to create a CSV file and save as "loan_data.csv". (4 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) Import the Python libraries and load loan data from the CSV file named "loan_data.csv" into a pandas DataFrame called "loan_data". (2 marks)

(c) Write a Python code to preprocess the "LoanStatus" column in order to encode it to binary values. (3 marks)

(d) Write a Python code to split the data into training and test sets. (3 marks)

(e) Write a Python code to train a decision tree classifier on the training data. (3 marks)

(f) Write a Python code to evaluate the classifier on the test data and print the accuracy. (2 marks)

(g) Write a Python code to visualise the decision tree. (3 marks)

Save Question 22 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 (b).

(a) Using SQL, create and insert data into the Employee and customer tables below. (10 marks)

Employee table

Empno	EmpLname	Age	Jobtitle	Salary
2567	Migwi	24	Programmer	120000
2570	Mutuku	32	Teacher	68000
2571	Musau	23	Cleaner	34000
2572	Wakayu	26	Secretary	54000
2573	Livanze	34	Accountant	102000
2574	Waigwa	34	Cleaner	26000
2575	Waema	26	Secretary	38000
2576	Matundura	34	Teacher	78000

Customer table

CustomerNo	CustomeFname	City	Empno
YT6998	Benard	Pretoria	2572
RF2344	Maurice	Pretoria	2573
RY5643	Juliet	Kampala	2573
GH4563	Noel	Kampala	2575
RT8976	Leonard	Durban	2576
LK4356	Hellen	Pretoria	2572
GH4568	Grace	Mombasa	2575
YT7998	Juliet	Pretoria	2572

(b) Write the SQL statement that uses the cube function to find the average salary of employees whose age is above 26 years. Group the result by employee last name and job title. (5 marks)

(c) Write the SQL statement that will display the customer number, city, employee number and employee job title for all customers from the city of Pretoria and whose employee job title is secretary. (5 marks)

Save Question 23 and upload.

(Total: 20 marks)

....

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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 shown at the end of the question.

Required Resources:

- **A computer**
- **SQL Server Management Studio**
- **Python program**
- **R Studio**
- **Orange Software**
- **Jupyter Notebook**
- **Ms Access 2016**

SECTION I (40 MARKS)

1. The ability of a data-mining algorithm to grow linearly in proportion to the dataset size while holding the available resources such as memory and the CPU constant is known as _________________. (2 marks)

2. Which layer in data warehouse architecture supports the storage of all data? (2 marks)

3. The comparison of general features of target class data object with the general features of an object from one or more sets of contrasting classes as used in data mining is called ________________. (2 marks)

4. In multidimensional data modeling, the visualisation technique that rotates data axes in view to provide an alternative presentation of the data is known as _________________. (2 marks)

5. The data mining technique used to examine the collection of items purchased by a customer in a single customer transaction to uncover the items purchased together is called ___________________. (2 marks)

6. The data transformation activity that involves dividing an attribute into several elements is referred to as __________________. (2 marks)

7. The routine data preprocessing technique the attempts to fill missing data values, smooth out noisy data and correct inconsistencies in data before loading it into a data warehouse is known as _________________. (2 marks)

8. A decision tree is a graphical representation of classification rules represented by nodes and edges. State the name given to the outgoing node without an outgoing edge ___________________. (2 marks)

9. Which data-mining tool enables data scientists to analyse complex data, discover patterns and build models so that they can easily detect fraud, anticipate resource demands and minimise customer attrition? (2 marks)

10. Write the python statement that will slice the numeric data into training and testing samples at a division point of 5. (2 marks)

11. The assortment of information that contains data around one or more business measurements in data warehouse design is known as __________________. (2 marks)

12. The technique used in data mining to mimic the workings of the human brain to recognise patterns and complex relationships in data to uncover hidden patterns is known as ________________. (2 marks)

13. Which data warehouse approach starts with data marts and then integrates them into a centralised data warehouse? (2 marks)

14. The process of using data mining techniques such as descriptive and predictive analytics to generate insights to support decision-making is referred to as ________________. (2 marks)

15. The type of cluster analysis where clusters are represented by a central entity which may or may not be part of the given data set is called _________________. (2 marks)

16. The technique used in data mining to transform numerical features on a similar scale is referred to as ___________________. (2 marks)

17. The data-preprocessing problem where the model memorises the noise in the data instead of learning meaningful patterns is known as _________________. (2 marks)

18. In a student data warehouse, the OLAP operations that allow users to start with a broad overview of student performance and gradually access more detailed data, ultimately leading to insights at the individual student level is known as ________________. (2 marks)

19. The process of deriving general principles from specific examples like developing predictive models from training data in machine learning is known as ________________. (2 marks)

20. The tools in OLAP that allow users to create custom queries without involving the IT department are 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 (e).

(a) Use the table below to create a Python CSV file called "customer_data.csv". (4 marks)

CustomerID	Age	AnnualIncome	SpendingScore
099	39	15000	30
100	41	35000	39
101	30	50000	81
102	23	20000	40
103	19	30000	30
104	21	40000	77
105	20	35000	40
106	23	20000	76
107	31	18000	35

(b) Import the Python libraries and load customer data from the CSV file named "customer_data.csv" into a pandas DataFrame called "customer_data". (3 marks)

(c) Write a Python code to explore the data as follows:

(i) To display the first five rows of the DataFrame to get a quick overview of the data. (2 marks)

(ii) To provide a summary of the DataFrame, including the data types of each column and the number of non-null values. (2 marks)

(d) Write a Python code to normalise the 'AnnualIncome' and 'SpendingScore' columns using Min-Max scaling. (4 marks)

(e) Write a Python code to perform K-Means clustering on the data to find the optimal number of clusters using the elbow method. (5 marks)

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

22. Create a word processing document named “Question 22” to save your answers to questions (a) to (f).

(a) Using a SQL database, create a new database named “EmployeeDB” using SQL statements. (2 marks)

(b) Create two tables called “DimEmployeee” and “FactSalary” respectively, given the metadata shown in the table below using relevant SQL Statements. (6 marks)

Tablename	**Column**	**Data Type**	**Description**
DimEmployee	EmployeeID	int	Primary Key
	Name	varchar(100)	Employee first name
	Department	varchar(100)	Department
FactSalary	SalaryID	int	Primary Key
	EmployeeID	int	Foreign Key
	Salary	Decimal(10,2)	Employee Salary
	DateEffective	date	Date of employment

(c) Write a SQL statement to populate the “DimEmployeee” with the following records. (3 marks)

EmployeeID	**Name**	**Department**
1	John Mocheche	Engineering
2	Nixon Muteti	Marketing
3	Jane Talo	Sales

(d) Write an SQL statement to populate the “FactSalary” with the following records. (3 marks)

SalaryID	**EmployeeID**	**Salary**	**DateEffective**
1	1	75000	2022-01-01
2	1	78000	2023-01-01
3	2	85000	2022-01-01
4	2	88000	2023-01-01
5	3	60000	2022-01-01
6	3	62000	2023-01-01

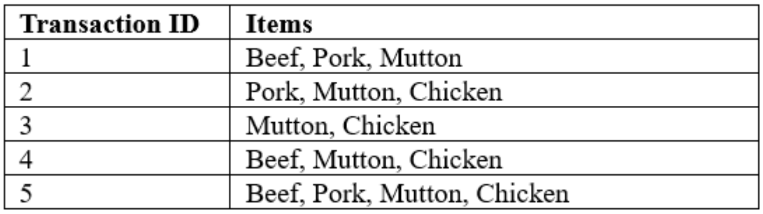
(e) Write the SQL statement to display all the salary data. (2 marks)

(f) Write an SQL statement query to find average salary per department. (4 marks)

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

23. Create a word processing document named “Question 23” to save your answers to questions (a) and (b).

Use the data in the table below to answer questions (a) and (b) below:



Transaction ID	**Items**
1	Beef, Pork, Mutton
2	Pork, Mutton, Chicken
3	Mutton, Chicken
4	Beef, Mutton, Chicken
5	Beef, Pork, Mutton, Chicken

(a) (i) Type the python code to import the necessary libraries and load the transaction data. (3 marks)

(ii) Write the python code to change the transaction data into a format suitable for the Apriori algorithm then apply the Apriori algorithm to find frequent item sets. (5 marks)

(iii) Write the python code to generate association rules from the frequent item sets. (4 marks)

(b) Give the python functions to perform the following tasks using the data frame called "**exams**":

(i) Create the exams data frame for the file "database.csv" (2marks)

(ii) Remove empty cells from the data frame. (2marks)

(iii) Return true for every row that is a duplicate, otherwise false. (2marks)

(iv) Remove duplicate rows from the data frame. (2marks)

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

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

……………………………………………………………………

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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

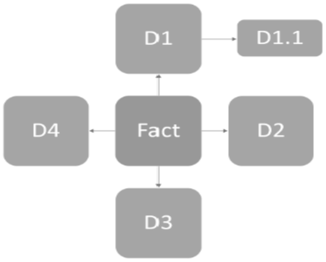
Answer ALL questions. This paper has two (2) sections. SECTION I has twenty (20) short response questions of two (2) marks each. SECTION II has three (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:

- **A computer**
- **SQL Server**
- **Python program**
- **R Studio**
- **Orange Software**

