UGANDA INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY
PROGRESSIVE ASSESSMENT EXAMINATIONS
ACADEMIC YEAR 2025/2026

PROGRAMME: DIPLOMA IN DATA SCIENCE MANAGEMENT & ANALYTICS (DSMA)

YEAR OF STUDY: YEAR ONE SEMESTER TWO

COURSE: PYTHON PROGRAMMING FOR DATA SCIENTISTS

COURSE CODE : DSA 1202

DATE: THURSDAY 25TH, JUNE 2026

TIME: 2:00 PM – 5:00 PM

DURATION: 3 HOURS

VENUE: LAB 1. LAB 2

INSTRUCTIONS

i. This is a practical examination
ii. Each question is worth 50 Marks
iii. Attempt any two questions

QUESTION 1:

The Ministry of Agriculture, Animal Industry and Fisheries (MAAIF) requires data-driven insights to understand how environmental factors impact coffee productivity across Uganda. To achieve this, the Ministry has partnered with a local AgriTech consultant to analyze daily IoT sensor data collected from various coffee-growing districts. The sensors track daily rainfall, temperature and soil moisture.

As a Data Scientist on this project, you are tasked with building an efficient data pipeline to clean the raw sensor logs, engineer performance metrics and generate visual reports for policymakers.

You are therefore provided with a dataset called "***coffee_farm_data.csv***". This dataset contains the following columns:

Column Name	**Description**
date	Observation Date
district	Farming District
farm_id	Unique Farm Identifier
rainfall_mm	Rainfall in millimeters (mm)
temperature_celsius	Temperature in Celsius
soil_moisture_percentage	Soil Moisture as a percentage
fertilizer_used	Fertilizers used
yield_kg	Coffee yields in kilograms (kg)

The Chief Technology Officer (CTO) emphasizes that your code must remain highly scalable. You must use vectorized operations (via Pandas/NumPy) and completely avoid inefficient row-by-row iteration (such as for loops or iterrows()), as this pipeline will eventually process millions of incoming nationwide IoT records.

Required Tasks (Total: 50 Marks):

(a) **Environment Setup:** Create a Jupyter Notebook named "***coffee_yield_analysis.ipynb***" to document your entire analysis. **(2 Marks)**

(b) **Data Ingestion:** Load the provided "***coffee_farm_data.csv***" dataset into a Pandas DataFrame. **(2 Marks)**

(c) **Data Type Optimization:** Convert the date column from a generic object/string type into a proper "***datetime64***" data type. **(2 Marks)**

(d) **Anomaly Detection & Filtering:** Identify and permanently remove all rows where the rainfall column contains negative values, as these represent physical sensor errors. **(6 Marks)**

(e) **Missing Data Imputation:** Handle missing (null) values across the dataset by replacing them with the respective column's average value. **(6 Marks)**

(f) **Feature Engineering:** Create a new column named "***productivity_score***" using a vectorized calculation. The score is defined as: **(8 Marks)**

$$productivityscore = (0.4 * soilmoisture) + (0.3 * rainfall) + (0.3 * temperature)$$

(g) **Data Aggregation:** Use the Pandas .groupby() method to calculate the monthly averages for rainfall, temperature and yield. **(8 Marks)**

(h) **Data Visualization:** Using the Seaborn library, construct a line plot that clearly illustrates the monthly trends of these averaged variables. **(12 Marks)**

(i) **Plot Annotation:** Add a descriptive title to your visual report reading: "***Monthly Environmental Trends*****"**. **(2 Marks)**

(j) **Exporting Output:** Save the final generated plot as a high-quality image file named "***monthly_environmental_trends.png*****"**. **(2 Marks)**

QUESTION 2:

A major telecommunications provider is facing high customer churn; many prepaid subscribers switch to competing networks shortly after purchasing promotional internet bundles. The Marketing Director requires a data-driven insights report to present to the executive board.

As the project Data Scientist, you have been provided with a subscriber dataset named "***internet_subscribers.csv***", with the following columns:

Column Name	**Description**
subscriber_id	Subscriber Number
region	Customer region
monthly_usage_gb	Data Usage per month in Gbs
complaints	Number of complaints from subscriber
monthly_spending	Monthly Expenditure
subscription_type	Daily, Weekly, Monthly
device_type	Smartphone, Router

You are required to analyze subscriber behavior, flag high-risk customers and visualize regional distribution trends to inform retention strategies.

Required Tasks (Total: 50 Marks)

(a) **Environment Setup**: Create a Jupyter Notebook named "***subscriber_analysis.ipynb***" to document your analysis and code. **(2 Marks)**

(b) **Data Ingestion**: Load the provided "***internet_subscribers.csv***" dataset into a Pandas DataFrame. **(2 Marks)**

(c) **Data Cleansing**: Identify and drop all rows containing missing or null values to ensure data quality. Indicate how many rows were dropped and how many have been kept. **(6 Marks)**

(d) **Feature Engineering**: Create a new column named "***departure_risk***". Use Pandas vectorization to populate it based on the following business rules: **(12 Marks)**

- Assign **"Likely"** if a subscriber's monthly_usage_gb is less than 2 **AND** their complaints count is greater than 3.
- Assign **"Stable"** for all other subscribers.

