DENNING
SAICS
SCHOOL FOR ARTIFICIAL INTELLIGENCE & COMPUTER SCIENCES

Techma Zone

Python DataQuest Competition: Day 4 - Task Breakdown

Objective:

Complete the three data manipulation tasks using Python tools such as Pandas, Matplotlib, and the CSV library. These tasks must be completed within a 2-hour timeframe, and teamwork is crucial for successful completion.

Task 1: Data Analysis and Visualization (Pandas + Matplotlib) – 5 Marks

Objective:

The dataset contains a date column that needs to be converted to a proper date format. Generate insights from the dataset by performing an analysis and visualizing the results.

Steps:

1. Convert the Date Column to Date Format:
- The column `dteday` in the dataset is in object format and needs to be converted to a `datetime` object for proper analysis.
Use the following code to perform the conversion:

```
df['dteday'] = pd.to_datetime(df['dteday'])
```

2. Data Analysis and Visualization:
- Analyze the dataset to generate meaningful insights. You can look for trends, distributions, correlations, etc.
- Use Matplotlib to visualize the findings. Some possible plots include:
  - Time-series plots showing changes over time.
  - Histograms or bar charts to show distributions of key variables.
  - Scatter plots for examining relationships between variables.
- Provide a report that summarizes the insights, with visualizations supporting the analysis.

Task 2: Data Cleaning and Transformation (Pandas) – 5 Marks

Objective:

Clean up certain columns by removing unwanted characters and converting columns into appropriate formats. Calculate the age of each artist based on their birthdate and submission date.

Steps:

1. Clean Up Nationality and Gender Columns:
- Remove unwanted characters such as parentheses from the `Nationality` and `Gender` columns.

2. Clean and Convert `BeginData` and `EndDate` Columns:
- Remove unwanted characters like "-" from the `BeginData` and `EndDate` columns.
- Convert these columns to `int` data type. If the data is missing or contains bad characters, handle these cases accordingly.
- Calculate the actual age for each artist by subtracting their `BeginData` from their `EndDate`. If no data is available, input a placeholder value such as "-" in the `actual_age` column.

3. Clean Date Column (Submission Date of Art):
- The `Date` column may have bad characters that need to be cleaned before proceeding with any calculations.

Task 3: Data Handling with CSV (Python's `csv` Library) – 10 Marks

Objective:

Work with a CSV file containing account transactions and perform specific data cleaning tasks.

Steps:

1. Extract Account Names Against Each Transaction Record:
- Open and read the CSV file using the `csv` library. Ensure that each transaction record includes the corresponding account name.
- Extract account names and their associated transactions.

2. Remove Total Rows for Accounts:
- The dataset includes rows with totals for accounts. These rows should be removed from the dataset.
- Filter out these rows by identifying them using specific criteria (e.g., rows containing the word "Total" or other distinctive features).

3. Remove Rows with "Beginning Balance" in the Date Column:

- If there are rows where the `Date` column has "Beginning Balance," remove those rows from the dataset as they are not relevant to the analysis.
- Filter out these rows based on this condition.

Collaboration and Task Management:

- Teamwork: The team must divide the tasks based on skillset and time constraints. It's essential that all team members collaborate efficiently to finish the tasks within the given 2-hour timeframe.