SECTION I (40 MARKS)

1. Identify the type of the data warehouse schema shown in the figure below: (2 marks)



2. Which type of data mining involves predicting a specific outcome based on historical data as used in data mining? (2 marks)

3. Which OLAP operation involves selecting two or more dimensions from the hypercube to create a new sub-cube for the given data? (2 marks)

4. The most common metrics used to evaluate the quality of clustering in data mining by quantifying how similar an object is to its own cluster compared to other clusters is known as ___________________. (2 marks)

5. The physical data warehouse that is installed and maintained on a company's server is referred to as ___________________. (2 marks)

6. Which data mining technique is commonly used in market segmentation to group customers with similar characteristics together? (2 marks)

7. During the process of building a data warehouse, the process of moving data from source systems into a data warehouse is called ___________________. (2 marks)

8. Which data mining platform supports multiple algorithms essential for machine learning, deep learning, text mining and predictive analytics? (2 marks)

9. The metric used to measure the proportion of true positive predictions among all positive predictions made by a classifier model is referred to as ___________________. (2 marks)

10. The storage system used in big data architectures to store raw, unstructured or semi-structured data in its original format, without the need for a predefined schema until it is processed for analysis is called ___________________. (2 marks)

11. An individual with multidisciplinary knowledge in computer science and statistics who asks unique and interesting questions of data based on formal or informal theory to generate rigorous and useful insights is known as ___________________. (2 marks)

12. You are tasked with designing a data warehousing solution for a large retail company that needs to integrate and analyse data from various sources, including sales transactions, inventory records and customer information. Your goal is to create an architecture that efficiently handles data staging, integration and access layer. What type of data warehouse architecture would you propose to meet these requirements? (2 marks)

13. Correlation is a technique used to measure the likelihood of two behaviours relating to each other. What type of correlation occurs when a value of one variable decreases with respect to another? (2 marks)

14. A pre-defined summary of data used to improve query performance in a data warehouse is called ___________________. (2 marks)

15. Analysing a large dataset to identify patterns and trends that can inform strategic decision-making for a retail company is essential for effective decision-making. Presenting the results of data mining analyses in a clear and intuitive manner for decision-making is called ___________________. (2 marks)

16. The process in data warehousing that involves periodically updating the data warehouse with new data is referred to as ___________________. (2 marks)

17. Which technique in data mining involves the use of algorithms to predict a target variable based on input variables? (2 marks)

18. The data mining technique that uses algorithms such as ID3, C4.5, and CART is called ___________________. (2 marks)

19. During the development process of a machine learning model, you are required to split the available dataset. One of the sets used to ensure the model's effectiveness and generalisation is known as___________________. (2 marks)

20. The interactive tools that allow users to explore data in more detail by navigating through different levels of granularity or dimensions in data warehouse presentation layers are 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 (c):

(a) Write the SQL statements to create and insert data in the "**Studentdata**" table below: (10 marks)

Studentname	Admission Year	Admission Month	StudentType	FeesPaid
Anita	2021	January	Full time	23000
Hellen	2021	January	Part time	37000
Anita	2021	January	Weekend	44000
Victor	2021	January	Distance learning	83000
Victor	2021	January	Distance learning	65000
Hellen	2020	September	Part time	77000
Hellen	2020	September	Part time	34000
Hellen	2020	September	Full time	53000
Hellen	2020	September	Part time	79000
Victor	2020	September	Full time	77000
Moses	2020	May	Distance learning	34000
Anita	2019	May	Distance learning	53000
Anita	2019	May	Distance learning	79000
Anita	2019	January	Distance learning	55000
Moses	2019	September	Weekend	78000
Moses	2019	May	Weekend	99000
Victor	2019	May	Part time	55000
Moses	2019	May	Part time	78000

(b) Write SQL statement that will find the average fees paid. Group the data by student name, admission year and student type. Display the results table. (5 marks)

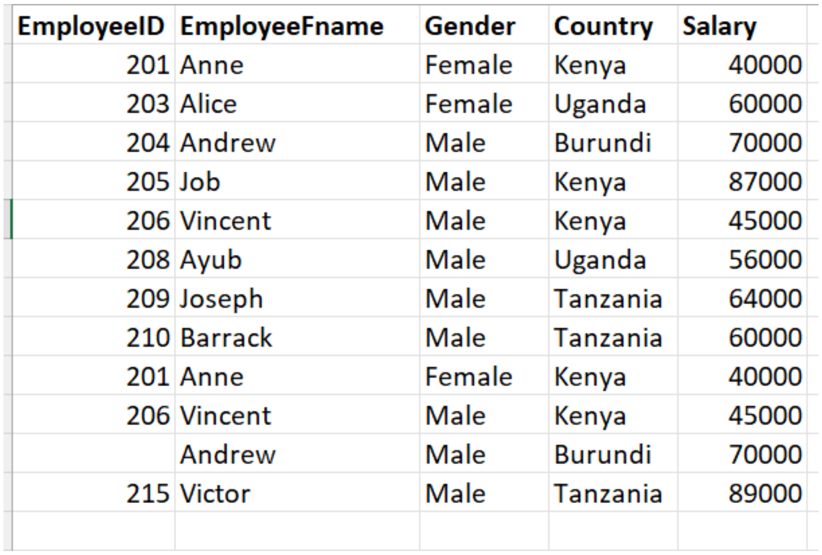
(c) Write SQL statement to use the cube function to find the sum of fees paid then group by admission month and student type. Display the results table. (5 marks)

Capture screenshots to show how you have performed the above task.
Save and Upload "Question 21".

(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 (b):

(a) Using a spreadsheet application, create the data set below and save it as "**Country**". (4 marks)



EmployeeID	EmployeeFname	Gender	Country	Salary
201	Anne	Female	Kenya	40000
203	Alice	Female	Uganda	60000
204	Andrew	Male	Burundi	70000
205	Job	Male	Kenya	87000
206	Vincent	Male	Kenya	45000
208	Ayub	Male	Uganda	56000
209	Joseph	Male	Tanzania	64000
210	Barrack	Male	Tanzania	60000
201	Anne	Female	Kenya	40000
206	Vincent	Male	Kenya	45000
	Andrew	Male	Burundi	70000
215	Victor	Male	Tanzania	89000

(b) Using the dataset in question (a) above, write a python code that will:

(i) Import the data set into a data frame and display the output after executing the code. (5 marks)

(ii) Return a data frame with no empty cells. (4 marks)

(iii) Remove duplicate entries from the data set. (4 marks)

(iv) Compute the total employee salary of the data set. (3 marks)

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

Save and Upload "Question 22" document.

(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 (b).

(a) Use the data below to produce a linear regression between the two variables length and width. (10 marks)

Width = 6,9,8,7,3,18,3,7,5,14,13,9,7

Length = 79,96,77,83,101,86,105,85,92,78,75,82,88

(b) The following array gives the ages of different staff in an organisation. Use the data to answer the following questions.

69,89,77,86,111,89,104,97,105,77,89,85,88,74,73

Required:
Write the python codes to perform the following:

(i) Find the mean. (2 marks)

(ii) Calculate the median. (3 marks)

(iii) Find the mode. (3 marks)

(iv) Count the number of occurrences of 77. (2 marks)

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

Save and Upload "Question 23" document.

(Total: 20 marks)

..

kasneb

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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:

- **A computer**
- **SQL Server**
- **Python program**
- **R Studio**
- **Orange Software**

SECTION I (40 MARKS)

1. _________ refers to the level of detail at which data is stored in a data warehouse ranging from fine-grained to coarse-grained. (2 marks)

2. The feature scaling method that subtracts the mean of each feature from its values and then divides by the range of the feature is referred to as? (2 marks)

3. A dimension reduction technique that performs aggregation on a data cube and makes the data less detailed is referred to as? (2 marks)

4. The name given to a technique used to analyse complex relational data, such as social networks, biological networks and transportation networks as used in technology trends in data mining is known as? (2 marks)

5. ___________ refers to the time delay between when data is generated or modified and when it becomes available for use in a data warehouse. (2 marks)

6. The type of data mart that combines data from the data warehouse and other operational sources and is best suited for multiple database environments is referred to as? (2 marks)

7. The data mining function which refers to the process of determining a model that describes and differentiates data classes with the intention of using it to predict a class of unknown objects is referred to as: (2 marks)

8. ________ analysis is used to group customers with similar characteristics or behaviours together for targeted marketing strategies. (2 marks)

9. The advanced analytics data mining tool intended to help users to quickly develop descriptive and predictive models through streamlined data mining processes is called? (2 marks)

10. The degree to which a human can understand and explain the reasoning behind the predictions or decisions made by a machine learning model or data mining algorithm is called? (2 marks)

11. An e-commerce platform operator wants to personalise product recommendations for each customer based on their browsing history and purchase behaviour. What type of data cube operation would be used to filter product recommendations by customer preferences? (2 marks)

12. State the SQL command operator to remove duplicates from a result set in SQL. (2 marks)

13. A schema is the logical description of data warehouse which is the blueprint for the entire data warehouse. How many fact tables are there in a Star schema? (2 marks)

14. The type of data warehousing schema, which consists of multiple fact tables and shared dimension tables, is referred to as_______________. (2 marks)

15. The architectural patterns for designing a data warehouse by identifying the main subject areas and entities the enterprise works with, such as customers, product and vendor is called? (2 marks)

16. The machine-learning paradigm, where an agent learns to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties is known as? (2 marks)

17. What is the primary modeling technique used in data warehousing? (2 marks)

18. An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification or the average prediction for regression is called? (2 marks)

19. ___________________ is the process of reducing the granularity of data for analysis purposes. (2 marks)

20. _________________ is the use of data mining algorithms for both signature-based and anomaly-based detection. (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 (e).

