Q 1(a) [12 Marks]

Using the topic from your CA682 visualisation assignment, apply the Generic Data Analytics Pipeline to describe how the data may have been Gathered, Processed, Analysed, Presented and Preserved. Give a brief description of the activities at each stage (1-2 sentences) and identify any specific tools that you did or would use. If you didn't specifically perform any stage then you can make assumptions or predictions about the actions and tools.

If you didn't complete a visualisation assignment then write about a scenario based on analysing *student feedback to DCU in relation to online teaching in 2020*.

1. Gathering

Description: The data was retrieved from a public GitHub repository containing Spotify song attributes, including track metadata, audio features, and popularity metrics.

Tools Used:

GitHub: For accessing the CSV file containing the dataset.

Python (Pandas): To load the data into a DataFrame for exploration.

Activities: Downloading the CSV file, verifying its contents, and ensuring it was compatible for analysis.

2. Processing

Description: The dataset was cleaned and preprocessed to make it analysis-ready. Missing values were removed, unrealistic thresholds were applied (e.g., filtering songs based on tempo and duration), and features were normalized.

Tools Used:

Python (Pandas): For data cleaning and manipulation.

Scikit-learn (MinMaxScaler): To normalize features like tempo, energy, and danceability.

Activities:

Removal of missing or irrelevant rows and columns.

Outlier detection and removal.

Normalization to ensure consistent feature scaling.

3. Analysis

Description: The data was analyzed to uncover correlations between audio features and song popularity. Trends across time and genres were examined to identify patterns in musical preferences.

Tools Used:

Python (Matplotlib, Seaborn): For creating exploratory visualizations like bar charts and scatter plots.

Descriptive Statistics: To summarize data properties.

Activities:

Correlation analysis to identify relationships between features (e.g., energy vs. popularity).

Summarizing trends in song attributes over time.

4. Presenting

Description: Visualizations were created to effectively communicate insights. A radar (spider) chart was chosen to display the multidimensional nature of the audio features, comparing recent songs to the entire dataset.

Tools Used:

Python (Matplotlib): For plotting the radar chart.

Canva: For prototyping and pre-designing the visualizations.

Activities:

Generating a static radar chart to compare recent tracks with the overall dataset.

Designing visually appealing and clear representations of trends.

Preserving

Description: The cleaned dataset and visualizations were saved for reproducibility and future analysis. Documentation of steps taken was provided to ensure clarity and ease of reusability.

Tools Used:

Python (Pandas): To save the cleaned data into a CSV file.

GitHub: For storing the processed data and code.

Activities:

Saving cleaned and normalized data to disk.

Documenting preprocessing steps and analysis in a Jupyter Notebook.

Archiving results and reports for future reference.

Gathering:

Activities: Information on stock prices, volume of trading and other related financial variables were obtained from [Alpha Vantage] through its (API). The data variables include historical and current stock data from Tesla's IPO and up to 2024.

Tools: API requests using Python `requests` library and data in JSON format received from the websites.

Processing:

Activities: The raw JSON data was migrated through an ETL process into a Relational Database Management System (RDBMS). Activities such as identifying and handling missing values, converting the data type to float, scaling the target variable closing prices for better comparison also came under data cleaning process. New columns created were `month_id`, `year_id`, and % change in closing price for better analysis blueprint.

Tools:

- For ETL: `SQLAlchemy`, MySQL.
- For processing: `pandas`, `ast`, and `json`.

Analysis:

Activities: Hence, normalized closing prices and trade volumes were used as variables, to match Tesla's ambitious growth progression to that of Amazon, Apple, IBM, and Microsoft. One of the reasons for normalizing the results was to compare the organizations with equal starting points with respect to price levels.

Tools: main data manipulation tool in this analysis was done with the help of `pandas`.

Presenting

Activities: Visualizations were created to highlight Tesla's unique growth, using a line chart to compare normalized stock performance over time. Design considerations included color-coding, subtle gridlines, and selective x-axis points for clarity.

Tools: `matplotlib`, `seaborn` libraries.

Preserving

Activities: The processed and visualized data, along with the analysis, was stored in a database for easy updates and reference. Incremental loading ensured that new data could be added without redundancy or system strain.

Tools: There was MySQL for data storage; Jupyter Notebook for documentation as well as reproducibility purposes.