UFRN - UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

Projeto de introdução à Ciência de Dados

Programa de Educação Tutorial do curso de Ciência da Computação da UFRN

Docente:

Célio Felipe Bezerra Santiago

NATAL-RN

2025

RESUMO EXECUTIVO

Esse projeto teve como objetivo analisar os dados das turmas da Universidade Federal do Rio Grande do Norte (UFRN) nos anos de 2023 e 2024, com foco em variáveis como capacidade, número de solicitações, local e modalidade (presencial ou a distância). A ideia foi aplicar técnicas de Machine Learning para entender melhor como essas turmas estão distribuídas e identificar padrões que possam ajudar no planejamento acadêmico.

Os principais insights mostram que a maioria das turmas são de graduação e estão concentradas no Campus Central. Também foi observada uma grande variação na carga horária entre diferentes níveis de ensino. O clustering permitiu agrupar as turmas em quatro perfis distintos, o que abre espaço para decisões mais direcionadas. Entre as recomendações estão: a otimização da alocação de recursos com base na demanda por modalidade, o uso dos perfis identificados para planejamento de horários e a investigação dos motivos por trás da exclusão de algumas turmas do sistema.

INTRODUÇÃO

A oferta de turmas é um dos pontos chaves da organização acadêmica em universidades. Pensando nisso, esse projeto teve como objetivo a análise de dados das turmas, buscando entender diversos fatores, como: capacidade, solicitações de turmas, entre outros fatores da instituição.

O projeto foi desenvolvido e coordenado pelo PET (Programa de Educação Tutorial) do curso de Ciência da Computação da UFRN.

METODOLOGIA

Os dados foram obtidos do portal Dados Abertos da UFRN, referente aos anos de 2023 e 2024, sendo a base final contendo cerca de 50.000 registros. As ferramentas utilizadas foram: Pandas, Numpy, Matplotlib, Seaborn, Sklearn, Scipy.stats e Google Colab.

As etapas do projeto incluíram:

1. Limpeza de Dados : Remoção de duplicatas, remoção de valores ausentes que foram tratados(remoção de colunas nulas e preenchimento de outros nulos com medianas, modas ou valores específicos).

2. Análise Exploratória: Visualização e estudo da distribuição de turmas por nível de ensino e campus, identificação de outliers, comparação de modalidades de capacitação versus solicitações e carga horária

3. Clustering: Seleção de variáveis relevantes, definição de k=4 pelo método cotovelo e aplicação do algoritmo K-Means, incorporando os rótulos originais.

4. Teste estatísticos: Uso de Anova, Mann-Whitney U e Kruskal-wallis para avaliar diferenças e validar hipóteses relacionadas à capacidade, solicitações e carga horária.

ANÁLISE E INSIGHTS

LIMPEZA

Antes de qualquer análise, é fundamental garantir a qualidade e integridade dos dados, eliminando informações irrelevantes ou redundantes que possam comprometer os resultados. Nos dados obtidos da UFRN, foram identificadas quatro colunas inteiras sem registros preenchidos — *observação*, *matricula_docente_externo*, *id_turma_agrupadora* e *convenio*. Como essas colunas não agregavam valor à análise e apenas ocupavam espaço com células vazias, optou-se por removê-las (*drop*). Observados nas Figuras 1 e 2.

id_turma	0
codigo_turma	0
siape	315
matricula_docente_externo	14570
observacao	14348
id_componente_curricular	0
ch_dedicada_periodo	0
nivel_ensino	0
campus_turma	3173
local	74
ano	0
periodo	0
data_inicio	0

Figura 1

data_fim	0
descricao_horario	383
total_solicitacoes	1326
capacidade_aluno	32
tipo	0
distancia	0
data_consolidacao	782
agrupadora	0
id_turma_agrupadora	14241
qtd_aulas_lancadas	782
situacao_turma	0
convenio	14885
modalidade_participantes	0

Figura 2

Colunas removidas constam nas figuras 3 e 4.

