📊 Projeto de Análise de Dados - Grupo 1_5

📜 Descrição

Este projeto analisa os dados educacionais de estudantes das escolas Mousinho da Silveira e Gabriel Pereira, com foco em desempenho acadêmico e fatores socioeconômicos.

🎯 Objetivos

- O objetivo deste projeto é analisar o desempenho dos alunos nas disciplinas de Matemática e Língua Portuguesa, observando como variáveis sociais e demográficas (como status dos pais, acesso à internet, tamanho da família e sexo) podem influenciar as notas (G1, G2, G3) e o histórico de reprovações.

🗂️ Fonte dos Dados

- Fonte: https://archive.ics.uci.edu/dataset/320/student+performance
- A base de dados original da disciplina matemática contém informações de 395 alunos, cada um representado por uma linha no DataFrame. Cada registro inclui diversas informações pessoais, familiares, escolares e as notas G1, G2 e G3, correspondentes ao desempenho do aluno na disciplina de matemática ao longo do período letivo. A base contém 396 linhas (incluindo cabeçalho) x 33 colunas.
- A base de dados original da disciplina Língua Portuguesa, por sua vez, traz informações de 649 alunos, cada um representado por uma linha no DataFrame. Assim como no anterior, cada registro inclui informações pessoais, familiares, escolares e suas respectivas notas (G1, G2 e G3), correspondentes ao desempenho do aluno na disciplina de Língua Portuguesa, distribuídas em 33 colunas. A base contém 650 linhas (incluindo cabeçalho) x 33 colunas.
- Após a importação, as bases de Matemática (395 registros) e Língua Portuguesa (649 registros) foram concatenadas, totalizando 1044 entradas antes da limpeza.

- Não há, em nenhuma das bases, campos vazios ou informações passíveis de descarte.
- Após um processo de limpeza dos dados, no qual foram removidos valores discrepantes (outliers), identificados com base no z-score (valores acima de 3 desvios padrão), a base passou a conter um total de 163 alunos. Essa etapa é fundamental para garantir uma análise mais fiel à realidade e evitar distorções causadas por outliers.

🛠️ Tecnologias e Bibliotecas Utilizadas

- Python 3.x
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Google Colab

🖌️ Etapas do Projeto

1. **Importação e Visualização dos Dados**
2. **Limpeza e Tratamento**
3. **Análise Exploratória (EDA)**
4. **Visualizações e Insights**
5. **Conclusões**

📊 Análises Realizadas

Para esta análise, foram levantadas algumas perguntas:

- O acesso à internet pode influenciar o desempenho dos alunos?
- As notas podem ser influenciadas pelo tamanho da família do aluno?
- Há diferença significativa de desempenho entre gêneros?

Como complemento das análises realizadas, foram utilizados gráficos de barras e boxplots. Esses gráficos permitem visualizar de maneira mais clara as diferenças de desempenho entre os grupos sociais observados, o que ajuda a ressaltar as tendências nas notas e no número de reprovações baseados nas variáveis disponíveis na base de dados.

Essas questões guiaram a análise exploratória a seguir:

Distribuição Etária

- **Análise:** A faixa etária dos alunos na base de dados filtrada varia principalmente entre 15 e 19 anos, com a maioria concentrada entre 16 e 18 anos. Essa concentração sugere que a base analisada após o filtro é composta principalmente por alunos do ensino médio, o que está de acordo com o contexto educacional da base original. Além disso, foram gerados gráficos de distribuição da idade dos alunos e gráficos de contagem para variáveis como sexo, status dos pais, acesso à internet e tamanho da família.Essas visualizações ajudaram a compreender o perfil dos estudantes na base de dados filtrada e contextualizaram melhor os grupos observados nas comparações de desempenho.

Status dos Pais (Pstatus)

- **Análise:** Ao comparar as médias das notas (G1, G2, G3) com o número de reprovações ('failures') entre alunos com pais que vivem juntos ('T') e pais separados ('A'), observamos que as médias das notas G1 e G2 foram ligeiramente maiores para alunos com pais vivem juntos. A média da nota G3 foi um pouco maior para alunos com pais separados. As diferenças nas médias das notas entre os grupos não foram muito grandes. A média de reprovações foi ligeiramente menor para alunos com pais que vivem juntos.

Acesso à Internet (internet)

- **Análise:** Ao comparar as médias das notas e reprovações entre alunos que têm acesso à internet ('yes') e aqueles que não têm ('no'), notamos que as médias das notas G1, G2 e G3 (e a média geral) foram ligeiramente maiores para alunos com acesso à internet. A média de reprovações foi um pouco maior para alunos com acesso à internet, o que pode parecer contraintuitivo e sugere a necessidade de investigação adicional.

Tamanho da Família (famsize)

- **Análise:** Ao comparar as médias das notas e reprovações entre alunos de famílias maiores ('GT3') e menores ('LE3'), observamos que as médias das notas G1, G2 e G3 (e a média geral) foram ligeiramente maiores para alunos de famílias menores. A média de reprovações foi muito similar entre os dois grupos.

Sexo (sex)

- **Análise:** Ao comparar as médias das notas e reprovações entre alunos do sexo feminino ('F') e masculino ('M'), notamos que as médias das notas G1, G2 e a média geral foram ligeiramente maiores para as alunas. A média da nota G3 foi similar. A média de reprovações foi visivelmente maior para os alunos do sexo masculino.

