EXTERNAL INTEGRATED SUMMATIVE ASSESSMENT

MOCK TEST

STUDENT NAME & SURNAME	
ID NUMBER	
EISA REGISTRATION NUMBER	
ASSESSMENT CENTRE	
ASSESSMENT CENTRE ACCREDITATION NUMBER	
QUALIFICATION	Occupational Certificate: Data Science Practitioner
SAQA ID	118708
CREDITS	185
PAPER	
DATE OF EISA	
DURATION	
TOTAL MARKS	

General EISA Rules

1. Students are **only** allowed to use the supplied EISA booklets.
2. Students are **only** allowed to use a black pen for their answers.
3. Students to ensure that their name, surname and EISA registration number appears on the front of your EISA booklet.
4. This is an open book examination.
5. All EISA booklets must be handed back to the invigilator intact. No pages may be torn off from the EISA booklet. The removal of EISA booklets from the examination room is prohibited.
6. Students may make use of a calculator in this EISA.
7. Unless this is an online examination where access to a computer will be made available to you; the use of any communication devices, including smart watches, cell phones, tablets, iPads, headphones and laptops are prohibited.
8. All cell phones are to be switched off for the duration of the EISA.
9. The invigilator will not assist you with the explanation of questions related to the EISA.
10. Students are prohibited from conversing in any manner with other students.
11. Students may not leave the examination venue within one hour of the start of the examination and in the last 10 minutes of the allotted examination period.
12. Students who are found to be disruptive and unruly in the assessment centre will be requested to leave the assessment centre by the invigilator.

I HEREBY CONFIRM THAT I HAVE READ THE ABOVE EISA RULES AND DECLARE THAT I UNDERSTAND AND ACCEPT THE RULES.

_____

SIGNATURE OF STUDENT

Candidate Instructions

- Candidates must complete all questions in this EISA.
- Candidates must ensure that they use only a black pen when completing this EISA.
- Should you require additional space to complete your answer, please request additional paper from your invigilator. Ensure that you indicate your name, surname and EISA registration number at the top of the additional paper. Also ensure that the question number is clearly marked on your additional paper.
- There is one handout in this paper, **Handout A – Project Charter (this must be handed in with your answer sheet).**

Section A: Theory Questions

1. Data Collection and Pre-Processing

1.1. Which of the following is a primary data source?

a) Social media posts
b) TV news broadcasts
c) News articles
d) Data from sensors

1.2. What is the first step in data preprocessing?

a) Data visualisation
b) Data cleaning
c) Data collection
d) Data analysis

1.3. Which of the following formats is commonly used to store structured relational data?

a) CSV
b) JPEG
c) MP4
d) JSON

1.4. True or False: Unstructured data is always stored in a database.

a) True
b) False

1.5. Which tool is most used for pre-processing data?

a) Excel
b) Python
c) SQL
d) Tableau

2. Data Analysis Techniques

2.1. Which of the following techniques is used to identify patterns in a dataset?

a) Data visualisation.
b) Data transformation.
c) Statistical analysis.
d) Data collection.

2.2. Which SQL command is used to retrieve data from a database?

a) INSERT
b) SELECT
c) UPDATE
d) DELETE

2.3. Which method would you use to detect trends in a time series dataset?

a) Histogram
b) Linear regression
c) Scatter plot
d) Pie chart

2.4. True or False: Data normalisation is a technique used to reduce data redundancy.

a) True
b) False

2.5. Which of the following tools is commonly used for statistical data analysis?

a) Excel
b) R
c) PowerPoint
d) Word

3. Data Visualization and Reporting

3.1. Which of the following is a common data visualisation tool?

a) Seaborn
b) Jupyter Notebook
c) SQL
d) Hadoop

3.2. True or False: Data storytelling involves explaining data patterns and trends through visualisations.

a) True
b) False

3.3. Which programming language is commonly used for creating data visualisations?

a) Java
b) Python
c) C++
d) HTML

3.4. Which type of chart is best for showing trends over time?

a) Bar chart
b) Pie chart
c) Line chart
d) Scatter plot

3.5. In a descriptive analytics report, what is the main purpose of a visualisation?

a) To summarise data insights
b) To store data
c) To clean data
d) To collect data

Section B: Practical Questions

1. SQL Queries

You are provided with a pre-populated SQLite database named sales.db. Your task is to explore this database and write a series of SQL queries to perform the tasks detailed below. Queries should be optimised to run within 20 seconds or less.