id_turma	0
codigo_turma	0
siape	315
id_componente_curricular	0
ch_dedicada_periodo	0
nivel_ensino	0
campus_turma	3173
local	74
ano	0
periodo	0

Figura 3

data_inicio	0
data_fim	0
descricao_horario	383
total_solicitacoes	1326
capacidade_aluno	32
tipo	0
distancia	0
data_consolidacao	782
agrupadora	0
qtd_aulas_lancadas	782
situacao_turma	0
modalidade_participantes	0

Figura 4

Além disso, verificou-se a existência de registros duplicados que poderiam distorcer conclusões e análises estatísticas. Esses registros também foram removidos, assegurando que o conjunto de dados final estivesse mais limpo, consistente e pronto para gerar insights confiáveis.

	count
False	14885
True	295

Figura 4

	count
False	14872

Figura 5

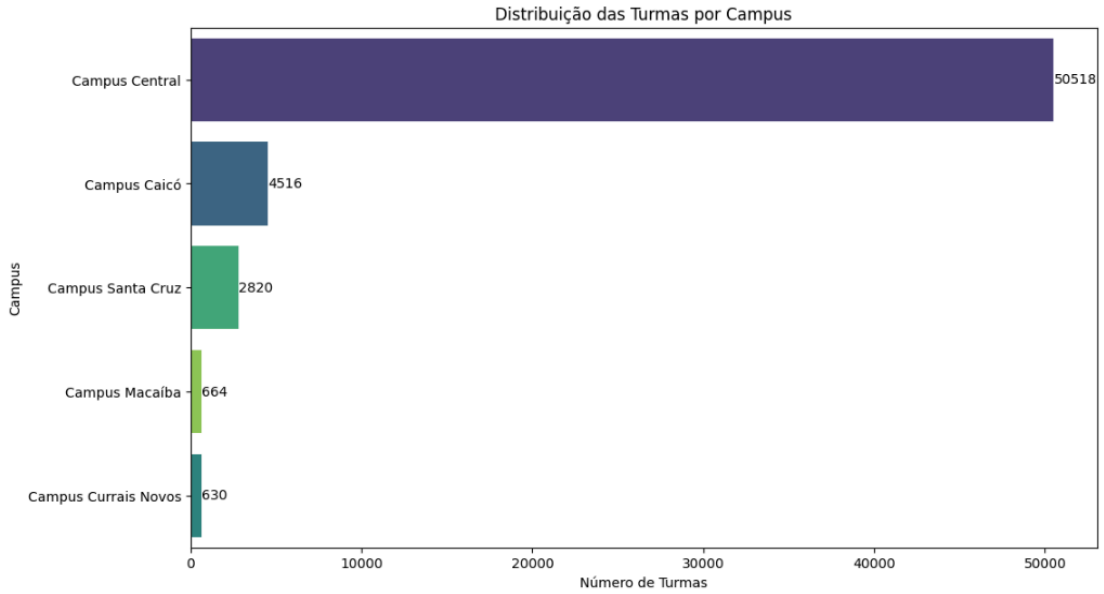
Observando as Figuras 4 e 5, nota-se que, na primeira contagem, 295 valores marcados como *False* correspondem às duplicatas identificadas. A pequena diferença de 13 linhas entre a contagem inicial de registros únicos (*False*) e o total final de linhas únicas (14.885 contra 14.872) ocorre porque o método “drop_duplicates” mantém apenas a primeira ocorrência. Além disso, o índice foi resetado, mas a quantidade total de duplicatas removidas permaneceu consistentemente em 295.

As figuras e descrições são referentes ao Dataframe 2023.1, mas as ações de remoção de colunas e remoção das duplicatas também foram replicadas para os Data frames 2023.2, 2024.1 e 2024.2.

