Project

Table of contents

Assignment / Project

Faculty:	Information Technology
Module Code:	ITAYA0-44
Module Name:	Data Analysis with Python
Content Writer:	Doreen Muderede
Internal Moderation:	Community of Practice
Copy Editor:	Ms Sabrina Govender
Total Marks:	100
Submission Week:	Week 7

This module is presented on NQF level 5.

5% will be deducted from the student's assignment mark for each calendar day the assignment is submitted late, up to a maximum of three calendar days. The penalty will be based on the official campus submission date.

Assignments submitted later than three calendar days after the deadline or not submitted will get 0%. [1]

This is individual assignment

Groups should consist of members.

This assignment contributes 30% towards the final mark.

[1] Under no circumstances will assignments be accepted for marking after the assignments of other students have been marked and returned to the students.

Instructions to Students

1. Remember to keep a copy of all submitted assignments.
2. All work must be typed.
3. Please note that you will be evaluated on your writing skills in all your assignments.
4. All work must be submitted through Turnitin and the full Originality Report should be attached to the final assignment. Negative marking will be applied if you are found guilty of plagiarism, poor writing skills or if you have applied incorrect or insufficient referencing. (See the table at the end of this document where the application of negative marking is explained.)
5. Each assignment must include a cover page, table of contents and full bibliography, based on the referencing method applicable to your faculty as applied at Pearson Institute of Higher Education.
6. Use the cover sheet template for the assignment; this is available from myLMS.
7. Students are not allowed to offer their work for sale or to purchase the work of other students. This includes the use of professional assignment writers and websites, such as Essay Box. If this should happen, Pearson Institute of Higher Education reserves the right not to accept future submissions from a student.

Section A

Learning Objectives:

By completing this assignment, students will be able to:

1. Analyse and solve systems of linear equations using matrices and matrix notation.
2. Demonstrate understanding of determinant properties and their significance in real-world systems.
3. Compare and contrast the structure and use of different vector spaces.
4. Apply mathematical modelling to social and applied contexts using concepts from linear algebra.
5. Reflect critically on the ethical and practical implications of mathematical solutions in society.

Assignment Topic:

Students will engage with practical scenarios where matrix algebra, determinants, and vector space theory are applied. Emphasis is placed on modelling, analysis, and ethical reflection. The assignment is designed to promote critical thinking and originality, discouraging rote copying.

Technical Aspects Covered:

- **Systems of Linear Equations:** Writing and solving equations using matrices.
- **Matrix Operations:** Matrix notation, inverse matrices, row reduction.
- **Determinants:** Calculation and interpretation in structural and applied contexts.
- **Vector Spaces:** Dimension, basis, operations, and applications.
- **Mathematical Modelling:** Using linear algebra to model and solve practical, real-world issues.
- **Ethical Reflection:** Considering data bias, fairness, and privacy in mathematical applications.

Marking Criteria		
Criteria	Weight (%)	Description
Mathematical Accuracy	30%	Correct use of matrix methods, equations, and algebraic processes
Application & Interpretation	25%	Real-world context relevance, explanation of how math applies
Research & Original Thought	20%	Use of credible sources, original problem-solving and modeling
Visuals & Representation	10%	Use of diagrams, graphs, or matrix visualizations to explain answers
Reflection & Ethical Considerations	10%	Depth of insight into limitations and consequences of mathematical models
Presentation & Referencing	5%	Clarity, organization, citation of sources, appropriate formatting
Total	100%	

Question 1 — 15 Marks

Study the scenario and complete the question(s) that follow(s):

Understanding Our Community

You've been hired by a nonprofit called ConnectSA, which provides digital literacy and streaming services to underserved communities in South Africa.

They have collected customer data from outreach efforts in Cape Town, Durban, and Soweto.

Your job is to clean, explore, and visualize this data to help them understand who their customers are and how engagement has grown over time.

Use the dataset titled customers.csv containing 100 customer records from a nonprofit streaming initiative.

1a. Load the dataset customers.csv using pandas. Display the first 5 rows and use .info() and .describe() to explore the structure of the data.

(3 marks)

1b. Use NumPy to calculate the mean and standard deviation of the age column. Then create a Boolean array to identify customers under the age of 25.

(4 marks)

1c. Use Boolean indexing to filter and display only the rows where customers are under 25.

