Project

Site:	Eduvos LMS	Printed by:	Thendo Siphuma
Course:	Python for Data Science Assessments	Date:	Monday, 17 November 2025, 1:37 PM
Book:	Project		

Table of contents

Assignment / Project

Faculty:	Information Technology
Module Code:	ITPFA0-44
Module Name:	Python for Data Science
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 an individual assignment.

This Choose an item: assignment 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[1] 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 *my*LMS.

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 Objective

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 modeling, 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

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

25 Marks

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

Who Are Our Customers?

You’ve been hired by a global tech company called GlobeStream, which provides cloud-based productivity tools to small businesses. Since launching in early 2020, GlobeStream has grown rapidly across more than 60 countries — but their customer data hasn’t kept up.

They’ve handed you a dataset of 100 customer records. Each row represents a customer who subscribed to GlobeStream between January 2020 and September 2025. The dataset includes basic information such as name, age, city, country, and subscription date.

However, the data is messy:

- The subscription_date column contains inconsistent formats (e.g., “2021-03-15”, “15/03/2021”, “March 15, 2021”)
- Some dates are missing or invalid (e.g., “2030-01-01” or “2019-12-31”)
- There’s no clear segmentation of customers by age or region
- The company has no idea when most customers joined or how their customer base has evolved over time

Your job is to:

- Explore and clean the dataset
- Extract useful time-based features (e.g., year, month, quarter)
- Segment customers into meaningful age groups using your own logic
- Identify patterns in customer sign-ups over time

- Save a cleaned version of the dataset with your new features

GlobeStream’s marketing team will use your analysis to plan targeted campaigns and understand how their global reach has changed over the years.

Use the dataset titled customers.csv *containing 100 customer records.*

1. Load the dataset and explore its structure.

a. What stands out to you? (5 Marks)

b. Clean the subscription date column and extract useful time-based features. (5 Marks)

c. Segment customers into meaningful age groups — define your own logic. (5 Marks)

d. Identify any patterns in customer sign-ups over time. Use visuals to support your findings. (5 Marks)

e. Save a cleaned version of the dataset with your new features. (5 Marks)

[Subtotal 25 Marks]

Rubric

Subtask	Criteria	Marks
1a	Data loaded, observations noted	5
1b	Date cleaned, features extracted	5
1c	Age groups defined and applied	5
1d	Trends identified, visualized	5
1e	Dataset saved correctly	5

End of Question 1

Question 2 25 Marks

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

The Post-Pandemic Product Puzzle

You've been hired by a mid-sized retail company called NovaMart, which operates across Southern Africa. After a tough few years of pandemic-related disruptions, NovaMart is trying to rebuild its product strategy. They've handed you a dataset of all products sold in the past 18 months — including categories, units sold, prices, and revenue.

But there's a catch:

- Some prices are missing due to system errors.
- Some products show suspiciously low sales.
- Revenue figures are inconsistent or missing.
- Management doesn't know which products are driving growth — or which ones are dragging them down.

Your job is to:

- Make sense of the data
- Identify top performers
- Flag underperformers
- Recommend which products to keep or discontinue
- Build a reusable tool that NovaMart can run weekly

Data Provided: products.csv Columns: product_id, product_name, category, units_sold, price, revenue.

Use the dataset titled Products.csv *containing.*

2a. Build a structure to represent the product data. (5 Marks)

2b. Calculate total revenue per product. Handle missing prices creatively (5 Marks)

2c. Flag products that may need to be discontinued (5 Marks)

2d. Visualize the top-performing products. Choose a chart that best tells the story (5 Marks)

2e. Write a short Python function that could help automate this analysis weekly (5 Marks)

[Subtotal 25 Marks]

Rubric

Subtask	Criteria	Marks
2a	**Structure chosen and justified**	**5**
2b	**Revenue calculated correctly**	**5**
2c	**Flagging logic applied**	**5**
2d	**Chart clear and informative**	**5**
2e	**Function reusable and documented**	**5**

End of Question 2

Question 3 25 Marks

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

The Clinic Without a Doctor

You've joined a public health initiative called CareTrack, which supports rural clinics in South Africa. One clinic in the Eastern Cape has no full-time doctor — just nurses and volunteers. They've been collecting patient data for months, hoping someone can help them identify which patients are at highest risk and need urgent care.

They've sent you a dataset with basic health metrics: age, sex, BMI, blood pressure, and a disease score (based on symptoms and lab results). They suspect that high BMI and high disease scores might indicate serious risk — but they need proof.

Your job is to:

- Explore the data

- Look for patterns between BMI and disease score
- Classify patients into risk levels
- Visualize the risk distribution
- Suggest one new feature that could improve their triage system

Data Provided: health_data.csv Columns: patient_id, age, sex, BMI, blood_pressure, disease_score

Use the dataset titled health_data.csv.

3a. Explore the dataset. What variables seem most useful for risk analysis? (5 Marks)

3b. Create a scatter plot to investigate BMI vs. disease score. (5 Marks)

3c. Define your own logic to classify patients into risk levels (5 Marks)

3d. Count how many patients fall into each risk level and visualize it. (5 Marks)

3e. Suggest one additional feature or transformation that could improve the model. Implement it. (5 Marks)

[Subtotal 25 Marks]

Rubric

Subtask	Criteria	Marks
3a	**Variables explored and discussed**	**5**
3b	**Plot created and interpreted**	**5**
3c	**Classification logic applied**	**5**
3d	**Counts and visuals accurate**	**5**
3e	**Feature added and justified**	**5**

End of Question 3

Question 4 25 Marks

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

The Eastern Cape Health Audit

The Eastern Cape Department of Health has launched a province-wide audit of patient records to improve data quality and identify trends in chronic illness. You've been contracted as a data analyst to help clean and analyse a sample dataset collected from clinics in Mthatha, Queenstown, and rural areas around Lusikisiki.

The dataset includes basic health metrics for 100 patients: age, sex, BMI, blood pressure, and a disease score based on nurse assessments. However, the data is messy:

- Some rows are duplicated due to manual entry
- Several fields have missing values
- BMI and disease scores vary widely, with some extreme outliers

Your job is to:

- Clean the dataset by removing duplicates, filling missing values, and normalizing key fields
- Randomly sample 20 patients and compare their statistics to the full dataset
- Create a frequency distribution of age or another variable to uncover patterns
- Apply a simple Naïve Bayes-style logic to classify patients as “Critical” or “Stable”
- Save the cleaned and classified dataset, and include a note in your code about ethical considerations when using health data for automated decision-making

The department will use your findings to guide future data collection and triage protocols in under-resourced clinics.

Data Provided: clinic_patients.csv Columns: patient_id, age, sex, BMI, blood_pressure, disease_score

Use the dataset titled The Eastern Cape Health Audit .csv.

4a. Clean the dataset: remove duplicates, fill missing values, normalize key fields. (5 Marks)

4b. Randomly sample 20 records and compare their statistics to the full dataset. (5 Marks)

4c. Create a frequency distribution of age or another variable. (5 Marks)

4d. Apply a simple Naïve Bayes-style logic to classify patients. (5 Marks)

4e. Save the cleaned and classified dataset. Include a note in your code about ethical considerations. (5 Marks)

[Subtotal 25 Marks]

Rubric

Subtask	Criteria	Marks

4a	**Duplicates removed, missing values filled, normalization applied**	**5**
4b	**Sample selected, comparison done with stats**	**5**
4c	**Histogram created, bins labelled, insights discussed**	**5**
4d	**Classification logic applied correctly**	**5**
4e	**Dataset saved, ethical note included**	**5**

End of Question 4