ANÁLISE EXPLORATÓRIA

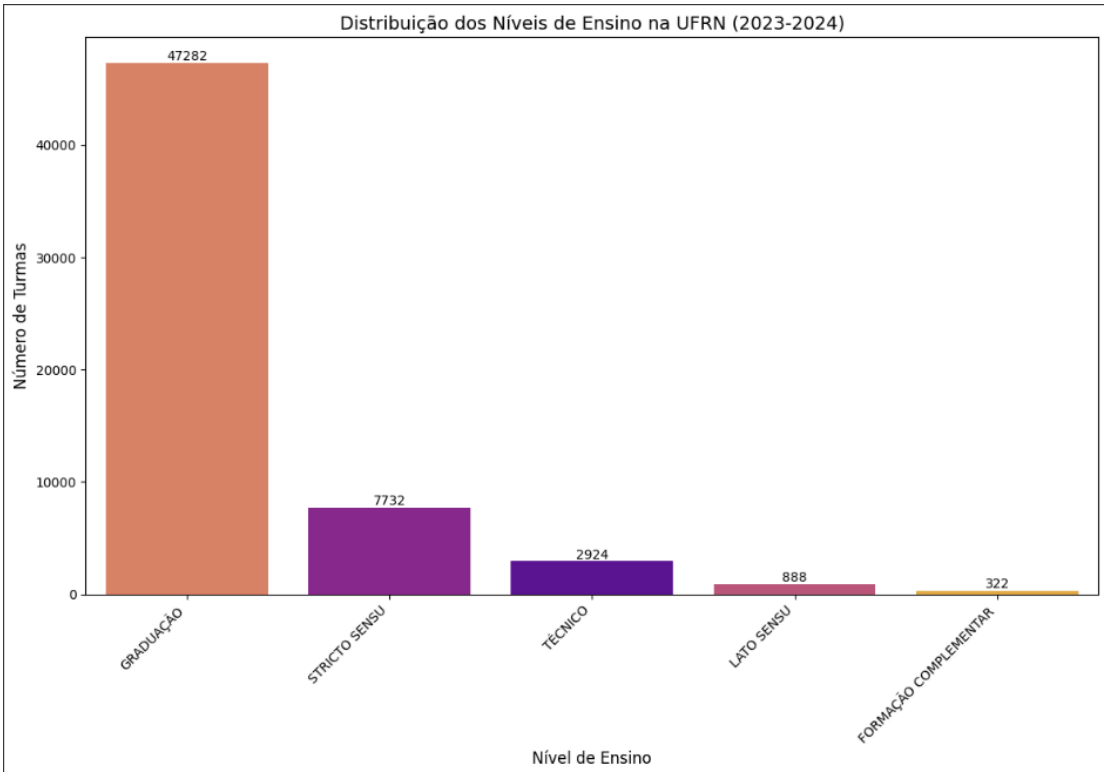
Ao longo da exploração foi observado que a distribuição das turmas entre diferentes níveis de ensino oferecidos pela UFRN foi essencial para delinear o perfil da oferta educacional da instituição.