3 marks)

1d. Create a new column called age_group using pandas. Segment customers into three groups:

- Youth: age < 25
- Adult: 25 ≤ age < 60
- Senior: age ≥ 60

(5 marks)

[Sub Total 15 Marks]

Question 2 — 35 Marks

Study the scenario and complete the question(s) that follow(s):

Tracking Digital Inclusion in South Africa

Background: The Department of Communications and Digital Technologies has launched a national initiative called ConnectSA, aimed at improving digital access in underserved communities. As part of this initiative, thousands of residents have signed up for subsidized internet services in Cape Town, Durban, and Soweto.

You have been hired as a data analyst to help evaluate the rollout. You have received a dataset of 100 customer records from the pilot phase.

Your job is to clean and analyse the data to uncover trends in digital adoption across regions and age groups. Use the dataset titled customers.csv containing 100 customer records from a nonprofit streaming initiative.

2a. Convert the subscription_date column to datetime format. Extract and create three new columns: year_joined, month_joined, and quarter_joined.

(7 marks)

2b. Use .groupby() to calculate the average age and average subscription year for each age_group. Display the results.

(7 marks)

2c. Create a bar chart showing the number of sign-ups per year. Label the axes and add a title.

(7 marks)

2d. Create a pie chart showing the distribution of customers by age_group.

(7 marks)

2e. Save the cleaned dataset as community_customers_cleaned.csv. Include comments in your code explaining each cleaning step.

(7 marks)

[Sub Total 35 Marks]

Question 3 — 25 Marks

Study the scenario and complete the question(s) that follow(s):

Data-Driven Health Decisions in the Northern Cape

You have joined a provincial analytics team working with mobile clinics in the Northern Cape. These clinics screen patients for chronic conditions and collect basic health metrics: age, sex, BMI, blood pressure, and a disease score based on symptoms and nurse assessments.

The Department of Health wants to use this data to:

- Identify high-risk patients
- Improve data quality
- Visualize regional health trends
- Integrate SQL-based reporting for decision-makers

Your task is to clean, transform, and analyse the dataset using advanced Python techniques and SQL queries. Your insights will guide triage protocols and resource allocation across rural clinics.

Use the dataset titled health_data.csv containing 100 patient records collected from mobile health clinics.

3a. Load the dataset health_data.csv and explore its structure. Identify missing values, outliers, and duplicates.

(5 marks)

3b. Clean the blood_pressure column using regular expressions to remove non-numeric characters and convert it to float.

(5 marks)

3c. Create a new column risk_level using a lambda function:

- If BMI > 30 and disease_score > 80 → "High"
- If BMI > 25 and disease_score > 60 → "Medium"
- Else → "Low"

(5 marks)

3d. Use .groupby() to calculate the average BMI and disease score per risk_level. Display the results.

(5 marks)

3e. Create a box plot of BMI by risk_level and a histogram of age. Label axes and add titles.

(5 marks)

[Sub Total 25 Marks]

Question 4 — 25 Marks

Study the scenario and complete the question(s) that follow(s):

Building a Health Risk Registry for Rural Clinics

Background: The Northern Cape Department of Health is developing a Health Risk Registry to monitor chronic illness trends across rural clinics. The registry will be powered by a relational database and used by nurses, doctors, and policymakers to identify high-risk patients and allocate resources.

You have been given a cleaned dataset of 100 patient screenings.

Your task is to load the data into a SQLite database, write SQL queries to extract insights, classify patients based on risk, and explain how indexing can improve performance as the registry scales.

Use the dataset titled health_data.csv containing 100 patient records collected from mobile health clinics.

4a. Load the cleaned dataset into a SQLite database using pandas and sqlite3.

(5 marks)

4b. Write a SQL query to count patients by sex and risk level. Display the results in Python.

(5 marks)

4c. Write a SQL query to calculate the average disease score per age group:

- Youth: age < 25
- Adult: 25 ≤ age < 60
- Senior: age ≥ 60

(5 marks)

4d. Write a SQL query using CASE to classify patients as "Critical" or "Stable" and export the results to a CSV.

(5 marks)

4e. In your notebook, explain how database indexes could improve query performance in large-scale health systems.

(5 marks)

[Sub Total 25 marks]