(e) **Data Aggregation**: Compute and display the average subscription spending (monthly_spend) grouped by each geographical region. **(12 Marks)**

(f) **Data Visualization**: Using Seaborn, construct a grouped bar chart. Plot the geographical region on the x-axis and the total subscriber count on the y-axis, segmenting the bars by departure_risk. **(12 Marks)**

(g) **Chart Annotation**: Add a descriptive title to your visualization reading: "***Customer Departure Risk by Region***". **(2 Marks)**

(h) **Exporting Output**: Save the final generated plot as a PNG image file named "***customer_departure_risk.png***". **(2 Marks)**

QUESTION 3:

The Bank of Uganda is conducting a sector-wide audit to analyze national mobile money usage patterns across the two dominant telecommunications operators: Trace Mobile and Spirit Telecom. As a Junior Data Analyst at the Central Bank, you have been provisioned with a core ledger export named "***mobile_money_transactions.csv***", containing the following columns:

Column Name	Description
date	the date on which a particular transaction was initiated
network	the mobile money company that initiated the transaction – Trace or Spirit
agent_id	the unique identifier for the mobile money agent
transaction_type	the transaction type – deposit or withdrawal
amount	the amount of money either deposited or withdrawn

All logged records represent fully completed, successful transactions.

Your task is to build an end-to-end analytical pipeline that cleans this raw transactional data, extracts key macro-economic indicators and generates executive-level visual insights for policy decisions.

Required Tasks (Total: 50 Marks)

(a) **Environment Setup**: Create a Jupyter Notebook named "***transaction_analysis.ipynb***" to document your analysis and code. **(2 Marks)**

(b) **Data Ingestion:** Load the provided "***mobile_money_transactions.csv***" dataset into a Pandas DataFrame. **(2 Marks)**

(c) **Data Cleansing**: Identify and drop all rows containing missing or null values to ensure data quality. **(2 Marks)**

(d) **Market Share Analysis**: Identify the mobile money company that had the most transactions (deposits and withdrawals combined). Show both the company name and the total number of transactions. **(6 Marks)**

(e) **Liquidity Inflow Analysis**: Identify the company that had the most amount of money deposited onto its network. Show both the company name and the total amount of money that was deposited. **(6 Marks)**

(f) **Macro Temporal Trends**: Identify the day of the week (i.e. Monday, Tuesday, Wednesday, …, or Sunday) when the most amount of money is withdrawn in this dataset (both companies combined). Show both the day of the week and the total amount of money withdrawn. **(8 Marks)**

(g) **Agent Monitoring**: Identify the agent from whom the most amount of money is withdrawn over the weekends (only Saturday and Sunday). Show both the agent ID and the total amount of money withdrawn. **(8 Marks)**

(h) **Data Visualization**: Using Seaborn, construct a grouped bar chart mapping the withdrawals. The x-axis must display the days of the week, the y-axis must show the aggregate withdrawal amounts and the bars must be grouped/colored by network provider. **(12 Marks)**

(i) **Chart Annotation**: Add a descriptive title to your visualization reading: "***Mobile Money Withdrawals per day***". **(2 Marks)**

(j) **Exporting Output**: Save the final generated plot as a PNG image file named "***daily_withdrawals.png***". **(2 Marks)**

QUESTION 4:

The Ministry of Health requires an analytical dashboard to evaluate patient distribution across various districts and health facility levels. This data-driven insight will guide the

ministry in optimizing resource allocation and improving healthcare service delivery nationwide. You are provided with a dataset named "***health_data.csv***" containing the following attributes:

Column Name	Description
date	the date on which the observation was made
district	the district from which the observation was made
facility_level	the level of the health facility
service	the service that was offered on that date and at that facility
number_patients	the number of patients who utilized that service at that facility on that date

Required Tasks (Total: 50 Marks)

(a) **Environment Setup**: Create a Jupyter Notebook named "***health_services_analysis.ipynb***" to document and execute your analytical workflow. **(2 Marks)**

(b) **Data Ingestion**: Load the provided "***health_data.csv***" dataset into a pandas DataFrame. **(2 Marks)**

(c) **Data Reshaping (Pivoting)**: Using the pivot_table method, transform the dataset into a summary table where each row represents a unique district, each column represents a service and the cells contain the total number of patients who utilized that service. Replace all missing values (NaN) with 0, to reflect zero utilization. **(10 Marks)**

(d) **Peak Utilization Analysis**: For every distinct healthcare service, identify the specific district that recorded the highest total patient volume. Your output must display both the district name and the corresponding patient count. **(10 Marks)**

(e) **District-Level Performance**: Calculate the average number of patients served across all service types for each individual district. Ensure that districts with no recorded visits for a specific service default to a value of 0. **(10 Marks)**

(f) **Time-Series Visualization**: Using the Seaborn library, construct a line chart illustrating the daily trend of the average number of patients seeking maternity

services across all health facilities. The x-axis must represent the chronological dates of the visit and the y-axis must represent the average number of patients on that day across all health facilities. **(12 Marks)**

(g) **Chart Annotation**: Enhance the visualization by adding a descriptive title: "***Average Maternity Patients per Day***". **(2 Marks)**

(h) **Exporting Outputs**: Save the finalized line chart as a png file named "***average_maternity_patients.png***". **(2 Marks)**

END