(a) Use the table below to create a CSV file and save it as “sales_data.csv”. (3 marks)

Date	**VehicleModel**	**SalesAmount**	**PurchaseFrequency**	**DiscountApplied**
2023-01-01	Toyota Camry	2,500,000	2	0.1
2023-01-05	Toyota Corolla	2,000,000	1	0.05
2023-01-10	Honda Civic	2,200,000	3	0.2
2023-02-02	Toyota Camry	2,600,000	2	0.1
2023-02-05	Ford Mustang	3,500,000	1	0.15
2023-02-15	Toyota Corolla	2,100,000	1	0.05
2023-03-03	Honda Civic	2,300,000	2	0.1
2023-03-08	Toyota Camry	2,700,000	3	0.2
2023-03-20	Ford Mustang	3,600,000	1	0.15

(b) Import the Python libraries and load sales data from the CSV file named “sales_data.csv” into a pandas DataFrame called “sales_data”. (2 marks)

(c) Write a Python code that will:

(i) Display the first few rows of the DataFrame to get a quick overview of the data. (1 mark)

(ii) Provide a summary of the DataFrame, including the data types of each column and the number of non-null values. (1 mark)

(d) Write a Python code to do the following preprocessing tasks:

(i) Remove rows with missing values from the DataFrame. (2 marks)

(ii) Remove duplicate rows from the DataFrame. (2 marks)

(iii) Convert the 'Date' column to datetime format. (3 marks)

(e) Write a Python code to do the following operations:

(i) Group the sales data by year and month and then calculates the total sales amount for each group. (2 marks)

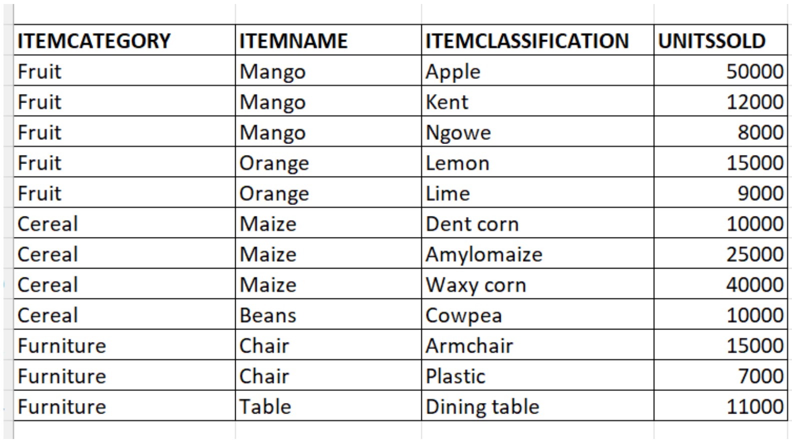
(ii) Segment customers based on their purchase frequency into three categories: one-time, regular and frequent buyers. (2 marks)

(iii) Count the occurrences of each vehicle model in the DataFrame, providing insights into popular models. (2 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 an SQL statements to create and insert data in the table below. (10 marks)



ITEMCATEGORY	ITEMNAME	ITEMCLASSIFICATION	UNITSSOLD
Fruit	Mango	Apple	50000
Fruit	Mango	Kent	12000
Fruit	Mango	Ngowe	8000
Fruit	Orange	Lemon	15000
Fruit	Orange	Lime	9000
Cereal	Maize	Dent corn	10000
Cereal	Maize	Amylomaize	25000
Cereal	Maize	Waxy corn	40000
Cereal	Beans	Cowpea	10000
Furniture	Chair	Armchair	15000
Furniture	Chair	Plastic	7000
Furniture	Table	Dining table	11000

(b) Using SQL rollup function, find the sum of item units sold by item category, item name and item classification.

Display the results table. (6 marks)

(c) Using SQL pivot operator, write an SQL statement that will pivot the total number of maize, beans and orange.

Display the results table. (4 marks)

Save "Question 22" 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 (g).

(a) Use the table below to create a CSV file and save it as “mobishop_data.csv”. (3 marks)

Product	Quantity	Revenue
Laptop	10	500000
Mouse	50	25000
Keyboard	30	90000
Monitor	20	300000
Laptop	5	250000
Keyboard	10	30000
Mouse	20	10000
Headphones	15	75000
Monitor	10	150000

(b) Import pandas libraries for handling data frames and mysql.connector for interacting with MySQL databases. (2 marks)

(c) Load data from a CSV file named “sales_data.csv” into a pandas DataFrame named “sales_data”. (2 marks)

(d) Establish a connection to MySQL database using the necessary credentials for the host, user, password and the database. (3 marks)

(e) Write SQL statements to write the DataFrame to MySQL table. (4 marks)

(f) Write SQL statements to clean the data by handling missing values. (2 marks)

(g) Write the SQL statement to calculate the top selling product. (4 marks)

Save “Question 23” and upload. **(Total: 20 marks)**

....

kasneb

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

MONDAY: 4 December 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:

- **A computer**
- **Python program**
- **R Studio**

SECTION I (40 MARKS)

1. A healthcare provider is seeking ways to enhance patient care and optimise hospital operations. You are a data scientist tasked with using data mining techniques to uncover valuable insights. Which data mining technique would you recommend to extract insights from unstructured medical notes, research papers, and patient feedback to improve treatments? (2 marks)

2. The procedure for converting and enhancing data in the staging layer to satisfy the needs of the data warehouse is known as: (2 marks)

3. OLAP can be used to analyse sales data by structuring it into a multidimensional cube. State the OLAP operation that would be used to change the orientation of the data cube to view sales from different perspectives, like switching from time-focused analysis to product-focused analysis. (2 marks)

4. Which type of machine learning algorithms, inspired by the human brain, is utilised for complicated data processing tasks including pattern recognition, picture and speech analysis, and others? (2 marks)

5. The data loading strategy where only new or changed data is added to the existing data in the warehouse is called: (2 marks)

6. What is the name of the statistical method employed in data mining to model the relationship between a dependent variable and one or more independent variables to be used for prediction? (2 marks)

7. The type of data mart that supports the combination of input from sources other than a data warehouse is known as: (2 marks)

8. The data mining algorithm that asserts "all of an item set's subsets are also likely to be frequent if an itemset is frequent" is known as: (2 marks)

9. The policies, procedures, and standards for managing and using data in an organisation is collectively known as: (2 marks)

10. The data mining technique that involves extracting patterns and insights from geographic and location-based data and is used in urban planning and environmental analysis is called: (2 marks)

11. The design and functionality of the graphical or command-line interfaces used in data mining tools and software is referred to as ___________________________ issues. (2 marks)

12. The process of intentionally introducing redundancy into a data warehouse schema to improve query performance and simplify data retrieval is called: (2 marks)

13. The name given to the type of records information about the day-to-day operations of the system, such as data loading schedules, errors, and system performance in a data warehouse is called: (2 marks)

14. When designing and implementing a data warehouse for the retail chain, several critical considerations should be addressed. State the name given to the consideration for the data warehouse having the ability to accommodate growth and changing data volumes. (2 marks)

15. The process of moving historical or infrequently accessed data to long-term storage to improve query performance and reduce storage costs in data warehouse is referred to as: (2 marks)

16. State the name given to the data warehouse characteristic which means that once data is loaded into the warehouse, it is not updated in place but instead a new data is appended. (2 marks)

17. The technique used to reduce storage space and improve query performance in a data warehouse is called: (2 marks)

18. The best name given to the data mining issue which relates to inability to work with parallel, distributed and incremental algorithms as well as inefficiency and difficulty in scalability is: (2 marks)

19. What defines the structure and relationships of the data in the warehouse, helping to organise and optimise data storage? (2 marks)

20. The type of data mining technique that is used to identify and analyse the relationship between variables 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).

(a) Using appropriate spreadsheet software, use the data provided below to create a dataset called "store.csv" about customers and save it. (5 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	Marries	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) Using Python panda library, load the dataset created in Question 21 (a) into a data frame object called "data".

(3

(c) Display the first row of the data frame in Python. (2 marks)

(d) Write a Python code to calculate the average income for each education level. (3 marks)

(e) Write a Python code to calculate and display the difference between the maximum and minimum values for the attribute "income". (3 marks)

(f) Write a Python code to calculate the variance and standard deviation of the income. (4 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 (f).

(a) Write SQL statements to create a database called "Data warehouse" and open the database for use.(2 marks)

(b) Use SQL statements to create a table called "Customers" to store the dataset shown below. (4 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

(c) Write SQL statement for extracting the first three records from the table created in Question 22 (b). (2 marks)

(d) Write the SQL statements to load a single record into the table customers (4 marks)

(e) Write an SQL statements to calculate the average income by 'MaritalStatus' in the dataset to illustrate aggregation as used in warehousing. (4 marks)

(f) Write SQL statements to calculate summary statistics for the "Income" column, including mean, median, and standard deviation. (4 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).

