Unit:
Data Analysis and Visualisation with Python

Assignment title:
A Comprehensive Data Analysis Project

Sample Assignment

Important notes

- Please refer to the *Assignment Presentation Requirements* for advice on how to set out your assignment. These can be found on the NCC Education website. Hover over 'About Us' on the main menu and then navigate to 'Policies and Procedures' then scroll to the 'Student Support' area.
- You **must** read the NCC Education document *Academic Misconduct Policy* and ensure that you acknowledge all the sources that you use in your work. These documents are available on the NCC Education website. Hover over 'About Us' on the main menu and then navigate to 'Policies and Procedures' then scroll to the 'Student Support' area.
- You **must** complete the *Statement and Confirmation of Own Work*. The form is available on the NCC Education website. Hover over 'About Us' on the main menu and then navigate to 'Policies and Procedures' then scroll to the 'Student Support' area.
- Please make a note of the recommended word count. You could lose marks if you write 10% more or less than this.
- You must submit a paper copy and digital copy (on disk or similarly acceptable medium). Media containing viruses, or media that cannot be run directly, will result in a fail grade being awarded for this assessment.
- All electronic media will be checked for plagiarism.

Scenario

You need to take on the role of a data analyst, tasked with selecting a dataset from a public source, conducting preliminary data exploration and cleaning, performing detailed analysis, constructing a predictive model, and finally evaluating the model's performance.

All code for this assignment should be written in Python using the Python libraries covered in this unit, and should be developed using Jupyter Notebook.

Each of the following tasks outlined requires both Python code for practical data analysis and the inclusion of thorough analysis, justification, and discussion (word limit) within your report to support your methodologies and findings.

Task 1 – 10 Marks

Dataset Selection

Identify and select a publicly available dataset. Go to Kaggle Datasets. Explore the available dataset related to decision trees and classification. Select a dataset with categorical target variable.

Provide a brief description of the dataset, including its source, the type of data it contains, and the potential questions or problems it could help address. (300 words)

Task 2 – 25 Marks

Data Exploration and Cleaning

Conduct an initial exploration to understand the structure, variables, and any patterns or anomalies within the dataset.

Using descriptive analytics to grasp the central tendency and dispersion.

Clean the data by handling missing values, outliers, and any inconsistencies to prepare it for analysis.

Additionally, employ appropriate visualisations to illustrate the dataset's characteristics, enhancing the understanding of its underlying patterns and anomalies.

If your dataset initially appears to have no missing or incorrect data, consider a hypothetical scenario where such issues are presented. Explain the strategies you would employ to identify and handle them.

Summarises your approach to data exploration and cleaning by discussing the initial exploration process to understand data structure and anomalies, the application of descriptive statistics to analyse the dataset's central tendency and dispersion, and the comprehensive cleaning process including handling of missing values, outliers, and inconsistencies. (700 words)

Task 3 – 15 Marks
Feature Selection and Scaling

After cleaning, use a correlation matrix to identify the relationships between variables. Select the features with strong correlations with the target variables.

Determine which features would benefit from scaling and apply appropriate scaling techniques to the identified features.

Additionally, analyse the correlation matrix thoroughly and justify your choice of selected features based on their relationship with the target variable. Critically analyse and explain your decision on whether or not scaling is necessary for each feature, considering the nature of the data and the potential impact on model performance. (700 words)

Task 4 – 25 Marks
Model Construction

Build a decision tree classifier to predict the target variable. Tune parameters for optimal performance.

In your approach, justify the decisions made in selecting the method for dividing your data, evaluating model reliability, and fine-tuning the model's settings to enhance its predictive accuracy. (600 words)

Task 5 – 25 Marks
Model Evaluation

Evaluate your model using appropriate metrics with justification. Discuss what these metrics tell you about your model’s effectiveness and any weaknesses.

Incorporate charts or diagrams to effectively demonstrate evaluation metrics, ensuring each is accompanied by a clear explanation of its significance and implications. (700 words)

Submission requirements

- Your program must be submitted as a zip file of the full project.
    - All code provided should be written in **Python using Jupyter Notebook** for this assignment. No marks will be awarded for code written in another language.
- Your report including diagrams and materials associated with the tasks above should be presented in a word-processed document. **(3000 words)**
- All references and citations must use the Harvard Style.

Candidate checklist

Please use the following checklist to ensure that your work is ready for submission.

Have you read the NCC Education document *Academic Misconduct Policy* and ensured that you have acknowledged all the sources that you have used in your work?	❑
Have you completed the *Statement and Confirmation of Own Work* form and attached it to your assignment? **You must do this.**	❑
Have you ensured that your work has not gone over or under the recommended word count by more than 10%?	❑
Have you ensured that your work does not contain viruses and can be run directly?	❑