Relacionamento amoroso (romantic)

- **Análise:** Os resultados mostram que os alunos sem relacionamento amoroso ("no") obtiveram uma nota média de 10.95, enquanto os alunos com relacionamento amoroso ("yes") obtiveram uma nota média de 10.14. Isso sugere que, neste conjunto de dados, os alunos sem relacionamento amoroso tiveram um desempenho ligeiramente superior nas notas em comparação com os alunos que estão em um relacionamento.

Relação entre faltas e notas (absences)

- **Análise:** Para visualizar a relação entre o número de faltas ('absences') e as notas (G1, G2, G3), foram criados gráficos de dispersão. Estes gráficos ajudam a identificar visualmente se há alguma tendência, mesmo que não seja uma correlação linear direta.

Relação entre tempo de viagem e notas (traveltime)

- **Análise:** Para visualizar a relação entre o tempo de viagem até a escola ('tempo_viagem_escola') e as notas (G1, G2, G3), foram criados gráficos de dispersão. Esses gráficos ajudam a identificar se há tendência, ainda que não seja uma correlação linear forte.

Relação entre Tempo de Estudo e Notas (studytime)

- **Análise:** Para visualizar a relação entre o tempo semanal de estudo fora de sala de aula (‘study_time’) e as notas (G1, G2, G3), foram usados, novamente, gráficos de dispersão. Esses gráficos ajudam a identificar se há tendência, ainda que não seja uma correlação linear forte.

🧠 Conclusões

Seguem abaixo o resumo das conclusões, após a análise realizada sobre o desempenho dos estudantes, com ênfase em como fatores sociais e demográficos podem influenciar suas notas e o número de reprovações, com base nos dados filtrados:

Distribuição Etária

- **Conclusão Resumida:** Após a remoção de outliers, a maioria dos alunos analisados está concentrada entre 16 e 18 anos, refletindo a predominância de estudantes do ensino médio na base filtrada.

Status de relacionamento dos pais (Pstatus)

- **Conclusão Resumida:** Alunos com pais que vivem juntos ('T') apresentaram médias ligeiramente maiores nas notas iniciais (G1 e G2) e menos reprovações do que alunos com pais separados ('A'). A diferença na nota final (G3) foi mínima.

Acesso à Internet (internet)

- **Conclusão Resumida:** Alunos com acesso à internet ('yes') tenderam a ter notas ligeiramente mais altas (G1, G2, G3) do que aqueles sem internet ('no'). No entanto, a média de reprovações anteriores foi maior para alunos com acesso à internet, uma relação que exigiria mais investigação.

Tamanho da Família (famsize)

- **Conclusão Resumida:** Alunos de famílias menores ('LE3') apresentaram médias de notas (G1, G2, G3) ligeiramente superiores em comparação com alunos de famílias maiores ('GT3'). O tamanho da família teve um impacto mínimo no número de reprovações anteriores.

Sexo (sex)

- **Conclusão Resumida:** Alunas ('F') tiveram médias de notas ligeiramente mais altas em G1 e G2 do que alunos ('M'). A diferença mais notável foi nas reprovações anteriores, com alunos do sexo masculino apresentando uma média visivelmente maior de reprovações.

Relacionamento amoroso (romantic)

- **Conclusão resumida:** Os resultados mostram que os alunos sem relacionamento amoroso ("no") obtiveram uma nota média de 10.95, enquanto os alunos com relacionamento amoroso ("yes") obtiveram uma nota média de 10.14. Isso sugere que, neste conjunto de dados, os alunos sem relacionamento amoroso tiveram um desempenho ligeiramente superior nas notas em comparação com os alunos que estão em um relacionamento.

Relação entre faltas e notas (absence)

- **Conclusão resumida:** Para visualizar a relação entre o número de faltas ('absences') e as notas (G1, G2, G3), foram criados gráficos de dispersão. Estes gráficos ajudam a identificar visualmente se há alguma tendência, mesmo que não seja uma correlação linear direta.

Relação entre tempo de viagem e notas (traveltime)

- Conclusão resumida: Para visualizar a relação entre o tempo de viagem até a escola ('traveltime') e as notas (G1, G2, G3), foram criados gráficos de dispersão. Esses gráficos ajudam a identificar se há tendência, ainda que não seja uma correlação linear forte.

Relação entre tempo de estudo e notas (studytime)

- **Conclusão resumida:** Para visualizar a relação entre o tempo semanal de estudo fora de sala de aula (‘studytime’) e as notas (G1, G2, G3), foram usados, novamente, gráficos de dispersão. Esses gráficos ajudam a identificar se há tendência, ainda que não seja uma correlação linear forte.

🚧 Limitações da Análise

- A base de dados foi reduzida para 163 alunos após a filtragem, o que limita a generalização dos resultados.
- Por serem autorreportadas, as variáveis sociais analisadas podem conter vieses.

- Não foi realizada nenhuma modelagem estatística ou teste de significância, então as relações observadas não confirmam causalidade.

Considerações finais

Após a análise dos dados disponíveis, concluiu-se que as notas (G1, G2, G3) e o histórico de reprovações não sofreram impacto ou influência das variáveis sociais e demográficas apresentadas nas bases de dados consultadas.

📁 **Como Acessar o Projeto**

Notebook Google Colab