CAB330: Assignment 2

Task Descriptions and Requirements

The purpose of this assignment is to help you understand how the methods learned in this unit can be applied to different types of datasets, including structured record data, transactional data, unstructured text data, and weblog data. The assignment consists of four tasks: **Clustering Analysis**, **Association Analysis**, **Text Analysis**, and **Weblog Analysis**. You need to use Python, along with any libraries covered so far, to answer the questions by presenting your response and findings.

Task 1: Clustering Analysis

The dataset "*Dataset1_AudioTrack.csv*" is from a music shop, which stores technical details of all audio tracks that it has been producing. The dataset includes various audio track and their features. Each row represents an audio track defined by its eight attributes. Detail of the dataset is given below:

Attribute	**Data Type**	**Description**
ID	String	The ID for the track
Name	String	Name of the track
Energy	Float	Energy is a perceptual measure of intensity and activity and is recorded in the range of 0.0 to 1.0. Typically, energetic tracks (with high values of this variable) feel fast, loud, and noisy. For example, if the audio track has a high energy, then it could indicate a "death metal" while a low value on the scale could indicate "bach prelude". Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Loudness	Float	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Typically tracks have this attribute values ranged between -60 and 0 dB.
Speechiness	Float	It detects the presence of spoken words in a track. The more exclusively speech-like a recording is (e.g., talk show, audio book, poetry), this attribute value is closer to 1.0. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Instrumentalness	Float	It indicates whether a track contains vocals or not. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
Type	String	It denotes the object type of the track.
Time_signature	Int	It is a notational convention used in Western music to specify how many beats are contained. The range is from 0 – 5.

The manager of the music shop plans to segment the audio tracks based on eight key attributes and requested help from ChatGPT, which provided an idea by choosing a clustering method to determine the optimal number of clusters and make each cluster represents a group of similar audio tracks.

Your task is to implement the idea by conducting ***k*-means clustering** on this dataset and find and describe the optimal **number of effective clusters**.

Answer the following seven questions in relation to this dataset and your clustering analysis.

1) Identify data quality issues in "*Dataset1_AudioTrack*.csv" file such as unusual data types, missing values and others. Describe your data cleaning approach and clean the dataset.
2) Which variables were included in your analysis? Justify your selection. Then, check whether all feature values fall within the required ranges, report their maximum and minimum values, and plot the distribution of each feature.
3) Build a default clustering model with $k = 3$ and answer the following sub-questions:
    a) How many data points (audio tracks) are assigned into each cluster?
    b) Plot the cluster distribution using a "pairplot". Describe the key characteristics of each cluster by summarizing their mean and range values in a table (as shown below).

Feature	**Cluster0_mean_range**	**Cluster1_mean_range**	**Cluster2_mean_range**
Energy	0.357 [0.014, 0.868]	…	…
Loudness	…	-24.47 [-39.97, 19.47]	…
…	…	…	…

4) Determine the optimal number of clusters (k) for the selected features using the Elbow Method and Silhouette Score.
5) Apply standardization and normalization to the optimal k-means model, and determine which method for scaling the features yields better clustering performance? Hint: Use StandardScaler and MinMaxScaler from the *sklearn* package to evaluate the impact of different scaling methods.
6) Regarding the better feature scaling method above, display the distributions of each feature in each cluster (Hint: you may use functions *histplot, kdeplot, etc.*).
7) Regarding the better feature scaling method above, describe what characterizes each cluster in plain English, and outline your interpretation approach step by step.

Task 2: Association Analysis

A supermarket is interested in identifying associations between items purchased by its customers. To do this, the store has decided to perform an association analysis using the dataset "Dataset2_TRANS.csv". Details of the dataset are provided below:

Attribute	Description
LOCATION	Point of sale device identification number. The range is from 1 - 10.
TRANSACTION_ID	Unique transaction identification number for a given sale. A sale may include several products and thus the same transaction id may occur over several rows.
TRANSACTION_DATE	Date of transaction
PRODUCT_NAME	Product Purchased
QUANTITY	Quantity of this product purchased (always set to 1 by a point-of-sale device)

Your task is to perform an association analysis on this dataset and answer the following seven questions based on the dataset and your analysis.

1) Identify data quality issues in the 'Dataset2_TRANS.csv' file, such as unusual data types, missing values, and other inconsistencies. Describe your data cleaning approach and clean the dataset accordingly.
2) Which variables (features) were included in your analysis? Justify your selection. Then, check if you need to convert some features' data type for preparing the transaction data.
3) Conduct association rule mining and display all rules with a minimum support of 0.015 and a minimum confidence of 0.12.
4) Identify the rule with the highest confidence and the rule with the highest lift from the resulting association rules.
5) Plot the confidence, support, and lift (for rules where lift ≥ 1.4) of the resulting association rules.
6) Find association rules that reveal which items are commonly purchased along with 'Shampoo'.
7) This association analysis helps the supermarket discover relationships between items that are frequently bought together. Provide two different applications or examples to illustrate how the store can benefit from your analysis and specify which rules are applied.

Task 3: Text Analysis

A cinema is planning to launch an online movie recommendation service for its customers. To support this initiative, it has compiled a metadata dataset, "Dataset3_Movie.csv", which includes key information about various movies.

In the context of movies, a *genre* refers to a category that groups films based on common characteristics such as narrative structure, subject matter, mood, or style. In the collected dataset, the **genre** column includes the following values: *Documentary*, *SciFi*, *Romance*, and *Kids&Family*.

