

REPORT

Análise Do Banco de Dados

“População Indígena Alfabetizada no Brasil”

Trabalho ponto extra
complementar de nota P1

para a disciplina:
Probabilidade e Estatística_Eng.Soft06_B_Noite

Professor: Rafael Mynssem

Grupo:

Denis Gomes Bomfim – 202211029

Kauã Arruda Machado - 202211399

Marcos Luiz de Sousa Reis - 202211072

Paulo Victor Melo Da Silva Ferreira - 202110217

Ano 2024-2

Sumário

1. Introdução
2. Metodologia
 - 2.1. Formação do Grupo e Definição de Tarefas
 - 2.2. Pesquisa e Coleta de Dados
 - 2.3. Análise e Tratamento de Dados
 - 2.4. Estudo dos Fenômenos Relacionados à Alfabetização
 - 2.5. Desenvolvimento do Código e do Notebook Jupyter
 - 2.6. Criação de Repositório Git e Submissão para Avaliação
3. Pré-processamento e Tratamento de Dados
 - 3.1. Tratamento de Valores Ausentes
 - 3.2. Salvamento dos Dados Tratados
4. Transformações de Variáveis
5. Geração de Gráficos para Análise Visual
 - 5.1. Boxplot para Identificação de Outliers
 - 5.2. Histograma da Distribuição da População Indígena Alfabetizada
 - 5.3. Gráfico de Barras Agrupadas por Faixa Etária e Sexo
 - 5.4. Matriz de Correlação entre Variáveis Numéricas
6. Estatísticas Descritivas
7. Resultados Esperados
8. Conclusões e Próximos Passos
9. Considerações Finais

Resumo

Este projeto/estudo tem como objetivo a “análise e tratamento de dados da população indígena alfabetizada no Brasil”, utilizando um banco de dados escolhido com informações categorizadas por município, sexo e faixa etária. O trabalho foi desenvolvido por um grupo de 4 membros engenheiros de software em formação, que dividiram as tarefas desde a pesquisa de dados até o desenvolvimento de um código em Python para análise e a geração do notebook Jupyter para submissão através de commit no repositório GitHub para avaliação.

As etapas do projeto incluíram:

- Coleta e limpeza dos dados, com o tratamento de valores ausentes e conversão de variáveis categóricas para numéricas.
- Análise estatística e visualização, utilizando gráficos como boxplots, histogramas e matrizes de correlação para identificar padrões e outliers.
- Desenvolvimento de um Jupyter Notebook para executar o código e realizar as análises de forma interativa.
- Criação de um repositório Git, permitindo a colaboração contínua e a submissão para avaliação.

Os resultados obtidos oferecem uma visão detalhada sobre a alfabetização da população indígena e podem embasar futuras políticas públicas voltadas à inclusão educacional. O estudo também abre portas para a ampliação da análise com variáveis adicionais e o uso de técnicas de machine learning para previsões mais avançadas.

1. Introdução

Este projeto visa analisar a população indígena alfabetizada no Brasil, segmentada por município, sexo e faixa etária. O objetivo é identificar padrões de alfabetização, detectar outliers e investigar possíveis correlações entre os grupos demográficos. A análise inclui o pré-processamento dos dados, visualização gráfica e cálculo de estatísticas descritivas, fornecendo uma base para futuras ações em políticas educacionais.

2. Metodologia

2.1. Formação do Grupo e Definição de Tarefas

A equipe foi organizada com cinco integrantes, cada um responsável por uma função específica, promovendo uma divisão eficiente de tarefas. A colaboração envolveu desde a pesquisa inicial até o tratamento dos dados, garantindo que cada fase do projeto fosse conduzida de maneira eficiente e organizada.

2.2. Pesquisa e Coleta de Dados

O grupo pesquisou e selecionou um banco de dados relevante para a análise da população indígena alfabetizada no Brasil. O foco foi em variáveis essenciais como município, sexo, faixa etária e alfabetização, proporcionando uma visão abrangente da situação de alfabetização no país. Durante a coleta, avaliou-se a qualidade dos dados e verificou-se a presença de valores ausentes ou inconsistências.

2.3. Análise e Tratamento de Dados

Após a coleta dos dados, foi realizada uma análise inicial para garantir a consistência do banco de dados. O tratamento dos dados incluiu as seguintes etapas:

Limpeza e pré-processamento: Correção de erros e preenchimento de valores ausentes.

Transformações de variáveis: Conversão de variáveis categóricas (como sexo) para valores numéricos, além do agrupamento de idades em faixas etárias.

Visualização dos dados: Uso de gráficos como boxplots e histogramas para identificar padrões e outliers.

2.4. Estudo dos Fenômenos Relacionados à Alfabetização

Foi realizado um estudo dos fatores que influenciam a alfabetização da população indígena, como aspectos demográficos, geográficos e socioeconômicos. A análise foi fundamentada em artigos acadêmicos e dados oficiais, permitindo a comparação dos resultados obtidos com estudos anteriores.

2.5. Desenvolvimento do Código e do Notebook Jupyter

O código foi desenvolvido em Python e implementado no Jupyter Notebook para facilitar a análise interativa. A equipe criou scripts para a limpeza, transformação e visualização dos dados, utilizando gráficos interativos para interpretar e visualizar as descobertas.

