NGEE ANN POLYTECHNIC

School of InfoComm Technology

Applied Analytics Assignment

Diploma in Cybersecurity & Digital Forensics
Diploma in Data Science
Diploma in Information Technology
Year 2/3 (2023/2024), Semester 3/5

INDIVIDUAL ASSIGNMENT 1

(30% of Applied Analytics Module)

Deadline for Submission:

10th Jun 2023 (Saturday), 23:59 HRS

Tutorial Group:	
Student Name:	
Student Number:	

Penalty for late submission:

10% of the marks will be deducted every day after the deadline.
NO submission will be accepted after 17th Jun 2023, 23:59.

1 Problem Statement

1.1 Objective

Sometimes, we have a group of observations and we need to split it into a number of subsets of similar observations. **Cluster analysis** is a group of techniques that will help you to discover these similarities between observations.

You are required to use the provided dataset and split it into smaller subsets (clusters) based on similar characteristics. You will utilise related Python Libraries (through Jupyter Notebook platform) to do this. You are also required to create cluster visualisations to help users explore the clustered data. You can use Python Libraries (e.g. matplotlib) or other suitable visualisation tools to do this. Finally, you are required to summarise and interpret the formed clusters.

Your task is to analyze the used car data and provide insights on key factors affecting sales, pricing, and demand for used cars in the market, and used the result to advice the used car reseller how they can leverage on the result to improve revenue. for example, to improve on marketing strategy to become more customer-centric, streamline purchasing of used car, pricing strategy etc

1.2 Dataset

The dataset contains 2059 information from a research company on used car profile based on past transaction. Each row has a set of features including Make, Model, Price, Year, Kilometer, Fuel Type, Transmission, Color, Owner, Seller Type, Engine (Cc), Max Power, Max Torque, Drivetrain, Length, Width, Height, Seating Capacity and Fuel Tank Capacity (a total of 19 columns). Please refer to the dataset (Used_Car_Data.csv) for more details.

1.3 Suggested Tasks

You are suggested to tackle this problem in 3 steps:

Step 1 – Clustering and visualisation of numerical data

- Perform simple data exploration to familiarise yourself with the dataset. You may perform descriptive analysis in Python using info(), describe(), histograms, boxplots etc. Are there any missing values or outliers. How did you handle these?
- Perform simple data manipulation on numerical data to prepare the data for clustering modeling. Do you need to scale the data or not? Did you ignore any numerical variables? Why?
- Build clustering models using numerical data only.
    - Build **TWO** different models: one is using K-means clustering technique and the other is using Hierarchical clustering technique
    - For both models:

- Present your choice for the optimal number of clusters. How did you arrive at this number?
- Evaluate the models using proper metrics, e.g. Sum of Squared Errors (SSE), Silhouette Scores etc.
- Analyse the formed clusters using proper visualisation tools (e.g. Matplotlib etc.)

- Compare the K-means clustering model & Hierarchical clustering model and select the preferred Model for the further exploration.

Step 2 – Summarise, Interpret and Reflect

- Summarise your findings in a table and provide names for clusters.
- Provide an interpretation of each cluster.
- Suggest possible further improvement(s). Reflect on the skills learnt and the skills you could have learnt better.

1.4 Suggested Report Format & Content Guidelines

Based on the above, write an **INDIVIDUAL** report with the following sections (see Table below). Sample content description is provided for each section. You are free to include other relevant information you deem necessary in the sections. You are strongly advised to use screenshots to capture details of work done.

*(Note: For a page with 1 inch margins, 11 point Calibri font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Table of Contents	NA
2.	Summary/Overview	500 words

3.	Build Clustering Models using Numerical Data
	• Data exploration and manipulation on numerical data
	• Build TWO different clustering models:
	  ○ K-means Clustering
	  ○ Hierarchical Clustering
	• Clustering Analysis using numerical data
	• Evaluate and Compare the Models
		Min: 1000 words
		Max: 2000 words

4.	Summary and Interpretation • Summarize your findings in a table and provide names for clusters • Provide an interpretation of each cluster	Min: 500 words Max: 1000 words

	Suggested Report Sections & Content Guidelines	Word Count
6.	Reflection • Suggest possible further improvement(s) to the current solution. • With reference to the module learning objectives stated, reflect on the skills learnt and the skills you could have learnt better.	Min: 500 words Max: 1000 words

2 Presentation and Demonstration

Each student will be required to submit a video recorded presentation to showcase and demo the work. The video recorded presentation should be not exceed 10 minutes. Video recorded presentations which exceed the allotted time will be penalized.

3 Deliverables

For this assignment, you must submit all the following:

1. **Recorded Video Presentation**
   - You are required to submit a video recorded presentation to showcase and demo your work using your Jupyter Notebook.
   - During the recording, your webcam must be turn on, clearly showing your face, for authentication.
   - The video recorded presentation should be not exceed 10 minutes.
   - Video recorded presentations which exceed the allotted time will be penalized.
   - You must record your video presentation using BrightSpace "AA_Assignment1 - VideoPresentation", the link can be found in Brightspace under Content/Assignment 1.
   - Deadline for the video submission is **Sat 10th Jun 2023, 2359 hours**

2. The **completed "AA_Assignment1_<*name*>.ipynb"** Jupyter Notebook File in POLITEMALL
   - This is the Jupyter Notebook which you use to conduct your recorded video presentation.
   - Deadline for Jupyter Notebook submission is **Sat 10th Jun 2023, 2359 hours**

3. A softcopy **Final Report** in POLITEMALL
   - Deadline for report submission is **Sat 10th Jun 2023, 2359 hours**

     Note: DO NOT PLAGIARIZE (please refer to Ngee Ann Polytechnic Plagiarism Policy webpage for more information)

4 Grading Criteria

	Grading Criteria	Component Weightage
Presentation	a) Quality of work b) Flow of presentation based on content guidelines (see section 1.4) c) Quality of presentation material d) Presentation and articulation skills	**30%**
Final Report	a) Quality of work b) Completeness of report based on suggested report sections and content guidelines (see section 1.4) c) Clarity of report, use of proper visual aids and use of proper grammar d) Quality of recommendations for further improvements	**70%**