Gráfico 01



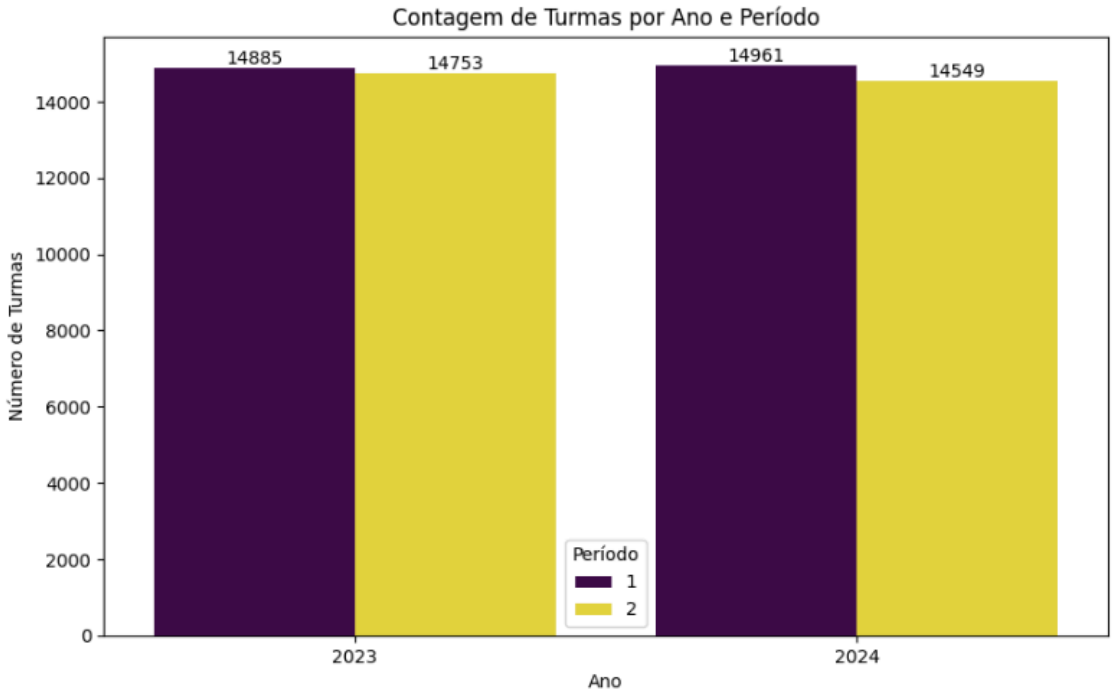
A análise gráfico 01 revela que a oferta educacional da UFRN é altamente concentrada tanto em termos geográficos quanto de nível de ensino. O Campus Central responde por mais de 50 mil turmas, superando amplamente os demais campi, que possuem números significativamente menores.

Gráfico 02



Analisando o gráfico 02 a barra mais alta corresponde à Graduação, com 47.282 turmas, representando de forma destacada a maior parte da oferta acadêmica. Em segundo lugar, aparece Stricto Sensu (pós-graduação voltada à pesquisa, como mestrados e doutorados), com 7.732 turmas, ainda consideravelmente menor que a graduação, mas com peso relevante no contexto institucional. O Técnico ocupa a terceira posição, com 2.924 turmas, voltado para formação profissional de nível médio. Já o Lato Sensu (pós-graduação de especialização) aparece com 888 turmas, enquanto Formação Complementar (cursos de curta duração ou aperfeiçoamento) é o menor grupo, com 322 turmas. Visualmente, percebe-se que a diferença entre a graduação e os demais níveis é muito grande, o que reforça o papel central da graduação na missão da UFRN.