2.6. Criação de Repositório Git e Submissão para Avaliação

Um repositório Git foi criado para compartilhar e versionar o trabalho de forma colaborativa. O uso do Git garantiu a integração contínua das contribuições de cada membro da equipe e facilitou a submissão do projeto para conhecimento e avaliação final.

3. Pré-processamento e Tratamento de Dados

3.1. Tratamento de Valores Ausentes

A variável “população indígena alfabetizada” apresentava valores ausentes, que foram preenchidos com a mediana. Essa escolha minimizou o impacto de outliers e garantiu a consistência dos dados.

3.2. Salvamento dos Dados Tratados

Os dados limpos foram exportados para um novo arquivo, `dados_limpos.csv`, para garantir que qualquer modificação futura possa ser feita de forma estruturada e controlada.

4. Transformações de Variáveis

Foram realizadas as seguintes transformações:

- **Agrupamento em Faixas Etárias:** As idades foram agrupadas para facilitar a análise visual.
- **Conversão de Sexo em Valores Numéricos:** Para facilitar os cálculos, o sexo foi convertido em valores numéricos (0 para masculino, 1 para feminino).

5. Geração de Gráficos para Análise Visual

5.1. Boxplot para Identificação de Outliers

O Boxplot revelou a presença de outliers, indicando municípios com populações indígenas alfabetizadas significativamente maiores que a maioria.

5.2. Histograma da Distribuição da População Indígena Alfabetizada

O histograma mostrou uma distribuição assimétrica, com a maioria dos municípios apresentando valores baixos e alguns municípios com valores excepcionalmente altos.

5.3. Gráfico de Barras Agrupadas por Faixa Etária e Sexo

Gráficos de barras permitiram identificar diferenças na alfabetização entre grupos de idade e sexo. Observou-se uma maior alfabetização em faixas etárias intermediárias e variação entre os sexos.

5.4. Matriz de Correlação entre Variáveis Numéricas

A matriz de correlação mostrou que a variável *população indígena alfabetizada* não apresentou correlações significativas com as variáveis categóricas transformadas, sugerindo que outros fatores influenciam a alfabetização.

6. Estatísticas Descritivas

Foram calculadas estatísticas como média, mediana, desvio padrão e variância para a variável “população indígena alfabetizada”. Esses cálculos forneceram uma visão clara da distribuição dos dados, com foco nas disparidades entre os municípios.

7. Resultados Esperados

Espera-se que a análise revele municípios com baixa taxa de alfabetização indígena, permitindo intervenções direcionadas. Além disso, padrões demográficos poderão ser identificados, sugerindo a necessidade de políticas públicas focadas em faixas etárias ou sexos específicos. A análise também fornecerá uma base para futuras expansões, como a inclusão de variáveis socioeconômicas.

8. Conclusões e Próximos Passos

O estudo mostrou uma distribuição desigual da alfabetização indígena no Brasil, com alguns municípios se destacando significativamente. Há variação na alfabetização por sexo e faixa etária, o que pode direcionar futuras políticas educacionais.

Próximos passos:

- Expansão de variáveis: Incluir dados adicionais, como aspectos socioeconômicos e localização geográfica.
- Machine Learning: Utilizar algoritmos de aprendizado de máquina para prever os níveis de alfabetização e identificar padrões mais complexos.

9. Conclusão – considerações finais

Este relatório sobre o projeto/estudo de “Análise e Tratamento de Dados da População Indígena Alfabetizada” oferece uma base sólida para a compreensão dos padrões de alfabetização entre os povos indígenas no Brasil. O trabalho realizado pelo grupo de engenheiros de software em formação, que incluiu o tratamento e a análise dos dados, bem como a visualização dos resultados, proporciona uma visão detalhada sobre a distribuição demográfica da alfabetização indígena, destacando tendências e discrepâncias importantes entre diferentes grupos etários e de gênero. Os resultados obtidos podem servir como um ponto de partida para investigações mais profundas e, potencialmente, embasar políticas públicas direcionadas à melhoria das condições de alfabetização nas comunidades indígenas. Ao identificar padrões e outliers, como a baixa alfabetização em determinadas faixas etárias ou a discrepância entre gêneros, este estudo fornece evidências que podem orientar ações governamentais e educacionais para reduzir as desigualdades e promover a inclusão educacional. Além disso, o projeto demonstra a importância de um tratamento rigoroso dos dados, evidenciando como o processo de limpeza, pré-processamento e análise estatística pode revelar insights que não seriam percebidos em uma abordagem superficial. O uso de ferramentas como o Jupyter Notebook e o repositório Git permitiu uma organização eficiente do trabalho, facilitando a colaboração entre os membros do grupo e garantindo a qualidade e reprodutibilidade dos resultados. Com base nesses achados, há potencial para expandir a

análise incluindo variáveis socioeconômicas e geográficas mais detalhadas, bem como a aplicação de técnicas de machine learning para prever a alfabetização indígena em diferentes cenários. Desta forma, este estudo contribui não apenas para a compreensão atual do fenômeno, mas também para o desenvolvimento de soluções sustentáveis e informadas.

Referências

<https://basedosdados.org/dataset/08a1546e-251f-4546-9fe0-b1e6ab2b203d?table=598a08fb-8e4e-430d-a46a-aa1c93448088> - Disponível em: [Censo 2022 – Base dos Dados](#), visitado em outubro/2024