The logical design for a data warehouse is represented in the following fact sales table and dimension tables for time and products.

fact _sales (sales_key, date_key product_key ,sales_amount)

dim_date (date_key, date, day,month,year)

dim_product(product_key,product_name,category,price)

(a) Use SQL statements to create the above three tables specifying the primary key for each table. (9 marks)

(b) Create a basic OLAP cube called "SalesCube" using SQL statements. Use the tables defined in question 23 (a). (4 marks)

(c) Perform a "slice" operation in SQL to view data from the SalesCube to display the sales for the month of July. (4 marks)

(d) Write a SQL statement to display all the records from the time dimension for the month of April. (3 marks)

Save and upload Question 23.

(Total: 20 marks)

……………………………………………………………………

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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:

- **A computer**
- **Python program**
- **R Studio**

SECTION I (40 MARKS)

1. The step in data warehousing that involves identifying where the critical information is and how to move it into the data warehouse structure is referred to as? (2 marks)

2. The data warehousing tools that support reporting and querying, application development and data mining are known as? (2 marks)

3. The data mining activity that involves the mining of data from mobile devices to get information about individuals is known as? (2 marks)

4. The data warehouse component that speeds up response and processing times, delivers data to users in easily digestible formats and stores profiles is called? (2 marks)

5. The component of a data warehouse architecture, which processes the data and organises it into tables, is known as? (2 marks)

6. Which term best describes an online transactional system that supports transaction-oriented applications in a 3-tier architecture? (2 marks)

7. The type of OLAP that stores data in the form of rows and columns and is capable of saving storage space while working with massive historical datasets that are not often queried is referred to as? (2 marks)

8. The type of OLAP operation that performs aggregation on a data cube and makes the data less detailed is called? (2 marks)

9. A regression technique that occurs when the regression line slopes upward with the lower end of the line at the y intercept of the graph and the upper end of line extending upward into the field, away from the x intercept is known as? (2 marks)

10. Which term best describes a series of activities that are essential to create a fully functioning data warehouse after classifying, analysing and designing the data warehouse with respect to the requirements provided by the client? (2 marks)

11. What is the name of the loading method where new data items are added from the operation data store to the data warehouse? (2 marks)

12. The type of data warehousing schema, which consists of multiple fact tables and shared dimension tables, is referred to as? (2 marks)

13. Which term best describes the data that provides information about one or more aspects of the data? (2 marks)

14. Which is the type of data mart that gets data directly from an operational source or external source? (2 marks)

15. The big data mining technique that is used to identify critical abnormalities in data that could be indicative of a deeper issue is known as? (2 marks)

16. The type of cluster analysis where clusters are represented by a central entity, which may or may not be part of the given data set, is called? (2 marks)

17. The knowledge discovery activity in which clever techniques are applied to extract patterns of data, which are useful, is referred to as? (2 marks)

18. Which term best describes the data warehousing architecture that incorporates real time and derived data? (2 marks)

19. The component of a data warehouse that analysts use to pull out insights from their data stored in the data warehouse is known as? (2 marks)

20. The data warehousing implementation technique that allows the customer to visualise the design of the architecture they have requested before the deployment takes place is referred to as? (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 (c).

(a) Describe **THREE** types of data marts. (6 marks)

(b) XYZ Company with branches in Kenya, Uganda and Rwanda wants to integrate its data.

Advise the company on the **SEVEN** steps that it will require to take in implementing a data warehouse. (7 marks)

(c) Given the data set X= 2, 5.5, 1.55, 6.1, 5.75, 6.8, 4.7, write the python codes to find the following:

(i) Mode. (2 marks)

(ii) Variance. (2 marks)

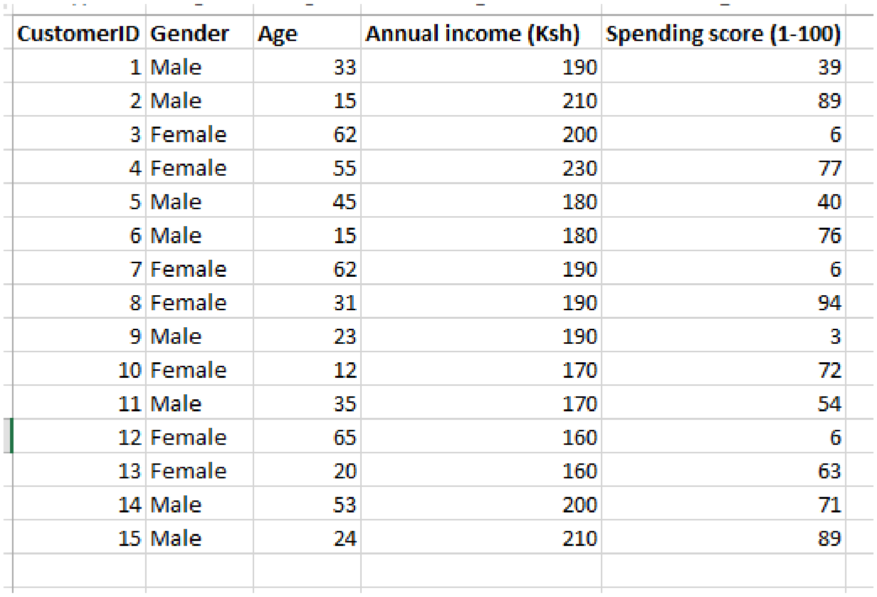
(iii) Standard deviation. (3 marks)

Save "Question 21" document and upload.

(Total: 20 marks)

22. Create a word processing document named “Question 22”. Use the word processing document to save your answers to questions (a) to (d).

(a) Create a folder on the desktop called “DDMA”. Create an excel document shown below, save it as a comma separated version (CSV) file named “**Customer**”. (4 marks)



CustomerID	Gender	Age	Annual income (Ksh)	Spending score (1-100)
1	Male	33	190	39
2	Male	15	210	89
3	Female	62	200	6
4	Female	55	230	77
5	Male	45	180	40
6	Male	15	180	76
7	Female	62	190	6
8	Female	31	190	94
9	Male	23	190	3
10	Female	12	170	72
11	Male	35	170	54
12	Female	65	160	6
13	Female	20	160	63
14	Male	53	200	71
15	Male	24	210	89

(b) Write a python code to import the necessary python libraries. (4 marks)

(c) Write python programming code that will retrieve data from the .csv file. (4 marks)

(d) Consider the two vectors below:

A1 <- c(5,7,8,7,2,2,9,4,11,12,9,6)

B1 <- c(99,86,87,88,76,74,87,94,78,77,85,86)

Required:
Write R Studio code that will display a scatter plot with a title “Student details”. Label the x-axis as “Plant age (weeks)” and y-axis as “Plant height”.

(8 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 (c).

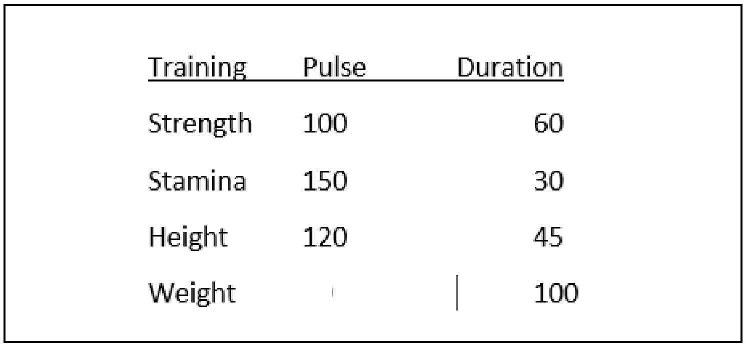
(a) Write R-studio code that will perform the following tasks:

(i) Creating a bar graph. (2 marks)

(ii) Producing a stem plot. (2 marks)

(iii) Displaying memory content. (2 marks)

(b) Using R Studio, create a data frame for the data shown below and use an appropriate function to summarise the values. (8 marks)



Training	Pulse	Duration
Strength	100	60
Stamina	150	30
Height	120	45
Weight		100

(c) Describe the following data mining techniques:

(i) Characterisation. (2 marks)

(ii) Discrimination. (2 marks)

(iii) Association analysis. (2 marks)

Save “Question 23” document and upload.

(Total: 20 marks)

....

