Unit:
Data Analysis and Visualisation with Python

Assignment title:
A Comprehensive Data Analysis Project

Sample Assignment

Markers are advised that, unless a task specifies that an answer be provided in a particular form, then an answer that is correct (factually or in practical terms) **must** be given the available marks. If there is doubt as to the correctness of an answer, the relevant NCC Education materials should be the first authority.

This marking scheme has been prepared as a **guide only** to markers and there will frequently be many alternative responses which will provide a valid answer.

Each candidate's script must be fully annotated with the marker's comments (where applicable) and the marks allocated for each part of the tasks.

Throughout the marking, please credit any valid alternative point.

Where markers award half marks in any part of a task, they should ensure that the total mark recorded for the task is rounded up to a whole mark.

Marker's comments:

Moderator's comments:

Mark:	**Moderated mark:**	**Final mark:**

Penalties applied for academic malpractice:

Task 1

(LO 1)– 10 Marks

Dataset Selection

The candidate class list should incorporate justification and discussion as to why each class was selected for inclusion, and how its relationship to other classes was derived. The class diagram should show attributes, operations, scope and relationship of classes to each other.

0-1 marks	2-3 marks	4-5 marks	5-6 marks	7-10 marks
Selects an inappropriate dataset that does not have a categorical target variable or is not relevant to decision trees and classification.	*Selects a dataset that somewhat meets the criteria but may lack a clear categorical target variable or full relevance to decision trees and classification*	*Selects an appropriate dataset with a categorical target variable, suitable for decision trees for classification, but the choice lacks optimal relevance or potential for insightful analysis.*	*Selects a suitable dataset with a clear categorical target variable, relevant and applicable to decision trees and classification, with some potential for insightful analysis.*	*Selects an ideal dataset with a well-defined categorical target variable, highly relevant and offering extensive potential for deep insights through decision trees and classification.*
Offers a vague or missing description of the dataset, lacking information on its source, data type, and any analytical potential.	*Provides a basic description, mentioning the source and type of data but fails to effectively outline the dataset's potential for addressing specific questions or problems.*	*Gives a clear description of the dataset, including source and type, with a general overview of potential analytical applications, but lacks depth or specificity.*	*Delivers a detailed description of the dataset, its source, the type of data it includes, and a solid rationale for its selection, including specific questions or problems it could help solve.*	*Presents an exhaustive and insightful description of the dataset, comprehensively detailing its source, data type, and elaborating on the dataset's capacity to provide answers to complex analytical questions or address particular problems.*

Task 2

(LO 2, 3, 4)– 25 Marks

Data Exploration and Cleaning – using appropriate visualisations, encoding categorical variables, scaling features if needed, handling outliers, anomalies, and missing/inconsistent/incorrect data.

0-7 marks	8-9 marks	10-14 marks	15-17 marks	18-25 marks
Demonstrates minimal exploration with no clear understanding of data structure.	*Demonstrates basic understanding of the dataset's structure with superficial exploration.*	*Demonstrates a satisfactory level of data exploration with some understanding of variables and patterns.*	*Demonstrates a thorough exploration and a good understanding of the dataset's structure and variables.*	*Demonstrates an extensive and detailed exploration and a deep understanding of the dataset's intricacies.*
Provides no descriptive statistics or incorrect measures of central tendency and dispersion.	*Basic use of descriptive analytics but understanding of central tendency and dispersion is superficial.*	*Adequate use of descriptive analytics with a fair understanding of measures of central tendency and dispersion.*	*Competent use of descriptive analytics with a good understanding of central tendency and dispersion.*	*Competent user of descriptive analytics demonstrating a deep understanding of measures of central tendency and dispersion.*
Incorrect use of visualisations. Charts, if used, are not appropriate for the data set.	*Limited use of visualisations. Basic boxplots, frequency tables or histograms are used but with limited effectiveness.*	*Adequate use of visualisations. Boxplots, and histograms are used correctly to show data distribution.*	*Good use of visualisations. Boxplots, bar charts, histograms, and other charts are well utilised to illustrate data characteristics.*	*Excellent us of visualisations. Boxplots, bar charts, histograms, and other charts are excellently crafted and highly informative.*
Visualisations do not aid in understanding the data, with no boxplots, frequency tables or	*Visualisations show some aspects of the data but are not fully utilised to enhance understanding.*	*Visualisations contribute to a better but not complete understanding of the dataset's characteristics*	*Visualisations are effectively used to identify underlying patterns and anomalies in the dataset if there are any.*	*Visualisations are integral to the exploration and provide deep insights into the dataset, enhancing the understanding of complex*

histogram present.				*patterns and relationships.*
Inadequate cleaning of data with significant missing values, outliers, and inconsistencies remaining, OR	*Some data cleaning is attempted, but several issues with data quality persist.*	*Data is cleaned to satisfactory level, but minor issues may still exist.*	*Data is cleaned effectively with most issues addressed, though slight room for improvement remains.*	*All known issues addressed, demonstrating a high standard of data preparation for analysis.*

Task 3

(LO 1, 2, 3, 5)– 15 Marks

A correlation matrix should be used to identify the features with strong correlations to the target variable. Z-score feature scale or other appropriate scaling techniques should be used.

