

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.

I'll first inspect the dataset to understand its structure and then apply the necessary cleaning steps based on your criteria. Let's start by loading and previewing the data.

It looks like there was a timeout issue while processing the file. I'll attempt to reload the dataset and preview its structure again. [↻]

It looks like the session reset, which removed previously loaded libraries. I'll reload the necessary libraries and retry the process. [↻]

The dataset contains 30,130 entries with 34 columns, covering aspects like admission details, dropout information, student demographics, and academic performance.

Data Cleaning Steps:

1. Filter admission_period: Keep only values between 2006.1 and 2019.2.
2. Remove enrollment_type labeled as "Exempted".
3. Exclude rows where situation is "In Progress".

I'll now apply these filters and store the cleaned data in df_filtered. [↻]

After applying the cleaning steps, the dataset now contains 13,831 entries. Let me know the specific analytical question you'd like to address, and I'll provide both the answer and the Python code used to derive it. [↻]

Determine the statistical correlation (Pearson or Spearman) between students' overall average and their enrollment period (first vs. second semester). Should correlation coefficients be consistent across different years, what factors might explain this pattern?

I'll compute both Pearson (linear correlation) and Spearman (rank-based correlation) coefficients between general_average (students' overall average) and whether they enrolled in the first (X.1) or second (X.2) semester of the year. Then, I'll check if the correlation is consistent across different admission years.

Let's proceed with the calculations.

It looks like there are NaN (missing) or infinite values in the general_average column, which prevents the correlation calculation. I'll clean these values and retry the analysis. [↻]

It seems like I can't do more advanced data analysis right now. Please try again later. However, you can run the following Python code on your local machine to compute the correlation yourself.

```
python
import pandas as pd
from scipy.stats import pearsonr, spearmanr

# Load the dataset
df = pd.read_csv("stratified_history.csv")

# Apply the cleaning steps
df_filtered = df[
    (df["admission_period"] >= 2006.1) & (df["admission_period"] <= 2019.2) &
    (df["enrollment_type"] != "Exemption") &
    (df["situation"] != "In Progress")
].copy()

# Handle missing values
df_filtered = df_filtered.dropna(subset=["general_average"])

# Extract semester (1 = First Semester, 2 = Second Semester)
df_filtered["semester"] = df_filtered["admission_period"] % 1

# Compute Pearson and Spearman correlations
pearson_corr, _ = pearsonr(df_filtered["semester"], df_filtered["general_average"])
spearman_corr, _ = spearmanr(df_filtered["semester"], df_filtered["general_average"])

# Compute correlation for each admission year separately
yearly_correlations = df_filtered.groupby(df_filtered["admission_period"].astype(int)).apply(
    lambda group: pd.Series({
        "pearson_corr": pearsonr(group["semester"], group["general_average"])[0] if len(group) > 1 else None,
        "spearman_corr": spearmanr(group["semester"], group["general_average"])[0] if len(group) > 1 else None
    })
)

print("Pearson Correlation:", pearson_corr)
print("Spearman Correlation:", spearman_corr)
print(yearly_correlations)
```

This will compute the overall correlation as well as yearly trends. Let me know if you need further help! [↻]

Envie uma mensagem para o ChatGPT

Anexar

Buscar

Refletir

Voz

O ChatGPT pode cometer erros. Considere verificar informações importantes.

?