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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:

- **A computer**
- **Python program**
- **R Studio**
- **Orange software**

SECTION I (40 MARKS)

1. The pre-calculated summarised data, from raw data that is used to improve query performance and reduce data storage in data warehouse is called? (2 marks)

2. The data warehouse component that performs all the operations associated with the management of data in the warehouse is referred to as? (2 marks)

3. The real-world datasets contain missing or incomplete data, which can impact the quality and accuracy of analysis and modeling. What is the name of the technique used to fill in missing or incomplete values in a dataset? (2 marks)

4. What name is given to a metric used to measure the proportion of positive predictions that are actually correct, given by the number of true positive predictions divided by the number of true positive and false positive predictions as used in data mining? (2 marks)

5. In data warehousing, the activity name that supports incremental loading of new data on a periodic basis within narrow time windows is? (2 marks)

6. A dimensional data model that has a fact table in the center, surrounded by normalised dimension tables is referred to as? (2 marks)

7. Which term best describes the analysis, dynamic synthesis and consolidation of large volumes of multidimensional data? (2 marks)

8. The database term that best describes the subset of a data warehouse that supports the requirements of a section of an organisation or business function is? (2 marks)

9. State the name given to the fact table in a data warehouse that does not contain any measures or facts but only the keys of the dimensions that participate in the fact table? (2 marks)

10. In data warehousing, which term best describes the strategy that represents an oversight for all domain specific strategies including business intelligence, big data and data management? (2 marks)

11. The data mining function that involves identifying data points that are unusual or deviate significantly from the norm, such as fraud detection or network intrusion detection is called? (2 marks)

12. The data mining functionality that is useful for discovering interesting relationships hidden in large data sets is referred to as? (2 marks)

13. Name two performance issues related to data mining. (2 marks)

14. The process that takes the raw results from data mining and carefully and accurately transforms them into useful and understandable information is referred to as? (2 marks)

15. Which is the term that best describes the data mining task which supports the finding of correlations between items in a database. (2 marks)

16. Name **TWO** data items that are removed during the data cleaning phase of the data mining process. (2 marks)

17. The term that best describes the series of activities that are essential to create a fully functioning Data Warehouse, after classifying, analysing and designing the Data Warehouse with respect to the requirements provided by the client is referred to as? (2 marks)

18. The data objects with characteristics that are considerably different from the other data objects in the data set are referred to as? (2 marks)

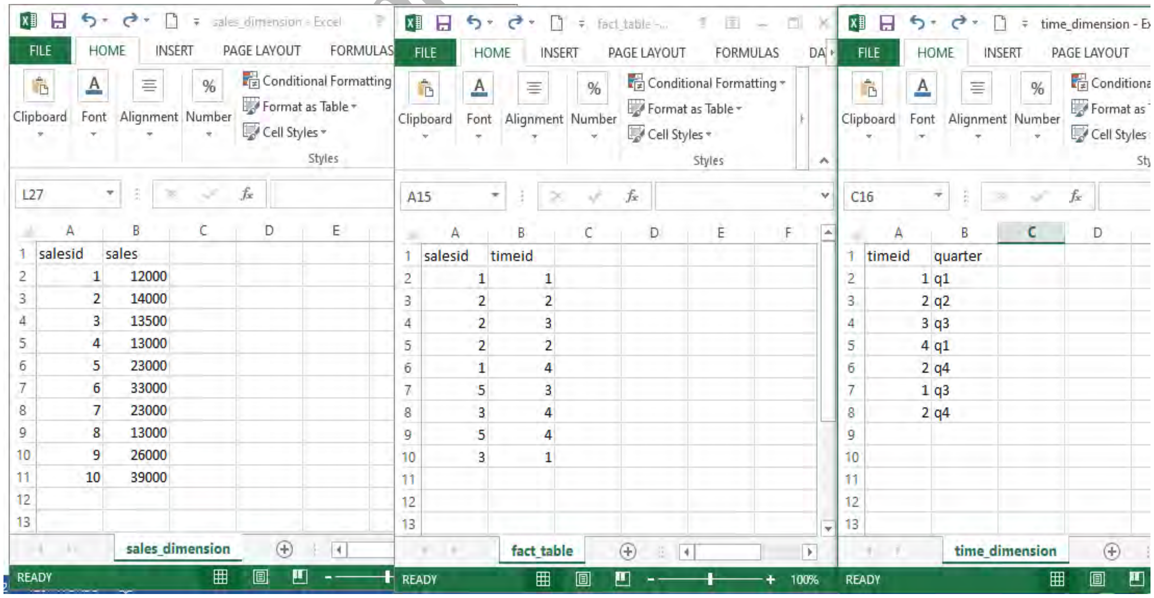
19. An important step in the data pre-processing phase that deals with identifying the most important variables in a dataset for data mining is known as? (2 marks)

20. Data warehousing typology where end users have direct access to the data stores using tools enabled at the data access layer 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 (e).

(a) Create a folder on the desktop called "DDMA". Using excel, create the three separate files shown below, saving them as a comma separated version (CSV) files named "sales_dimension.csv", "fact_table.csv" and "time_dimesion.csv" respectively: (6 marks)



salesid	sales
1	12000
2	14000
3	13500
4	13000
5	23000
6	33000
7	23000
8	13000
9	26000
10	39000

salesid	timeid
1	1
2	2
2	3
2	2
1	4
5	3
3	4
5	4
3	1

timeid	quarter
1	q1
2	q2
3	q3
4	q1
2	q4
1	q3
2	q4

(b) Use the relevant library and load the "sales_dimension", "fact_table" and "time_dimension" tables into pandas dataframes. (4 marks)

(c) Join the "sales_dimension" and "fact_table" tables using the "salesid" applying relevant function and display the results of the merged table. (3 marks)

(d) Join the merged table in question 21(c) above and the “time_dimension” table using “timeid” applying relevant function and display the results of the final merged table. (3 marks)

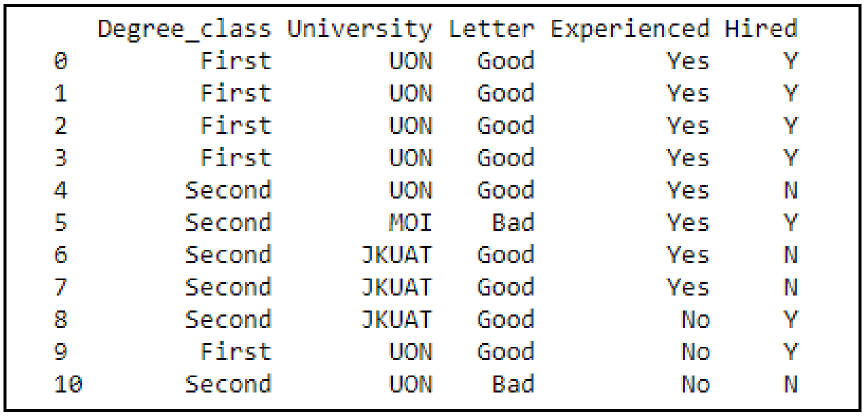
(e) Group the final table by quarter and calculate the sum of sales applying relevant python and display the results of quarterly sales. (4 marks)

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 (h).

(a) Type the following dataset into Excel worksheet. Create a folder in drive C: and name it **Analytics**. Save the Excel worksheet in **Analytics** folder as a comma separated version (CSV) file and name it ***Colleges.csv***. (2 marks)



```
   Degree_class University Letter Experienced Hired
0         First        UON   Good         Yes     Y
1         First        UON   Good         Yes     Y
2         First        UON   Good         Yes     Y
3         First        UON   Good         Yes     Y
4        Second        UON   Good         Yes     N
5        Second        MOI    Bad         Yes     Y
6        Second      JKUAT   Good         Yes     N
7        Second      JKUAT   Good         Yes     N
8        Second      JKUAT   Good          No     Y
9         First        UON   Good          No     Y
10       Second        UON    Bad          No     N
```

(b) Write the python code to import the necessary python libraries. (2 marks)

(c) Write the python code to extract data from the source and print the output. (2 marks)

(d) Write the python code to transform the data into a numeric array and print the output. (3 marks)

(e) Write the python code to separate independent variables from dependent variable and print the output. (2 marks)

(f) Write the python code to normalise the data and print the output. (3 marks)

(g) Apply K-means algorithm to create clusters (3 marks)

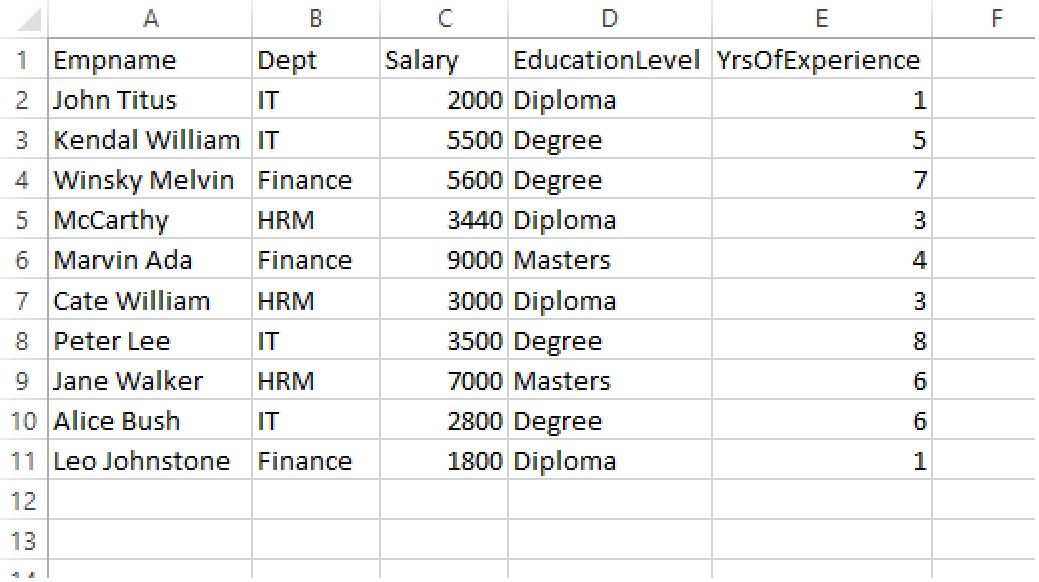
(h) Write the python code to visualise the generated clusters. (3 marks)

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 (g). You are required to use either Python/R programming languages or Orange studio to answer the questions.

