DATA SCIENCE 500

RICHFIELD
richfield.ac.za

Name and Surname: _____

Student ITS No: _____

Qualification: _____ **Year of Study:** _____ **Semester:** _____

Assignment due date: _____ **Date submitted:** _____

QUESTION	EXAMINER MARKS	MODERATOR MARKS	REMARKS

ASSIGNMENT INSTRUCTIONS

Please tick each box to confirm completion.

Use Times New Roman font, size 12, with 1.5 line spacing throughout the document.
Apply Harvard Referencing Style for all citations and references.

For essay-style assignments, please include the following sections:

Table of Contents
Introduction
Main Body (with relevant subheadings)
Conclusion
References

Submit the assignment in PDF format on Moodle.
Use the specified cover page provided.
Include a signed declaration of originality.

DECLARATION OF ORIGINALITY:
I hereby declare that this assignment is my own work and has not been copied from any other source except where due acknowledgment is made. I affirm that all sources used have been properly cited and that this submission complies with the institution's policies on academic integrity and plagiarism.

Student Signature: _____ **Date:** _____

Question 1 (20 marks)

A data science team at a large online retailer is analysing website activity to assess customer engagement and predict future behaviour. The company tracks two events for every visitor:

- Event A: The visitor clicked on a promotional banner during their session.
- Event B: The visitor returned to the website within 7 days (repeat visitor).

Based on historical web traffic logs from 10,000 visitors, the following frequencies were recorded:

	Returned (B)	**Did Not Return (¬B)**	**Total**
Clicked Banner (A)	2,000	1,000	3,000
Did Not Click Banner (¬A)	2,500	4,500	7,000
Total	4,500	5,500	10,000

a. What is the probability that a randomly selected visitor clicked on a banner and returned within 7 days? **(3 marks)**

b. What is the probability that a visitor either clicked a banner or returned within 7 days (or both)? **(3 marks)**

c. Given that a visitor returned within 7 days, what is the probability that they had clicked on a banner? **(4 marks)**

d. If the marketing team claims that banner-clicking leads to more returning users, do the probabilities support this claim? **(5 marks)**

e. Are the events "clicking on a banner" and "returning within 7 days" independent? Show your working and explain what this means for the business. **(5 marks)**

Question 2 (25 marks)

A fitness study tracked the daily step counts of 15 randomly selected participants over 14 consecutive days. Each participant was previously assessed and assigned a fitness level reflecting their typical physical activity. Your task is to simulate realistic daily step count data for each participant based on their fitness level and then perform a detailed analysis of the generated data.

Part 1: Dataset Creation **(5 marks)**

1. Randomly assign 15 participants a fitness level:
   - 0 = Low
   - 1 = Moderate
   - 2 = High
2. Based on each participant's fitness level, generate 14 daily step counts using a normal distribution with:
   - Low: mean = 6000 steps, SD = 600
   - Moderate: mean = 7500 steps, SD = 500
   - High: mean = 9000 steps, SD = 700
3. Round the step counts to the nearest integer, and clip values to the range 3000 to 15000 to maintain realistic daily steps.
4. Store the data in a NumPy array of shape (15, 14), where rows represent participants and columns represent days.
5. Display:
   - The fitness level assigned to each participant
   - The full generated dataset (step counts for all participants over 14 days)

Part 2: Data Analysis **(20 marks)**

Using the dataset created in Part 1, perform the following:

a. Calculate the average daily steps per participant. Sort these averages in descending order and display the top 5 participants along with their averages.

b. Calculate the overall mean and standard deviation of all step counts combined. Round both to the nearest whole number.
c. Compute the median daily steps for each participant. Identify and display the participant(s) with the highest and lowest median values, including their participant numbers and median values.
d. Count and display how many participants have an average daily step count above 8000 over the 14 days.
e. Compute and display the 25th, 50th (median), and 75th percentiles of all step counts combined across all participants and days.

Question 3 (25 marks)

You are a junior data analyst at a government agency tasked with understanding how work patterns and education levels vary across income brackets. You will be working with a messy dataset that requires thorough cleaning before analysis can begin. Once the data is properly prepared, you will create a series of subplots to uncover key trends and insights about income distribution patterns.

Dataset: https://www.kaggle.com/datasets/jainaru/adult-income-census-dataset

Part 1: Dataset Preparation and Cleaning (5 marks)

1. Import the dataset using the Pandas library
2. Locate and address missing data points (indicated by "?" values in the dataset)
3. Apply appropriate imputation methods based on data type:
   - Categorical variables: Replace missing values with the most frequently occurring value (mode)
   - Numerical variables: Replace missing values with the middle value (median)
4. Eliminate any irrelevant entries or duplicate records discovered during the cleaning process

Part 2: Data Visualisation **(20 marks)**

Create a 2×2 grid of Matplotlib subplots with the following graphs:

1. **Top-Left: Stacked Bar Chart**
   Showing the distribution of individuals by gender, with income categories (<=50K and >50K) represented as stacked segments.
2. **Top-Right: Line Graph**
   Tracking the relationship between age and average hours worked per week, displaying separate trend lines for each income bracket.
3. **Bottom-Left: Histogram**
   Comparing the distribution patterns of weekly work hours across both income groups using distinct colours and a comprehensive legend.
4. **Bottom-Right: Grouped Bar Chart**
   Presenting average education level (education-num) for different occupational categories, with side-by-side bars comparing the two income groups.

Each graph should have a clear, descriptive title, properly labelled axes with units and descriptions, legends where multiple data series appear, and strategic annotations highlighting significant patterns or outliers. The overall design should use harmonious colour schemes, legible typography, and appropriate whitespace to maintain a contemporary, professional appearance that enhances readability and visual appeal.

Question 4 (30 marks)

Analyse how years of experience, education level, and work location affect the salary of software engineers. Use the following data:

- Years of Experience
  [1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 13, 15, 16, 18, 20]
- Education Level (0: Bachelor's, 1: Master's, 2: PhD)
  [0, 0, 1, 1, 0, 2, 1, 2, 1, 2, 0, 1, 2, 0, 1]

- Location (0: Remote, 1: On-site, 2: Hybrid)

  [0, 1, 1, 2, 0, 2, 2, 1, 1, 0, 2, 2, 1, 0, 1]
- Salary (in thousands of USD)

  [48, 53, 60, 65, 68, 80, 78, 88, 90, 100, 92, 105, 108, 115, 120]

1. Create a 3-colour scatter plot (using matplotlib) showing years of experience vs salary.
2. Use different marker shapes for education level and different colours for location.
3. Add lines of best fit for each location group (Remote, On-site, Hybrid).
4. Fit a separate line per location using only experience as the independent variable.
5. Create a multiple linear regression model using sklearn where:
   - Independent variables: Years of Experience, Experience^2, Education Level, Location
   - Target: Salary
6. Display the model's coefficients and intercept. Briefly explain which variable has the greatest impact.
7. Use your model to predict salary for:
   - A software engineer with 9 years of experience, a Master's degree, and a Remote job.
   - A software engineer with 14 years of experience, a PhD, and a Hybrid position.
8. Create a combined graph that includes:
   - The original scatter points (coloured and shaped)
   - Lines of best fit for location (from step 2)
   - The two predicted points from step 5 (use special markers like stars)