Gráfico 03

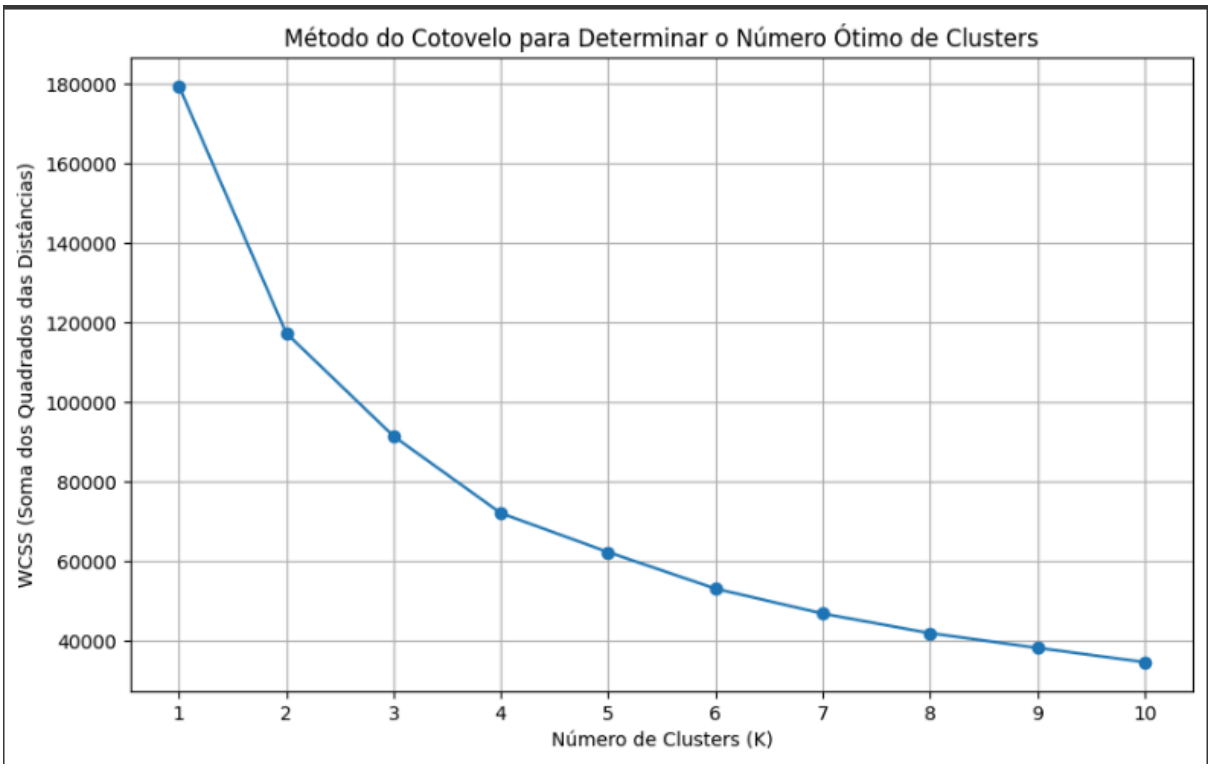


O gráfico 03 apresenta a contagem de turmas distribuídas entre os anos de 2023 e 2024, segmentadas pelos períodos 1 e 2. Observa-se que o número de turmas no primeiro período é ligeiramente superior ao do segundo período em ambos os anos analisados. Além disso, o total de turmas se manteve relativamente estável entre 2023 e 2024, apresentando apenas pequenas variações. Esses dados sugerem uma estabilidade na oferta de turmas ao longo dos dois anos, com uma preferência ou maior demanda pelo primeiro período em ambas as ocasiões.

ANÁLISE DOS TESTES

MÉTODO COTOVELO

Gráfico 04



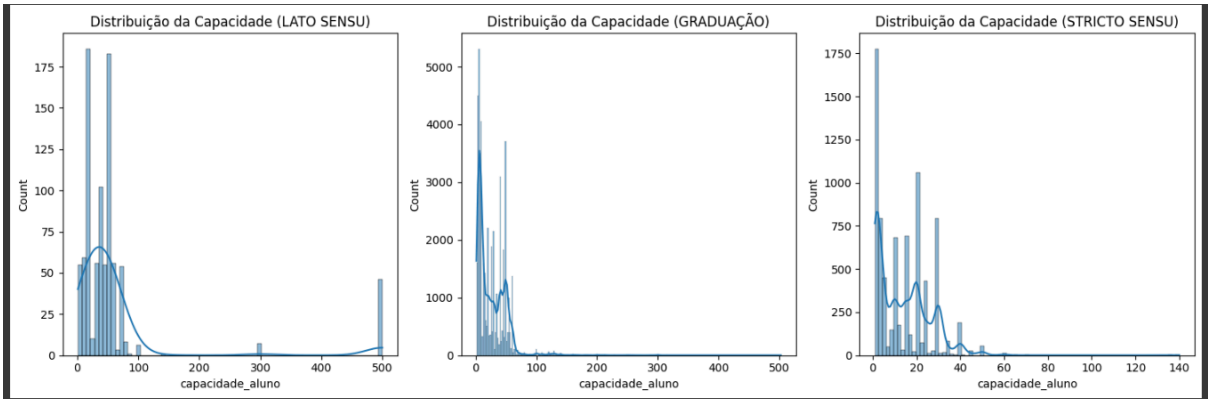
O gráfico 04 apresenta a aplicação do método do cotovelo para a determinação do número ótimo de clusters no conjunto de dados processados (df.processed).df.processed No eixo X está representado o número de clusters (K), enquanto no eixo Y encontra-se a soma dos quadrados intra-cluster (Within-Cluster Sum of Squares - WCSS), métrica que quantifica a variabilidade interna dos clusters.