(a) Create a CSV dataset shown below and save it in DDMA folder as "Workers.csv". (4 marks)



	A	B	C	D	E	F
1	Empname	Dept	Salary	EducationLevel	YrsOfExperience	
2	John Titus	IT	2000	Diploma	1	
3	Kendal William	IT	5500	Degree	5	
4	Winsky Melvin	Finance	5600	Degree	7	
5	McCarthy	HRM	3440	Diploma	3	
6	Marvin Ada	Finance	9000	Masters	4	
7	Cate William	HRM	3000	Diploma	3	
8	Peter Lee	IT	3500	Degree	8	
9	Jane Walker	HRM	7000	Masters	6	
10	Alice Bush	IT	2800	Degree	6	
11	Leo Johnstone	Finance	1800	Diploma	1	
12						
13						

(b) Load the dataset "workers.csv" using the relevant commands. (2 marks)

(c) Use a relevant function to clean the dataset by the drop missing values. (2 marks)

(d) Get the statistics about the dataset "workers.csv" and display the results of the console screen. (2 marks)

(e) Use a relevant function to filter and display the data for all employees earning a salary greater or equal to 5000 dollars. (3 marks)

(f) Calculate and display the total salary paid to the staff by department and education level. (4 marks)

(g) Create and display a bar plot of salary by departments with a well labelled graph showing the title, x and y axis. (3 marks)

Save "Question 23" and upload.

(Total: 20 marks)

....

kasneb

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

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 (3) practical questions of sixty (60) marks. Marks allocated to each question are shown at the end of the question.

Required Resources:

- **A computer**
- **Python program**
- **R Studio**

SECTION I (40 MARKS)

1. The process for finding hidden patterns and relationships in a set of data to unearth previously unknown insights is called? (2 marks)

2. What is the name given to the open standard process model, used in data mining to describe the steps in a data mining process and is not specific to any particular industry or tool. (2 marks)

3. The component of the data warehouse that ensures reliable extraction and loading of data in the database is referred to as? (2 marks)

4. To support quick access to various business applications in data mining, a multidimensional cube is used. Name the type of analytical OLAP multidimensional operation that performs aggregations on the data either by moving up the dimensional hierarchy or by dimensional reduction. (2 marks)

5. Which data warehouse characteristics is described by the statement “data is not updated in real time but is refreshed from operational systems on a regular basis”. (2 marks)

6. The data warehouse component, whose purpose of is to show the pathway back to where the data began, so that the warehouse administrators know the history of any item in the warehouse is called? (2 marks)

7. A machine learning approach characterized by the imitation of human brains using a multilayer perceptron is called? (2 marks)

8. What name is given to the dimensional data model design schema that has a fact table at the center with normalised dimension tables connected to it? (2 marks)

9. Write python programming function to display the number of rows and columns in a data frame. (2 marks)

10. Which tier of the three-tier architecture tools is used for high-level data analysis, querying, reporting, and data mining? (2 marks)

11. What name is given to the knowledge discovery in database process stage, where data is consolidated into forms appropriate for data mining, by performing summary and aggregation operations. (2 marks)

12. The conversion of continuous data to discrete data, such as the age of between 1 to 2 years into “toddlers” or years 10 to 12 years into “teenage” is called? (2 marks)

13. The store for performance measurements as a result of business process events in a data warehouse is called? (2 marks)

14. The data mining algorithm that classifies and categorizes a data point by calculating the distances between the data point and other data points in the training data set and assign the data point to the class using proximity is known as? (2 marks)

15. The Python library used to support arrays and matrices operations is called? (2 marks)

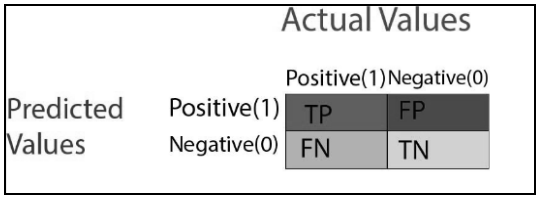
16. What is the name given to process of extracting essential data from the standard language text? (2 marks)

17. Pruning is the technique of removing the unused branches from the decision tree as decision tree might represent noisy or outliers. What do you call the pruning approach where the construction of the decision tree is stopped early? (2 marks)

18. What do we call the facts that cannot be summed up for any of the dimensions present in the fact table as used in data warehouse logical design? (2 marks)

19. Give the list of data for ages of five people as: 23,56,55,34,25, and their weights: 34,78,72,65,34. State two vectors for ages and weights using R programming. (2 marks)

20. Identify the data mining evaluation tool below used to measure the performance of a data mining model: (2 marks)

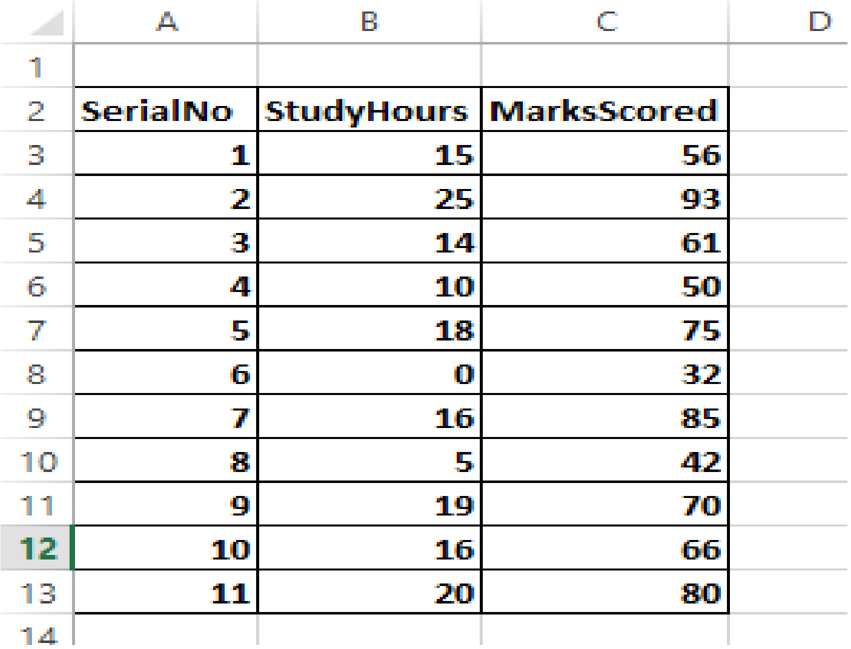


		Actual Values	
		Positive(1)	Negative(0)
Predicted Values	Positive(1)	TP	FP
	Negative(0)	FN	TN

SECTION II (60 MARKS)

21. Create a word processing document named “Marks analysis” and use the word processor document to save your answers to questions (a) to (b).

(a) Create a folder on the desktop named “DDMA”. In the folder you have created, create an excel worksheet shown below and save it as a comma separated version (CSV) file named “**Studentmarks**”. (3 marks)



	A	B	C	D
1				
2	SerialNo	StudyHours	MarksScored	
3	1	15	56	
4	2	25	93	
5	3	14	61	
6	4	10	50	
7	5	18	75	
8	6	0	32	
9	7	16	85	
10	8	5	42	
11	9	19	70	
12	10	16	66	
13	11	20	80	
14				

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

(b) Write python programming codes that will perform the following tasks:

(i) Retrieve data from the “**Studentmarks**” file and print a summary of its description. (4 marks)

(ii) Split the data into training and testing sets. (4 marks)

(iii) Split each set into input and output attributes data. (4 marks)

(iv) Display the relationship between the attributes in training and test data sets using a scatter graph. (4 marks)

(v) Build the linear regression object and train it using the training set. (4 marks)

(vi) Create a Linear Regression Object of the two variables. (2 marks)

