

## 3.2 - diag 2

17 de julho de 2024



historico\_estratificado.csv  
Planilha

Analise os dados e limpe as colunas da seguinte forma:  
período\_ingresso: de 2006.1 até 2019.2  
tipo\_matricula: remover "Dispensa"  
situacao: remover "Em Curso"  
o novo dataframe após a limpeza deve se chamar df\_filtrado



Primeiro, vou carregar e visualizar os dados do arquivo `historico_estratificado.csv` para entender melhor sua estrutura e conteúdo. Em seguida, realizarei as limpezas solicitadas nas colunas `período_ingresso`, `tipo_matricula` e `situacao`. Vamos começar!

Os dados foram carregados com sucesso e o DataFrame possui várias colunas. Agora, vou realizar as seguintes limpezas:

1. **Filtrar `período_ingresso`**: Manter registros de 2006.1 até 2019.2.
2. **Remover linhas em `tipo_matricula`** onde o valor é "Dispensa".
3. **Remover linhas em `situacao`** onde o valor é "Em Curso".

Vou aplicar esses filtros agora. [\[↗\]](#)

A limpeza dos dados foi realizada com sucesso. O novo DataFrame `df_filtrado` contém 13.831 registros, após aplicar os seguintes filtros:

- **Período de ingresso** entre 2006.1 e 2019.2.
- Remoção de registros onde **tipo de matrícula** era "Dispensa".
- Remoção de registros onde a **situação** era "Em Curso".

Se precisar de mais análises ou ajustes neste DataFrame, estou à disposição para ajudar! [\[↗\]](#)

Existe uma diferença estatisticamente significativa nas notas entre os alunos de matrícula Normal e Extra Curricular?



Para determinar se existe uma diferença estatisticamente significativa nas notas (representadas pela coluna `media_final`) entre os alunos de matrícula "Normal" e "Extra Curricular", podemos usar um teste estatístico. Um teste comum para comparar as médias de dois grupos independentes é o teste t de Student, desde que as premissas para sua aplicação sejam atendidas (normalidade dos dados e homogeneidade de variâncias).

Primeiro, vamos verificar a normalidade dos dados de cada grupo. Se os dados não forem normalmente distribuídos, poderíamos usar um teste não paramétrico, como o teste de Mann-Whitney.

Vamos começar verificando a normalidade e a homogeneidade das variâncias.

Os resultados dos testes indicam o seguinte:

1. **Teste de Shapiro-Wilk** para normalidade:
  - Grupo "Normal": p-valor = 0.0, indicando que as notas não são distribuídas normalmente.
  - Grupo "Extra Curricular": p-valor muito próximo de zero, indicando que as notas também não são distribuídas normalmente.
2. **Teste de Levene** para homogeneidade das variâncias:
  - p-valor  $\approx 5.60 \times 10^{-8}$ , que é menor que 0.05, indicando que as variâncias entre os grupos não são homogêneas.

Dado que ambas as premissas do teste t de Student não são atendidas (normalidade e homogeneidade de variâncias), é mais adequado usar um teste não paramétrico. Vamos realizar o

**teste de Mann-Whitney** para comparar as medianas das notas entre os dois grupos. Este teste não assume normalidade dos dados e é adequado para comparar duas amostras independentes. Vamos proceder com esse teste agora. [\[↗\]](#)

O resultado do teste de Mann-Whitney mostra um p-valor aproximadamente igual a  $2.80 \times 10^{-7}$ , que é significativamente menor que 0.05. Isso indica que há uma diferença estatisticamente significativa nas notas entre os alunos de matrícula "Normal" e "Extra Curricular".

Portanto, podemos concluir que as medianas das notas dos dois grupos diferem de maneira significativa. Se precisar de mais detalhes ou análises adicionais, estou à disposição! [\[↗\]](#)

Comece com Data Analyst

[Denunciar conteúdo](#) | [Termos de uso](#) | [Política de privacidade](#) |

?