The cinema aims to enhance its recommendation system by using an AI-based method to identify more specific sub-categories within each genre. This task involves performing text analysis on the movie descriptions to discover clusters of movies that share similar topics within each genre.

Details of the dataset are provided below:

Attribute	Description
Cast1, Cast2, Cast3, Cast4, Cast5 and Cast6	The group of popular actors/actresses who acted in the movie
Description	This provides a short synopsis of the movie
Director 1, Director 2, Director 3	The list of directors for this movie. If it is directed by only one director then Director 2 and Director 3 will have "Director Not available".
Genre	A movie genre is a motion picture category based on similarities in either the narrative elements or the emotional response to the film. It has values like Documentary, SciFi, Romance, and Kids&Family.
Rating	Using the Motion Picture Association of America (MPAA) film rating system, each movie is rated for its suitability for certain audiences based on its content. It includes G, NC17, NR, PG, PG-13 and R
Release Date	The date of release for the movie.
Runtime	Runtime is the time between the starting of the movie up to the end of the credits scene.
Studio	The facility that was used to make that movie.
Title	Title of the movie
Writer1, Writer2, Writer 3, Writer 4	A list of screenplay writers or the scriptwriters or scenarists who has written the screenplay for this movie.
Year	The Year the movie was released

Answer the following six questions in relation to this dataset and your text analysis.

1) Data Exploration: Display the value of the *Description* field for the 24th row (index=23); calculate the average length (in words) of the texts in the *Description* column; and report the number of rows for each category in the *Genre* column.
2) Define a text preprocessing function that splits an input document (string) into tokens and processes them by removing unnecessary characters, stopwords, digits, punctuation, and short tokens (i.e., tokens with length less than 3). Test the function by applying it to the "Description" column to transform each entry into tokens; save the processed tokens in a new column named "Parsed_description"; and display the value of "Parsed_description" for the 24th row (index=23).
3) Define a vectorization function that converts the "Parsed_description" column into a matrix of token counts using a suitable vectorizer. The function should return a list of terms, where each term is represented in the format:{'term': ..., 'idx': ..., 'tf': ..., 'df': ...}
    - term: the token (word)
    - idx: the index/position in the vocabulary
    - tf: total term frequency across all rows
    - df: document frequency

    Then, visualize (plot) the frequency distribution of terms following Zipf's Law.
4) For each "Genre" category (*Documentary*, *SciFi*, *Romance*, *Kids&Family*), extract its corresponding sub-dataframe (Pandas DataFrame) and store them in a list named sub_data, such that:
    - sub_data[0] corresponds to "Documentary"
    - sub_data[1] corresponds to "SciFi"
    - sub_data[2] corresponds to "Romance"
    - sub_data[3] corresponds to "Kids&Family"

    Then, for each genre category, visualize (plot) the frequency distribution of terms according to Zipf's Law.

5) Generate 5 clusters for each genre category (resulting in a total of 20 clusters across all 4 categories). For each cluster, visualize it by displaying the top 12 most representative terms that characterize the cluster.
6) Assign a meaningful sub-category name to each cluster within all categories. Clearly explain the method or rationale you used to determine each cluster name.

Task 4: Web Log Analysis

An e-commerce company has collected a web server logfile, "Dataset4_weblog.txt", which records all registered events. A dash ('-') in any field indicates missing data. The log file contains information about website visitors, their behavior, and crawler activity.

The goal of this task is to gain insights (or patterns) into user visits, where a "visit" (or a session) is defined as a continuous period of activity by the same user on the website, grouped together from multiple log entries.

The following table shows the possible variables (attributes) used in the logfile.

Attribute	**Data Type**	**Description**
IP address	String	Client's IP address
Remote Log	String	Remote name of the User performing the request.
User ID	String	The ID of the user making the request.
Timestamp	Date	The date and time of the request are represented in UTC format as follows: Day/Month/Year: Hour:Minutes:Seconds +Time-Zone-Correction.
Request Type	Categorical	The type of request (GET, POST, PUT, DELETE) that the server received.
Path	String	The path of the website related to the request.
Protocol	String	Protocol used for connecting to the server and its version.
Status	Integer	The server returns a status code for each request.
Response size	Integer	The data in bytes that was sent back to the client.
Referrer	String	The source of the user's referral to the current website, represented as "-" if there's none.
User agent	String	The user agent string provides details of the browser and host device, such as the name, version, and device type.

Your task is to pre-process the given dataset (a text file) and apply an appropriate data mining technique to the raw log data. Then, answer the following five questions based on the dataset and the analyses you have conducted.

1) Pre-process the weblog data to identify useful attributes (or variables) from the weblogs.txt file, such as *IP Address*, *Remote Log Name*, *User ID*, *Timestamp*, Request *Type*, *Path*, *Protocol*, *Status*, *Size*, *Referrer*, and *User-Agent*. Then, save the parsed logs into a Pandas DataFrame.
2) Select useful features for grouping logs into sessions, where each session is identified by a session key in the format "*IP + Timestamp*". Remove entries in the *Path* column

that reference graphics or audio files, and report how many rows and columns are removed after completing this question.

3) Select a suitable data mining method to the processed dataset. Explain the rationale for selecting this method.
4) For the selected data mining method, describe its goal, and outline the corresponding steps of your data analysis as implemented in your Python code.
5) Discuss the results obtained, and the applicability of your findings. You should include only a high-level managerial kind of discussion on the findings. It should be more than just an interpretation of the method output.