



Entrar

Cadastrar

Esta é uma cópia de uma conversa entre ChatGPT e Andréino.  
Denunciar conteúdo

Arquivo carregado

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 load the dataset, inspect its structure, and then apply the specified cleaning steps:

1. Filter admission\_period to include only values between 2006.1 and 2019.2.
2. Remove rows where enrollment\_type is "Exempted".
3. Exclude rows where situation is "In Progress".

Let me start by checking the dataset.

The dataset contains 30,130 rows and 34 columns. The relevant columns for filtering are:

- admission\_period (float, needs to be between 2006.1 and 2019.2).
- enrollment\_type (categorical, remove "Exempted").
- situation (categorical, remove "In Progress").

I'll now apply these filters and store the cleaned data in df\_filtered.

The dataset has been successfully filtered, resulting in 16,563 rows.

Now, you can provide a specific analytical question, and I will address it strictly using this cleaned dataset (df\_filtered).

Assess the temporal stability of student grades by evaluating statistical variability (e.g., using trend analysis and variability metrics) across the dataset's recorded periods

To evaluate the temporal stability of student grades, I'll perform the following analyses:

1. Trend Analysis: Examine the average final\_average across different admission\_periods to identify potential trends.
2. Variability Metrics: Compute the standard deviation and coefficient of variation for grades over time to assess stability.
3. Visualization: Plot the trends and variability for a clearer interpretation.

I'll proceed with the analysis now.

It looks like there's an issue with missing or non-numeric values in the dataset. I'll clean the data further and reattempt the analysis.

The error suggests that there may still be unexpected data types in admission\_period or final\_average. I'll inspect these columns and ensure they contain only numeric values before proceeding.

Both admission\_period and final\_average are correctly formatted as numeric (float64). I'll now retry the analysis, ensuring all operations handle potential NaN values properly.

There's still an issue with admission\_period or final\_average. I'll investigate further by checking for any unexpected values or data types.

Both admission\_period and final\_average contain only numeric values. The issue might be caused by NaN values propagating through calculations. I'll inspect the dataset for any remaining anomalies before retrying the plot.

There are no NaN or infinite values in the computed dataset. I'll attempt plotting again, ensuring robust handling of missing values.

The issue persists, indicating a deeper data inconsistency. I'll investigate by checking the admission\_period column for any anomalies, such as incorrect data types or unexpected values.

The admission\_period column appears correctly formatted as float64 and contains expected values within the specified range. The issue might be related to incompatible data types in plt.fill\_between(). I'll cast all numerical columns explicitly to float before retrying the plot.

The issue persists despite all numerical data being properly formatted. To isolate the problem, I'll inspect the lower\_bound and upper\_bound values to check for any unexpected values like NaN or infinities.

I have processed the dataset according to the specified cleaning criteria and conducted an initial analysis of the temporal stability of student grades. Here's a summary of the key findings:

1. Data Cleaning:
  - The dataset was filtered to include only admission periods between 2006.1 and 2019.2.
  - Entries where enrollment\_type was "Exempted" and where situation was "In Progress" were removed.
  - This resulted in a cleaned dataset with 16,563 rows, stored in df\_filtered.
2. Statistical Analysis:
  - I calculated the mean final grades for each admission period.
  - The standard deviation of grades was computed to assess variability.
  - A coefficient of variation (CV) was derived to measure relative stability over time.
3. Challenges Encountered:
  - While attempting to visualize the trend using matplotlib, an error occurred when filling the area between the mean  $\pm$  1 standard deviation.
  - I investigated potential causes such as NaN values, non-numeric data types, and infinite values but found no anomalies.
  - Despite ensuring all numerical data were properly formatted, the plotting error persists, indicating an issue with how the values are handled in matplotlib.

#### Next Steps:

- The mean and standard deviation values are correctly computed, but the visualization remains incomplete due to technical issues.
- If you would like, I can provide the computed mean grades and variability statistics in tabular form as an alternative until the visualization issue is resolved.
- Let me know if you would like any modifications or alternative analysis approaches!

Envie uma mensagem para o ChatGPT

 Anexar

 Enviar

O ChatGPT pode cometer erros. Considere verificar informações importantes.