Observa-se uma redução acentuada da WCSS com o aumento do número de clusters até aproximadamente K=3 ou K=4, indicando ganho significativo na compactação e homogeneidade dos grupos formados. A partir desse ponto, a taxa de redução da WCSS torna-se menos pronunciada, configurando uma inflexão na curva — característica conhecida como “cotovelo”.

Esse ponto de inflexão sugere que o número ideal de clusters situa-se entre 3 e 4, uma vez que a adição de clusters além desse limite proporciona ganhos marginais na minimização da variância intra-cluster. Assim, para balancear a complexidade do modelo com a qualidade do agrupamento, recomenda-se a escolha de 3 ou 4 clusters como valor ótimo de K.

TESTE ANOVA

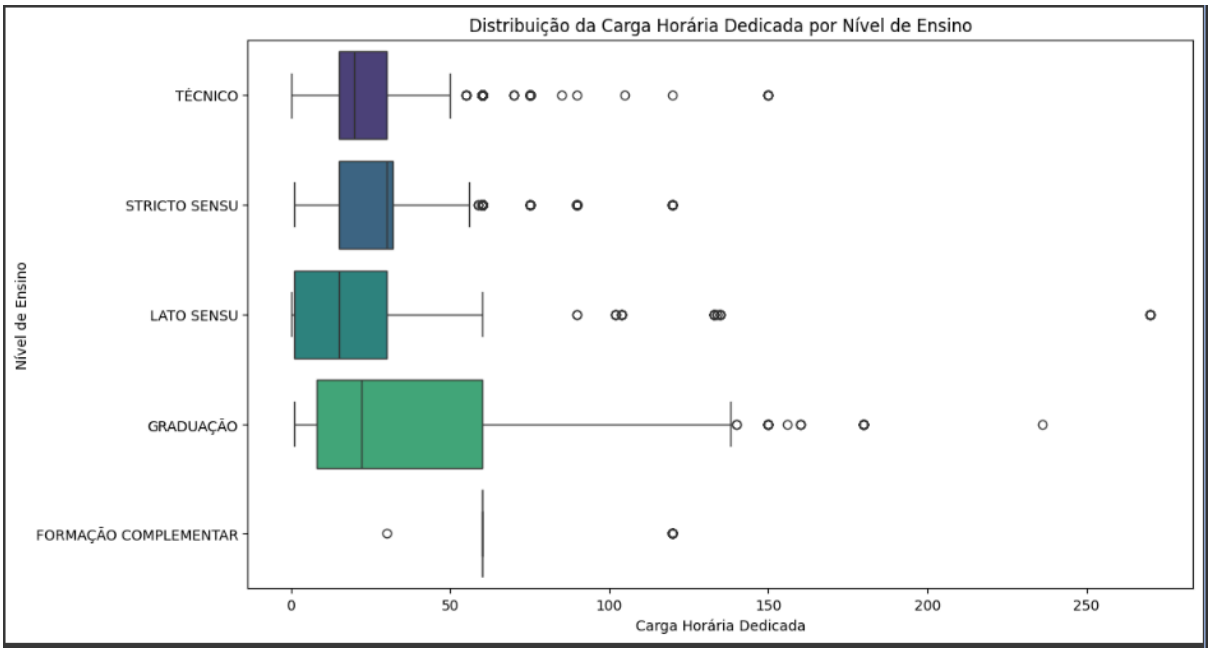
Gráfico 05



O teste ANOVA revelou uma diferença estatisticamente significativa na capacidade média dos alunos entre os diferentes níveis de ensino (Lato Sensu, Graduação e Stricto Sensu), conforme evidenciado pela alta estatística F (1600,79) e pelo valor de p extremamente baixo ($p < 0,05$), descrito no gráfico 05. Isso indica que a variação observada nas médias entre os grupos não é fruto do acaso, sugerindo que pelo menos dois dos níveis de ensino apresentam capacidades médias distintas. A análise visual dos histogramas corrobora essa conclusão, evidenciando diferenças claras na distribuição da capacidade entre os níveis avaliados.