0-3 marks	**4-5 marks**	**6-8 marks**	**9-10 marks**	**11-15 marks**
Show minimal effort in applying the correlation matrix. *Lacks identification of key features with strong correlations to the target variable.* *No attempt at feature scaling or inappropriate scaling techniques used.* *Fails to analyse the correlation matrix or justify feature selection and*	*Applies the correlation matrix but with limited understanding.* *Identifies some features with correlations to the target variable but with inaccuracies.* *Attempts feature scaling but with significant misunderstandings of appropriate techniques.* *Provides a basic analysis of the correlation matrix with minimal*	*Correctly uses the correlation matrix with adequate explanation.* *Selects features with strong correlations to the target variable with minor errors.* *Applies basic feature scaling techniques correctly to some, but not all, identified features.* *Provides a clear analysis of the correlation matrix and justifies the selection of*	*Uses the correlation matrix effectively to select features.* *Accurately identifies and selects features with strong correlations to the target variable.* *Demonstrates an understanding of feature scaling with the correct application of techniques to most identified features.* *Thoroughly analyses the correlation matrix and provides well-reasoned justifications for*	*Demonstrates a comprehensive understanding of the correlation matrix's use in feature selection.* *Correctly identifies all relevant features with strong correlations to the target variable.* *Applies feature scaling with an advanced understanding of appropriate techniques, improving model performance.* *Presents a comprehensive and insightful analysis of the correlation matrix, with compelling*

scaling decisions. *Lacks understanding of the relationships between variables and their impact on model performance*	*justification for feature selection and scaling.* *The explanation lacks depth, and the impact on model performance is not adequately addressed.*	*features and the need for scaling.* *The explanation may lack detail in places or not fully consider the impact on model performance.*	*feature selection and scaling decisions.* *Demonstrates an understanding of their potential impact on model performance, with minor gaps in analysis.*	*justifications for feature selection and scaling.* *Demonstrates a deep understanding of how these decisions affect model performance, with a critical evaluation of all aspects.*

Task 4

(LO 1, 2, 3, 5)– 25 Marks

Build a decision tree – dataset split, k-fold cross validation, constructing a decision tree, and parameters tuning.

0-7 marks	**8-9 marks**	**10-14 marks**	**15-17 marks**	**18-25 marks**
No or little evidence of splitting of dataset into training and test sets.	*Basic splitting of dataset but with a skewed or inappropriate ratio.*	*Adequate splitting of the dataset with a reasonable train-test ratio.*	*Correctly splits the dataset with a justified train-test ratio that reflects best practices.*	*Demonstrates a sophisticated approach to splitting the dataset.*
No use of cross-validation, leading to potential overfitting or underfitting.	*Limited use of cross-validation, with poor rationale for chosen method.*	*Uses cross-validation, but the approach may not be fully justified or optimised.*	*Implements cross-validation effectively to assess model robustness.*	*Utilises advanced cross-validation techniques, like K-fold, providing a thorough evaluation of model performance.*
Decision tree classifier is either not implemented or is implemented without any consideration of parameters, resulting in poor model performance.	*Decision tree classifier is fitted to the data, but with default parameters and no tuning, leading to suboptimal performance.*	*Decision tree classifier is correctly fitted to the data, with some parameters adjusted based on intuition rather than a systematic approach.*	*Decision tree classifier is well-fitted to the data with parameters like Gini index, entropy, and max tree height being tuned based on a structured approach.*	*Decision tree classifier is expertly fitted with a deep understanding of the impact of parameters. All relevant parameters, including Gini index, entropy, and*

				max tree height, are tuned for optimal performance.
No parameter tuning attempted or tuning done without any understanding of parameters like Gini index, entropy, or max tree height.	Attempts at parameter tuning are made, but with limited success due to poor choice of values or misunderstanding of the parameters' impact.	Some parameter tuning is evident, with a focus on one or two parameters like Gini index or entropy, but not all relevant parameters are considered.	Parameter tuning is performed, resulting in a noticeable improvement in model performance.	Comprehensive parameter tuning is conducted with a justified approach, leading to superior model performance.
No or little justification for the method of data division, model reliability evaluation, or parameter tuning, lacking critical reflection on choices made.	Limited justification for decisions regarding data division, model evaluation, and parameter adjustments, with minimal reflection on their impact on model performance.	Provides an adequate justification for chosen methods of data division, model reliability evaluation, and parameter tuning, with some reflection on their effectiveness and limitations.	Provides a well-reasoned justification for the approach to data division, evaluating model reliability, and fine-tuning parameters, clearly connecting decisions to model performance outcomes.	Provides a comprehensive and insightful justification for all decisions made throughout the model construction process, from data division to parameter tuning, demonstrating a deep understanding of their impact on predictive accuracy and model reliability.