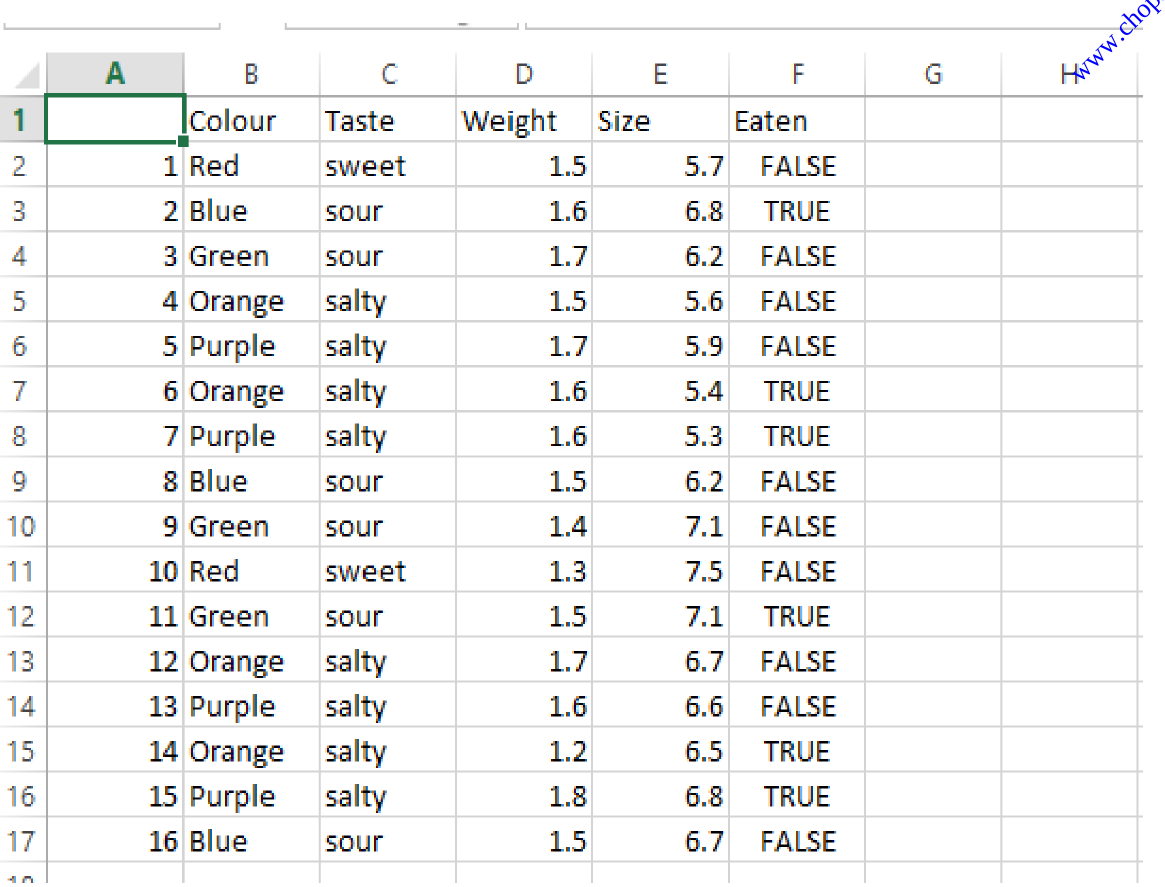
Capture and save screenshots to demonstrate how you have performed the above task.

Upload “Marks analysis” document.

(Total: 25 Marks)

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

(a) Create an excel worksheet shown below and save it in “DDMA” folder on the desktop. Save the file you have created as a comma separated version (CSV) file named “COLORTASTE”. (3 marks)



	A	B	C	D	E	F	G	H
1		Colour	Taste	Weight	Size	Eaten		
2	1	Red	sweet	1.5	5.7	FALSE		
3	2	Blue	sour	1.6	6.8	TRUE		
4	3	Green	sour	1.7	6.2	FALSE		
5	4	Orange	salty	1.5	5.6	FALSE		
6	5	Purple	salty	1.7	5.9	FALSE		
7	6	Orange	salty	1.6	5.4	TRUE		
8	7	Purple	salty	1.6	5.3	TRUE		
9	8	Blue	sour	1.5	6.2	FALSE		
10	9	Green	sour	1.4	7.1	FALSE		
11	10	Red	sweet	1.3	7.5	FALSE		
12	11	Green	sour	1.5	7.1	TRUE		
13	12	Orange	salty	1.7	6.7	FALSE		
14	13	Purple	salty	1.6	6.6	FALSE		
15	14	Orange	salty	1.2	6.5	TRUE		
16	15	Purple	salty	1.8	6.8	TRUE		
17	16	Blue	sour	1.5	6.7	FALSE		

(b) Import the data set into R-Studio. (4 marks)

(c) Write the R-studio codes that will perform the following tasks:

(i) Display a scatterplot representing “size” compared to “weight”. (5 marks)

(ii) Return the first 6 rows. (4 marks)

(iii) Display a histogram of the frequency of each weight. (4 marks)

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

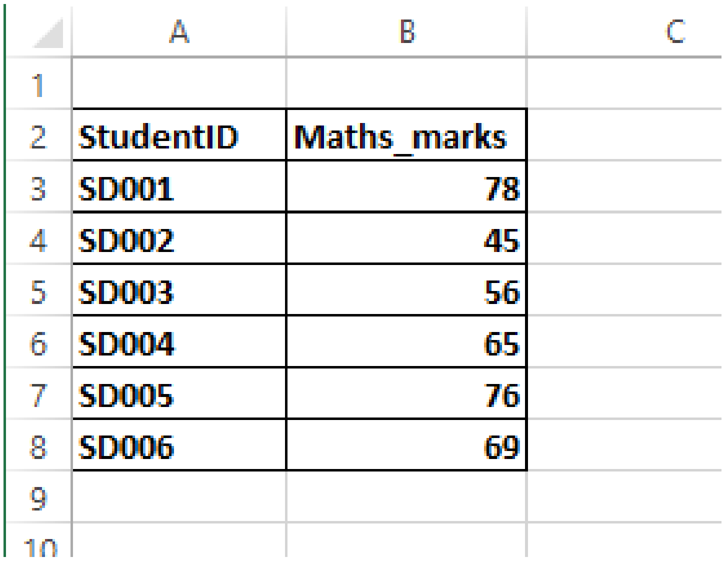
Upload “Colour” document.

(Total: 20 Marks)

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

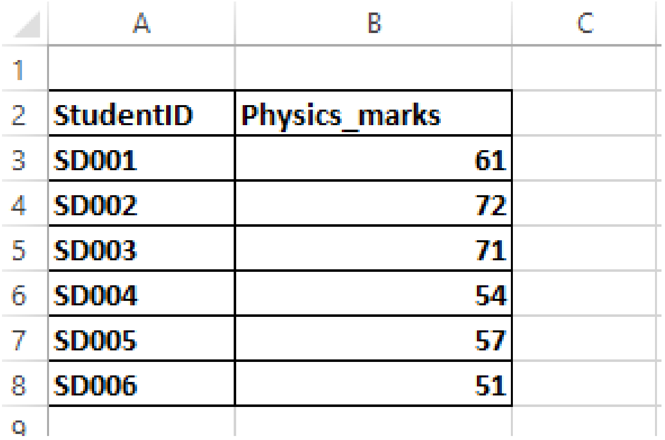
(a) Create excel document shown below and save it in “DDMA” folder on the desktop. Save the excel worksheets as a comma separated version (CSV) file named **physicsmarks** and **mathematicsmarks** respectively. (3 marks)

mathematicsmarks



	A	B	C
1			
2	StudentID	Maths_marks	
3	SD001	78	
4	SD002	45	
5	SD003	56	
6	SD004	65	
7	SD005	76	
8	SD006	69	
9			
10			

physicsmarks



	A	B	C
1			
2	StudentID	Physics_marks	
3	SD001	61	
4	SD002	72	
5	SD003	71	
6	SD004	54	
7	SD005	57	
8	SD006	51	
9			

(b) Write python programming codes that will load the two data sets on question (a) above, convert both data sets into data frames and merge them. (6 marks)

(c) The speed of 15 cars is given by Speed = [99,86,87,88,111,86,103,87,94,78,77,85,86,85,97]. Write the python code to find the following:

(i) The median

(ii) The mean

(iii) The mode (6 marks)

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

Upload Marks document.

Total: 15 Marks)

……………………………………………………………………………

kasneb

DIPLOMA IN DATA MANAGEMENT AND ANALYTICS (DDMA)

LEVEL II

WAREHOUSING AND DATA MINING

MONDAY: 1 August 2022. Afternoon paper. **Time Allowed: 3 hours.**

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

SECTION I

1. You have been provided with a dataset that is not labelled, to discover hidden patterns in data using machine learning. Which tasks of machine learning is the most suitable for this kind of a task? (2 marks)

2. Name the component of the star schema that is used to store the measurements and the unique identifier as used in logical design of the data warehouse (2 marks)

3. A ________________ is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test and each leaf node holds a class label. (2 marks)

4. A ________________ schema, is a multidimensional representation which contains two facts table that shares dimensions between them. (2 marks)

5. The process of grouping related values together to reduce the number of distinct values for an attribute in data mining is referred to as ________________________. (2 marks)

6. _____________ is the process of the staging layer of the data warehouse architecture that does the aggregation, summarisation and change the format, structure and sift through data. (2 marks)

7. The data warehouse characteristic where data is maintained via different intervals of time such as weekly, monthly, or annually is known as ______________________. (2 marks)

8. Write a python programming code to slice a list called “mynumbers”, to display all the elements except the first element. (2 marks)

9. The process of discovering potentially useful, interesting, and previously unknown patterns from a large collection of data useful for business decision making is known as ___________________________. (2 marks)

10. The number of times the algorithm sees the entire data set in during the training process is called _______________. (2 marks)

11. The Online Analytical Operation (OLAP) that perform aggregation on a data cube by climbing up a concept hierarchy for a dimension is called ________________. (2 marks)

12. The data classification method that classifies a point by calculating the distances between the point and points in the training data set by assigning the point to the class that is most common is referred to as _____________________.

