Data Manipulations with Pandas II

Task 1

Given a DataFrame with multiple categorical columns, find all unique combinations of values from these columns that occur at least 10 times. This Task involves:

- Identifying combinations across several columns
- Counting occurrences efficiently
- Ensuring optimal performance for large datasets.

Task 2

You are provided with a DataFrame containing missing data across many columns. Your task is to:

- Identify the top 3 columns that have the highest percentage of missing values.
- Remove rows with missing values in those top 3 columns, but leave the rest of the DataFrame intact.

Task 3

Given a DataFrame of 100,000 rows with mixed data types, find the most frequent value for each column without using the mode function.

- Focus on alternative Pandas built-in methods.

Task 4

In a DataFrame containing employee work hours (with columns 'Employee ID', 'Date', and 'Hours Worked'), identify employees who have worked for 5 consecutive days.

- Consider time series data handling
- Properly analyze date intervals to ensure accuracy.

Task 5

You have a time-series DataFrame indexed by dates. Your task is to:

- Resample the data to monthly intervals

- Focus only on the first quarter of each year (Jan, Feb, Mar)

- Sum the values for each month in the first quarter.

Task 6

Given a sales transactions DataFrame, remove the top 5% of transactions by value. Ensure that:

- The remaining data keeps its original index positions

- You properly exclude rows based on percentile thresholds without sorting the entire DataFrame.

Task 7

Create a moving average calculation for a time series that has irregular time intervals.

- Ensure the rolling window takes into account the gaps in time

- Handle any missing values efficiently.

Task 8

Given a DataFrame of financial transactions, detect when the cumulative sum of previous transactions exceeds a given threshold for the first time.

- Focus on an efficient way to track this threshold without iterating row by row.

Task 9

With a DataFrame containing a 'Category' column and a 'Value' column, your task is to:

- Group by 'Category'

- Find the rows where the 'Value' is the maximum for each category.

- Avoid using lambda or apply functions.

Task 10

You have a DataFrame where some columns contain nested JSON-like data (dictionaries). Your task is to:

- Flatten the DataFrame by creating new columns for each key in the nested data

- Ensure the original data structure remains intact while expanding the JSON fields.