Dataset Cleaning and Analysis

21/03/2025

stratified_history.csv
8.1 MB

Analyze the provided dataset to identify its key characteristics for data-driven decision-making. Upon receiving a specific analytical question, address it strictly using the dataset. Your deliverables must include a clear, concise answer to the question and the Python code that extracts and processes the data used to derive your answer. The dataset should be cleaned by filtering admission_period to include only values between 2006.1 and 2019.2, removing entries where enrollment_type is labeled as "Exempted", and excluding those where situation is "In Progress". The resulting cleaned dataset should be stored in a new DataFrame named df_filtered.

qwen-max-latest

To analyze the provided dataset (`stratified_history.csv`) and address your requirements, I will follow these steps:

1. **Load the Dataset** : Read the CSV file into a Pandas DataFrame.
2. **Clean the Dataset** :
   - Filter `admission_period` to include only values between 2006.1 and 2019.2.
   - Remove rows where `enrollment_type` is labeled as "Exempted".
   - Exclude rows where `situation` is "In Progress".
3. **Store the Cleaned Data** : Save the cleaned dataset in a new DataFrame named `df_filtered` .
4. **Provide Python Code** : Include the code used for cleaning and processing the data.

Below is the Python code to perform these steps:

```python
import pandas as pd

# Step 1: Load the dataset
df = pd.read_csv('stratified_history.csv')

# Step 2: Clean the dataset
# Filter admission_period to include only values between 2006.1 and 2019.2
df_filtered = df[(df['admission_period'] >= 2006.1) & (df['admission_period'] <= 2019.2)]

# Remove rows where enrollment_type is "Exempted"
df_filtered = df_filtered[df_filtered['enrollment_type'] != 'Exempted']

# Exclude rows where situation is "In Progress"
df_filtered = df_filtered[df_filtered['situation'] != 'In Progress']

# Step 3: Display the cleaned dataset (optional, for verification)
print(df_filtered.head())

# Step 4: Save the cleaned dataset to a new CSV file (optional, if needed)
# df_filtered.to_csv('cleaned_stratified_history.csv', index=False)
```

Explanation of the Code:

1. **Loading the Dataset** : The `pd.read_csv()` function reads the CSV file into a Pandas DataFrame.
2. **Filtering** `admission_period` : The condition `(df['admission_period'] >= 2006.1) & (df['admission_period'] <= 2019.2)` ensures that only rows with `admission_period` values within the specified range are retained.
3. **Removing "Exempted" Enrollment Types** : The condition `df_filtered['enrollment_type'] != 'Exempted'` removes rows where the `enrollment_type` column has the value "Exempted".
4. **Excluding "In Progress" Situations** : The condition `df_filtered['situation'] != 'In Progress'` excludes rows where the `situation` column has the value "In Progress".
5. **Resulting DataFrame** : The cleaned dataset is stored in `df_filtered` .

Key Characteristics of the Cleaned Dataset:

- The dataset now contains only relevant rows based on the specified filters.
- It excludes irrelevant or incomplete data (e.g., "Exempted" enrollment types and "In Progress" situations).
- The filtered dataset can be used for further analysis or decision-making.

If you have a specific analytical question to address using this cleaned dataset, please provide it, and I will derive the answer using `df_filtered`.

copyright © 2025