13. The subject-oriented data repository used for the analytical purposes of the specific group is called a ____________ (2 marks)

14. The data warehouse control panel that describes the data and analytics system is known as ____________. (2 marks)

15. In machine learning, after the development of a classifier model, you need to measure the performance of the classifier. Name the technique which generates a tabular summary of the number of correct and incorrect predictions made by a classifier. (2 marks)

16. ________________ is the data warehouse operation that selects two or more dimensions from a given cube and provides a new sub-cube. (2 marks)

17. ____________ algorithm is a probabilistic classifier that uses probability of a prediction from the underlying evidence as used in data mining techniques. (2 marks)

18. A _____________ is a structure that categorises data in order to enable users to answer business questions such as sales, inventory or marketing as used in data warehouse schema design. (2 marks)

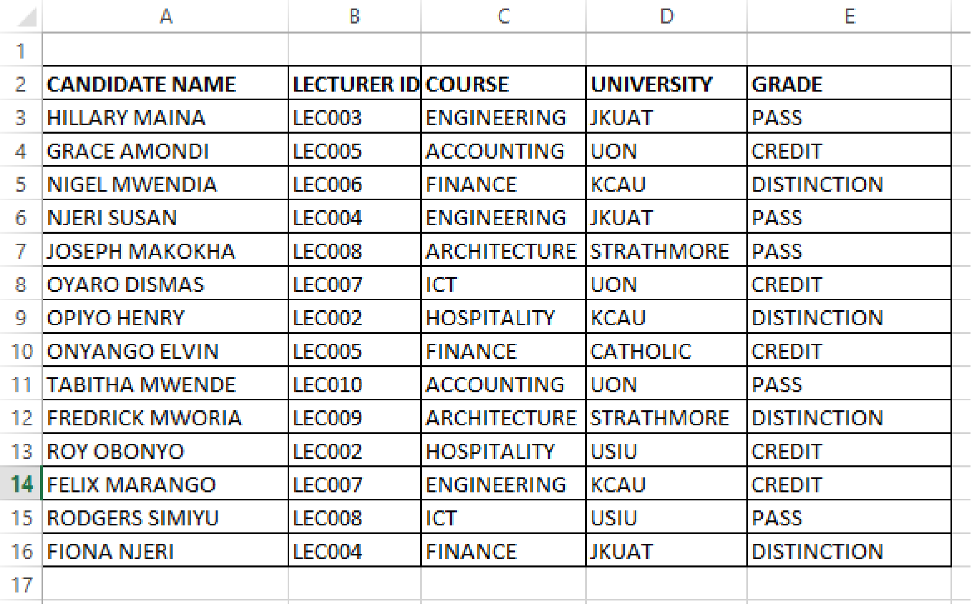
19. An ______________________ is an observation or a data point that lies an abnormal distance from other values in a random sample from a population. (2 marks)

20. A ________________ schema is a database organisational structure optimised for use in a data warehouse or business intelligence that uses a single large fact table to store transactional or measured data and one or more smaller dimensional tables that store attributes about the data. (2 marks)

SECTION II

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

(a) Create a folder on drive C: and call it "ICT". In "ICT" folder, create an excel document shown below, save it as a comma separated version (CSV) file and name it "Grades". (3 marks)



	A	B	C	D	E
1					
2	CANDIDATE NAME	LECTURER ID	COURSE	UNIVERSITY	GRADE
3	HILLARY MAINA	LEC003	ENGINEERING	JKUAT	PASS
4	GRACE AMONDI	LEC005	ACCOUNTING	UON	CREDIT
5	NIGEL MWENDIA	LEC006	FINANCE	KCAU	DISTINCTION
6	NJERI SUSAN	LEC004	ENGINEERING	JKUAT	PASS
7	JOSEPH MAKOKHA	LEC008	ARCHITECTURE	STRATHMORE	PASS
8	OYARO DISMAS	LEC007	ICT	UON	CREDIT
9	OPIYO HENRY	LEC002	HOSPITALITY	KCAU	DISTINCTION
10	ONYANGO ELVIN	LEC005	FINANCE	CATHOLIC	CREDIT
11	TABITHA MWENDE	LEC010	ACCOUNTING	UON	PASS
12	FREDRICK MWORIA	LEC009	ARCHITECTURE	STRATHMORE	DISTINCTION
13	ROY OBONYO	LEC002	HOSPITALITY	USIU	CREDIT
14	FELIX MARANGO	LEC007	ENGINEERING	KCAU	CREDIT
15	RODGERS SIMIYU	LEC008	ICT	USIU	PASS
16	FIONA NJERI	LEC004	FINANCE	JKUAT	DISTINCTION
17					

(b) Write a python code to retrieve and display data from the file named Grades.csv that is located in the folder called ICT. Capture and save the screenshot of the resulting output. (3 marks)

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

(c) Using the data in the table (a) above, write python programming codes that will perform the following tasks. Capture and save the screenshots of the resulting outputs.

(i) Output 8 rows randomly from the entire dataset. (3 marks)

(ii) Transform the data into a numeric array and print the output. (3 marks)

(iii) Separate independent variables from dependent variable and print the output. (3 marks)

(iv) Normalise the data and print the output (2 marks)

(v) Compress the data to two attributes (3 marks)

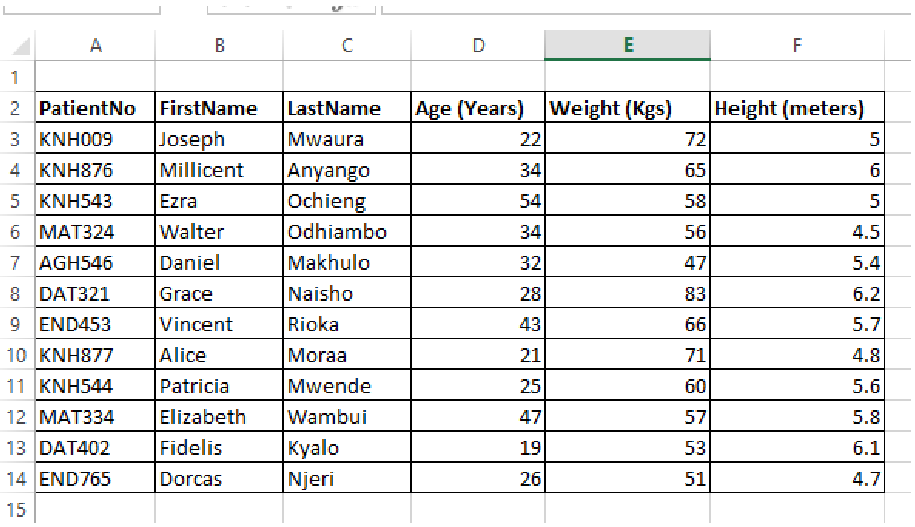
Capture a screenshot to demonstrate how you have performed the above task.

Upload Question 21.

(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) Create an excel document shown below and save it as a comma separated version(CSV) file named "Patient" in the ICT folder of question 1. (3 marks)



PatientNo	FirstName	LastName	Age (Years)	Weight (Kgs)	Height (meters)
KNH009	Joseph	Mwaura	22	72	5
KNH876	Millicent	Anyango	34	65	6
KNH543	Ezra	Ochieng	54	58	5
MAT324	Walter	Odhiambo	34	56	4.5
AGH546	Daniel	Makhulo	32	47	5.4
DAT321	Grace	Naisho	28	83	6.2
END453	Vincent	Rioka	43	66	5.7
KNH877	Alice	Moraa	21	71	4.8
KNH544	Patricia	Mwende	25	60	5.6
MAT334	Elizabeth	Wambui	47	57	5.8
DAT402	Fidelis	Kyalo	19	53	6.1
END765	Dorcas	Njeri	26	51	4.7

Capture screenshots and save them in Question 22 document.

(b) Import the data set created in (a) above into R-Studio

(c) Write the R-studio codes that will perform the following tasks:

(i) Code to return the first 9 rows. (3 marks)

(ii) Code to return the dimensionality of data. (2 marks)

(iii) Code to display the number of rows in the dataset. (2 marks)

(iv) Code to display the variables names. (3 marks)

(v) Code to display the data structure. (2 marks)

(vi) Code to calculate the total weight of all the patients. (2 marks)

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

Upload Question 22 document.

(Total: 20 marks)

23. Create a word processing document named Question 23 and use the word processor document to save your answers to question (a) to (g) below.

Explain how you would perform the above task.

(a) Import the libraries numpy for numerical analysis and matplotlib for visualisation. (3 marks)

(b) Load the sklearn linear regression library (2 marks)

(c) You are given a list of integer numbers (3,21,22,34,54,34,55,67,89,99) and (1,10,14,34,44,36,22,67,79,90). Explain how you would use python to define two lists named X and Y respectively and convert them into an array (3 marks)

(d) Write a python code to display the lists created in (c) on the console screen (2 marks)

(e) Create a well labelled scatter plot for the two lists to illustrate their correlation on a 2D plane (2 marks)

(f) Create an instance of class “LinearRegression” and fit it to the data created in (c) (4 marks)

(g) Display the coefficient of determination of the model using the score function and display the model intercept and coefficient. (4 marks)

Upload Question 23 document.

(Total: 20 marks)

..