7CS033 Marking Scheme / Rubric

Week 1

Workshop 1 – Data analysis with Pandas.

Complete the workshop task and compile a 600-words report based on the workshop task by answering all the questions throughout the Notebook including in the Report section.

Criteria	Points
Dataset: • Dataset correctly loaded. • Personalised dataset is correctly generated.	2
All the given code attempted and executed on the personalised dataset. Description of code is provided as necessary.	7
All the questions are answered appropriately.	7
Overall presentation.	3
Reference: Use Harvard referencing format as appropriate (not included in word limit).	1

Week 2

Workshop 2 – Loan prediction using decision tree classifier.

Complete the workshop task and compile a report based on the workshop task. Report should discuss the practical task and results focusing on:

Criteria	Points
Personalised dataset is correctly generated.	1
Demonstrated appropriate use of • Dataframe functions/properties • Feature selection • LabelEncoder	6
Develop a decision tree classifier.	3
Feature Selection: Appropriate features selected and comparison of two models based on feature selection.	4

Result: Discussion of the results based on the evaluation	4
Overall presentation and reference, use of Harvard referencing format as appropriate (not included in word limit	2

Week 3

Workshop 3 – Sentiment Analysis

Complete the workshop task and compile a report based on the workshop task.

Report should discuss the practical task and results focusing on:

Criteria	**Points**
Dataset collection/selection.	4
Demonstrated understanding of text mining and text pre-processing.	4
Visualisation. Performed appropriate and meaningful visualisation methods to inform the data analysis and result and discussion.	2
Sentiment analyser: Discussed how sentiment analysis model is developed using external library.	3
Result and discussion: • Evaluation of the sentiment analyser. • Compare and discuss the results obtained.	5
Overall presentation and reference, use of Harvard referencing format as appropriate (not included in word limit).	2

NOTE: Optional Workshops 4 and 5.

You are required to submit 'Workshop 4' or 'Workshop 5' for the assessment NOT both.

Suggestion: Work on both the workshops and submit the one in which you have more confidence.

Week 4

Workshop 4 – K-Means clustering

Complete the workshop task and compile a report based on the workshop task.

Report will be a comparative analysis and should discuss the experiments conducted including investigation of the given problem based on the chosen durations (it is expected to compare two durations), and interpretation and comparison of the results. Students are encouraged to consider the following points while compiling the report.

Criteria	**Points**
Dataset: Which data points (durations) are considered for clustering (e.g. cluster highest and lowest risk areas) using K-means clustering and why?	3
Data processing: Brief description of the analysis and pre-processing performed and reasons why these processes are required (e.g., it's impact on the results). HINT: you can include the below points while discussing your work: Geolocation: How latitude and longitude are retrieved and if the latitude and longitude are provided in the given dataset, are the retrieved values are same? Data transformation: Briefly describe data transformation as appropriate (Expected steps – Geolocation inserted, Label encoder, Normalisation).	6
K-Means algorithm: Justification of the use of the K-Means algorithm for the given problem and implementation details.	3
Discussion and interpretation of the results: Describe your understanding of the clustering results based on the maps produced. Discuss whether you agree or disagree with the results	5

achieved? Provide evidence to support the findings, e.g., a news website supporting or contrasting the clustering results.	
Overall presentation and reference, use of Harvard referencing format as appropriate (not included in word limit).	3

Week 5

Workshop 5 – Hierarchical clustering

Identifying best and the worst place to live in West Midlands based on the street crime data using hierarchical clustering.

Complete the workshop task and compile a report based on the workshop task.
Write a report of 600-words that should discuss the experiments conducted including investigation of the given problem based on the dataset, and interpretation and comparison of the results. You are encouraged to consider the following points while compiling the report:

Criteria	**Points**
Which data points (duration) are considered for clustering (e.g. cluster highest and lowest risk areas using Hierarchical clustering and why?	3
Brief description of the analysis and pre-processing performed and reason why these processes are required (e.g., it's impact on the results).	5
Demonstrated understanding of the different processes involved (data transformation, normalisation, cluster assignments).	5
Result and conclusion: Discuss the results (your understanding) of the clustering algorithm on clustering crime dataset on selected location (town other than Wolverhampton) with evidence.	5
Overall presentation and reference, use of Harvard referencing format as appropriate (not included in word limit).	2

Week 6

Workshop 6 – Market Basket Analysis using association rule mining.

Complete the workshop task (Jupyter Notebook) and compile a report based on the workshop task.

Criteria	**Points**
Dataset: Dataset correctly generated.	2
Demonstrated understanding of the overall process.	4

Discussion of the proposed three different settings, including parameter selected.	4
*Discussion of the rules generated and reasoning of why those rules were generated.	4
Findings (code and discussion) of the selected duration.	4
Overall presentation and reference, use of Harvard referencing format as appropriate (not included in word limit).	2

***HINT:** Discussion should be based on the relation of the *most frequent items bought and frequent items bought together*.

Example:

> Consider the rules for Setting '*1*'
> Display the *n* (*n* = 10) most frequent items in the dataset. Discuss, how many of these items can be found in rules generated for setting '*1*'? Discuss, are all the top *n* items included in the rules? Provide an explanation as to why these items may be missing/present.