Task 5

(LO 2, 3, 4, 5)– 25 Marks

Model evaluation – Selecting appropriate classification metrics based on the dataset chosen by student. Evaluate the model's significance and implications.

0-7 marks	8-9 marks	10-14 marks	15-17 marks	18-25 marks
Metrics used are not appropriate for the model or the problem.	*Basic metrics such as Accuracy are used, but more comprehensive metrics are missing.*	*Appropriate metrics are used based on the dataset chosen by student.*	*Comprehensive metrics such as Accuracy, Confusion Matrix, F1-score, or ROC AUC are used.*	*All appropriate metrics are used, including Accuracy, Confusion Matrix, F1-score, and ROC AUC.*
Little to no explanation is provided for the metrics chosen.	*Limited explanation is provided for the metrics chosen.*	*Adequate justification is provided for the metrics chosen.*	*Thorough justification is provided for the metrics chosen.*	*In-depth justification is provided for the metrics chosen.*
Charts or diagrams are absent or do not effectively demonstrate evaluation metrics.	*Charts or diagrams are present but lack clear explanations or relevance.*	*Charts or diagrams are used to demonstrate evaluation metrics but may lack detailed analysis.*	*Charts or diagrams effectively demonstrate evaluation metrics and are accompanied by clear explanations.*	*Charts or diagrams are used excellently to enhance the understanding of evaluation metrics.*
No discussion on the significance or implications of the metrics.	*Attempt made to explain the significance of the metrics, but the explanation is superficial or lacks depth.*	*Significance and implications of metrics are explained, but may not cover all relevant aspects.*	*Limitations are acknowledged, and a reasonable level of critical analysis is applied to the model's weaknesses.*	*Critical evaluation of the model's performance is insightful, considering both the strengths and limitations of the model.*

Learning Outcomes matrix

Task	Learning Outcomes assessed	Marker can differentiate between varying levels of achievement
1	1	Yes
2	2, 3, 4	Yes
3	1, 2, 3, 5	Yes
4	1, 2, 3, 5	Yes
5	2, 3, 4, 5	Yes

Grade descriptors

Learning Outcome	Fail	Referral	Pass	Merit	Distinction
Demonstrate knowledge and apply principles of data science to analyse business problems	Has basic recognition of data science principles but cannot apply them to business problems.	Shows limited understanding of data science principles with minimal application to simple business problems.	Adequately applies data science principles to analyse standard business problems.	Soundly applies data science principles and techniques to a range of business problems, showing good understanding and application.	Applies a comprehensive range of data science principles to analyse complex business problems innovatively and effectively.
Be able to produce, comprehend and run Python code using Jupyter Notebook	Can produce basic Python code but struggles with comprehension and execution in Jupyter Notebook.	Can produce and run Python code in Jupyter Notebook but with limited understanding and numerous errors.	Adequately writes, understands, and executes Python code in Jupyter Notebook with few errors.	Proficiently writes, comprehends, and runs Python code in Jupyter Notebook, demonstrating good coding practices.	Excellently writes, understands, and executes Python code in Jupyter Notebook with high efficiency and adherence to advanced coding standards.
Be able to analyse and interpret data using relevant Python packages	Has basic knowledge of Python packages but cannot effectively analyse or interpret data.	Has limited ability to use Python packages for data analysis and interpretation with frequent misunderstandings.	Can use Python packages to adequately analyse and correctly interpret data.	Effectively utilises Python packages to analyse and interpret data with insightful understanding.	Masterfully uses Python packages to perform sophisticated data analysis and interpretation, providing deep insights.
Create effective visualisations using Matplotlib	Can create only basic visualisations with significant guidance needed.	Creates simple visualisations but lacks effectiveness and clarity.	Produces clear and adequate visualisations using Matplotlib and Seaborn.	Creates detailed and highly effective visualisations, demonstrating a good	Designs exceptional visualisations with Matplotlib and Seaborn that are

and Seaborn				understanding of visual design principles.	insightful and tailored to specific data narratives.
Apply data analysis and visualisation techniques to real-world datasets	Demonstrates minimal understanding of applying data analysis and visualisation techniques to real-world data.	Applies basic data analysis and visualisation techniques to real-world data but with limited effectiveness.	Adequately applies data analysis and visualisation techniques to real-world datasets with competent execution.	Effectively and confidently applies data analysis and visualisation techniques to real-world datasets, showing good interpretative skills.	Excellently applies a wide range of data analysis and visualisation techniques to real-world datasets, showcasing creativity and high-level proficiency.