DESCRIPTION OF COURSEWORK

Course Code	AIT403
Course Name	Advanced Data Analytics
Lecturer	Teo Bee Guan
Academic Session	2025/09
Assessment Title	Test 1

A. Introduction/ Situation/ Background Information

This course introduces algorithmic techniques that form the foundation for processing and analysing massive datasets in various formats. It prepares students to apply these techniques to real-world problems. Through this assignment, students will gain hands-on experience in solving practical challenges using advanced data analysis methods.

B. Course Learning Outcomes (CLO) covered

At the end of this assessment, students are able to:

CLO 1 Demonstrate familiarity with fundamentals for processing massive datasets.

CLO 2 Evaluate various algorithmic design techniques covered in the syllabus to process massive datasets.

CLO 3 Apply the learned techniques to design efficient algorithms for massive industrial; datasets.

CLO 4 Solve certain algorithmic problems for a given dataset using appropriate software.

C. University Policy on Academic Misconduct

1. Academic misconduct is a serious offense in Xiamen University Malaysia. It can be defined as any of the following:
   i. **Plagiarism** is submitting or presenting someone else's work, words, ideas, data or information as your own intentionally or unintentionally. This includes incorporating

published and unpublished material, whether in manuscript, printed or electronic form into your work without acknowledging the source (the person and the work).

ii. **Collusion** is two or more people collaborating on a piece of work (in part or whole) which is intended to be wholly individual and passed it off as own individual work.

iii. **Cheating** is an act of dishonesty or fraud in order to gain an unfair advantage in an assessment. This includes using or attempting to use, or assisting another to use materials that are prohibited or inappropriate, commissioning work from a third party, falsifying data, or breaching any examination rules.

2. All assessments submitted must be the student's own work, without any materials generated by AI tools, including direct copying and pasting of text or paraphrasing. Any form of academic misconduct, including using prohibited materials or inappropriate assistance, is a serious offense and will result in a zero mark for the entire assessment or part of it. If there is more than one guilty party, such as in case of collusion, all parties involved will receive the same penalty.

D. Instruction to Students

a) This is an in-class test assessment. Each student must work independently and apply appropriate data analysis techniques to complete the test **within class hours**.

b) Submission Requirements: Please upload the following materials to **Moodle**:

   a. **Cover Page:** Include your name, student ID, course title, and assignment title.

   b. **Jupyter Notebook File (.ipynb):**

      i. Your notebook should contain all code and analysis (in Python comments or Markdown text format).

      ii. Save the file as: StudentID.ipynb

c) **Submission Deadline:** Submit all required files by **13 Nov 2025, 9:00 (GMT+8)**. Late submissions will **not** be accepted.

E. Evaluation Breakdown

No.	Component Title	Mark
1.	Data Exploration and Data Cleaning	20
2.	Clustering	40
3.	Cluster Behaviour Analysis	40
	TOTAL	**100**

F. Task(s)

Part 1: Data Exploration and Data Cleaning

You are given a dataset, "**CC GENERAL.csv**", and you are expected to perform the following tasks:

Explore the data and fix any observed issues such as abnormal values, missing values, incorrect data type, etc, using relevant Pandas built-in functions.

a) Record your observations in Markdown cells or Python comments (e.g., "Column X stored as strings while it is supposed to be a float" or "Column Y has about 10% of missing values", etc). Your record should also note when no issues are found, for example, stating that all fields are fully populated and no missing values were detected for that specific data-quality check (if this is the case).

b) For every issue found, apply appropriate fixes, such as correcting data types, handling missing values, or treating outliers wherever applicable.

[20 Marks]

Part 2: Clustering

Apply any **two clustering algorithms**, such as K-Means/Hierarchical Clustering/DBScan, to segment customers based on their card usage behaviour. Determine the optimal number of clusters for each chosen algorithm using a relevant method. (If you use DBScan, you need to determine an appropriate epsilon and min samples parameter value)

Note: *Before clustering, you must also check if feature scaling is required. If so, apply the appropriate method to standardize the data.*

Write your rationale for choosing a specific number of clusters or your choice of parameter for epsilon & min sample in Python comments/Markdown cell.

[40 Marks]

Part 3: Cluster Behaviour Analysis

Interpret the characteristics of each cluster by comparing their feature values (e.g cash advance vs credit limit) with the help of appropriate charts. Record all observations of the cluster behavior in Markdown cells or Python comments as below:

> ***Cluster 1:*** *This group has people with high credit limits who take more cash in advance*

Note:

a) You are only required to pick **ONE** of the clustering results from Part 2 to perform cluster behaviour analysis
b) You need to give **FIVE** observations of the cluster behaviour.

[40 Marks]

APPENDIX 1

MARKING RUBRICS

Component Title	Assignment 1				Mark	100	
Criteria	Score and Descriptors					Weight	Marks
	Excellent (17-20)	Good (13-16)	Average (9-12)	Need Improvement (5-8)	Poor (0-4)		
Data Exploration and Data Cleaning							
Part 1	Excellent use of Pandas functions with detailed observations on all columns; correctly identifies and fixes data issues with clear explanations.	Good use of Pandas functions with minor omissions; most data issues identified and reasonably fixed.	Basic use of Pandas functions; only partial data issues identified or fixed.	Limited inspection; few functions used, minimal fixes, or unclear explanation.	Little to no inspection or fixes	**20**	
Clustering							
	Excellent (32-40)	**Good (24-31)**	**Average (16-23)**	**Need Improvement (8-15)**	**Poor (0-7)**		
Part 2	Correctly applies 2 algorithms, uses proper scaling, justifies parameters clearly, and produces accurate, well-explained results	Applies 2 algorithms correctly with reasonable parameter justification; minor gaps in explanation or scaling	Clustering done but with limited justification, partial errors, or basic explanations	Significant mistakes in clustering steps, weak or missing justification, unclear results	Clustering incorrect, incomplete, or missing	**40**	
Cluster Behaviour Analysis							
	Excellent (32-40)	**Good (24-31)**	**Average (16-23)**	**Need Improvement (8-15)**	**Poor (0-7)**		
Part 3	Five clear, insightful observations supported with suitable charts	Five reasonable observations with charts; mostly correct but less detailed.	Observations are basic, partially correct, or weakly supported	Few or unclear observations; charts missing or poorly used	Behaviour analysis mostly incorrect, incomplete, or missing	**40**	
					TOTAL	**100**	