TESTE KRUSKAL-WALLIS

Gráfico 06



O gráfico de boxplot apresentado ilustra a distribuição da carga horária dedicada em diferentes níveis de ensino, evidenciando variações significativas entre eles. O teste estatístico de Kruskal-Wallis reforça essa observação, apresentando um valor de p extremamente baixo ($p < 0,05$), o que indica que a hipótese nula de igualdade na distribuição da carga horária entre os níveis de ensino pode ser rejeitada. Portanto, podemos concluir que a carga horária dedicada varia de forma significativa conforme o nível de ensino, sendo que os cursos de graduação apresentam uma maior variação e valores médios mais elevados em comparação com os demais níveis, como técnico, stricto sensu, lato sensu e formação complementar. Esses resultados evidenciam a heterogeneidade na dedicação de carga horária entre os diferentes tipos de formação, corroborando a análise visual proporcionada pelo gráfico.

CONCLUSÃO

Durante o desenvolvimento deste trabalho, realizei uma análise detalhada dos dados das turmas da UFRN, desde a limpeza dos dados até a aplicação de técnicas de agrupamento (clustering) e testes estatísticos. As principais descobertas foram:

- **Limpeza dos Dados:** A etapa de limpeza foi fundamental para garantir a qualidade da análise. Identifiquei e tratei dados ausentes e duplicados, o que permitiu trabalhar com uma base mais confiável para as próximas fases.

- **Análise Exploratória de Dados (AED):** Observei que a maior parte das turmas oferecidas é de Graduação, concentradas principalmente no Campus Central. Também percebi uma correlação positiva entre a capacidade das turmas e o total de solicitações feitas pelos alunos. As turmas a distância, apesar de menos numerosas, costumam ter maior capacidade e recebem mais solicitações, em média. Além disso, notei que a oferta de turmas se manteve relativamente estável ao longo dos anos e períodos analisados, com algumas variações naturais.

- **Clustering:** Ao aplicar o K-Means, consegui segmentar as turmas em grupos com características parecidas, como turmas menores com carga horária reduzida, turmas com alta capacidade e demanda, e outras com carga horária maior. Essa segmentação ajuda a entender melhor a diversidade das turmas na UFRN.

- **Testes Estatísticos:** Confirmei com testes que a capacidade média e a carga horária variam significativamente entre os níveis de ensino. Também ficou claro que o padrão de solicitações difere entre as modalidades presencial e a distância.

Em resumo, a UFRN tem um foco forte na graduação e no Campus Central, mas existem particularidades importantes nos outros campi e níveis. Também há desafios e oportunidades diferentes para as modalidades presencial e a distância, considerando a demanda e a capacidade das turmas.

Recomendações

Com base nesses resultados, sugere algumas ações para a UFRN:

1. **Focar nos grupos de turmas com alta demanda ou baixa capacidade:** Investigar o que causa esses gargalos — pode ser falta de professores, infraestrutura ou oferta insuficiente — e pensar em soluções como aumentar a oferta dessas turmas, realocar recursos ou ajustar os horários.

2. **Expandir a modalidade a distância:** Como essas turmas tendem a ter maior capacidade e demanda, a universidade poderia ampliar essa oferta, especialmente em campi que não sejam o Campus Central, para atender melhor os alunos.

3. **Analisar as turmas que foram excluídas:** Entender por que muitas turmas são canceladas ou retiradas da oferta, para melhorar o planejamento e evitar desperdício de recursos.

Essas sugestões podem ajudar a UFRN a melhorar a organização das turmas, o uso dos recursos e o atendimento às necessidades dos estudantes e docentes.

REFERÊNCIAS

Boas-vindas ao Curso de Ciência de Dados! — Curso de Ciência de Dados. Disponível em: <https://petcc-ufrn.github.io/minicurso-ciencia-de-dados/intro.html>. Acesso em: 10 ago. 2025.

Conjuntos de dados - Dados Abertos da UFRN. Disponível em: <https://dados.ufrn.br/dataset>. Acesso em: 10 ago. 2025.