1.1. Determine which product has been purchased the most
Write a query to find the product that has been purchased the most in terms of quantity. The result should display the product name, and the total quantity sold.

1.2. Identify which store has the widest range of products in stock
Write a query to find out which store carries the most unique products in its inventory (stock). The query should return the store name and the number of different products available.

1.3. Identify which product was purchased by the most individual users
Write a query to find which product was purchased by the largest number of different users. The result should include the product name and the number of individual users who bought it.

1.4. Identify the most popular shipping method
Write a query to find the most popular shipping method used in all sales. The result should display the shipping method and the total number of times it has been used.

1.5. Find the product with the highest quantity in inventory (stock) across both store and warehouse
Write a query to determine which product has the highest combined quantity in both store and warehouse inventories. The result should display the product name and the total quantity in inventory.

2. Database Population and Data Validation

2.1. Database Population

You are provided with a file named books_data.json which contains unstructured data from an online bookstore. Your task is to explore this file, structure, and validate the data as needed to perform the tasks detailed below.

2.1.1. Write a python script that creates an SQLite database named books.db with the following tables and the appropriate FKs (foreign keys) to maintain relationships between the tables

- Categories: Stores the category name of each book
- Books: Stores the title, price, rating, category ID, and description of each book.
- Stock: Stores the stock status (in stock or not) and the stock count for each book.
- ProductIDs: Stores the unique product ID (UPC) for each book.

2.1.2. Write a python script to structure and insert the unstructured data from books_data.json into books.db

- Data should be written to the most suitable table.
- Data should be correctly linked via suitable foreign keys.

2.2. Data validation

Using books.db, write a series of SQL queries to perform the tasks detailed below to validate the data.

2.2.1. Get all books in a category

Write a Python function that, given a category name, returns a list of books (title and price) that belong to that category from the Books table.

2.2.2. Check stock status for a book

Write a Python function that, given a book title, checks if the book is currently in stock by querying the Stock table.

2.2.3. Find the average rating for each category

Write a Python function that, given a category name, calculates the average rating of all books in that category.

2.2.4. Find books that are below a certain price

Write a Python function that retrieves the title and price of all books that are priced below a given threshold.

2.2.5. Identify the best-stocked book within a specific category

Write a Python function that, given a category name, finds the book with the highest stock count in that category.

3. Wine Quality Dataset Analysis

You are provided with a file named wine_quality.zip which contains two datasets related to red and white vinho verde wine samples, from the north of Portugal.

Your goal is to explore this dataset. Clean it if necessary. Identify any biases. Then perform a series of regression analyses to uncover trends. Detect potential biases due to excluded variables, and finally make predictions on wine quality.

3.1. Loading the Dataset and Data Preparation

3.1.1. Data loading

Load the dataset using pandas.DataFrame and ensure the data is correctly formatted (i.e., all columns should be numeric).

3.1.2. Data Preparation

Write a function to separate the target variable (quality) from the features and prepare the data for further analysis (e.g., scaling or normalisation if necessary).

3.2. Visualisations and Exploratory Data Analysis

Analyse the distribution of classes and the relationships between features to identify patterns or biases in the data.

3.2.1. Class Distribution Visualisation

Task: Implement a function that visualises the distribution of the wine quality classes. This will help identify any imbalances in the dataset.

Question: Is the dataset imbalanced in terms of wine quality classes? What impact could this have on your regression models?

3.2.2. Feature Distribution and Outlier Detection

Task: Create box plots for each feature to visualise the distribution of values and detect any outliers.

Question: Which features have significant outliers? How might these outliers affect model performance?

3.2.3. Correlation Heatmap

Task: Generate a heatmap that shows the correlation (relationship) between features and the target variable (quality). This will help you understand which features are most important to the prediction of wine quality.

Question: Which features are strongly correlated with wine quality? Do these correlations suggest any biases or trends that need further investigation?

3.3. Regression Models

Apply different regression techniques to predict wine quality based on the dataset, evaluate the performance of their models, and visualise the results.

3.3.1. Class Distribution Visualisation

Implement a simple linear regression model using one feature (e.g., alcohol) to predict wine quality. Visualise the regression line and the residuals.

3.3.2. Multiple Linear Regression

Write a function to perform multiple linear regression using several features (e.g., alcohol, volatile acidity, sulphates, citric acid) to predict wine quality. Visualise the model coefficients.

3.3.3. Polynomial Regression

Create a polynomial regression model using one feature (e.g., alcohol) with a degree of 2 or higher. Compare its performance to the simple linear regression model.