Ericky Dias, Felipe Freitas, Filipe Pereira, Guilherme Farçoni, Guilherme Sitton, Izaque Israel e Julia Issa

IA na Educação: Transformando a Avaliação para Combater a Formação Sem Base
A Revolução da IA na Avaliação Educacional: Personalização para Alunos

Brasil
2023

Resumo

Estudantes da TripleTen foram desafiados para um hackathon. Alunos da área de dados e web dev, se juntaram formando a organização Sons of Node, para cumprir um desafio. Trazer luz acerca do tema Inteligência Artificial. Neste relatório contém toda a análise desenvolvida pela área de dados, desde a obtenção de dados até seu dicionário, análise exploratória, data vis e conclusões geradas. Nosso projeto se baseia na defesa do avanço da IA, principalmente no ramo da educação, onde analisamos e entendemos melhor sobre o problema da formação de alunos do ensino médio sem base. (explicar mais a análise)

Palavras-chave

Sumário

Introdução

Hackathon é um projeto multidisciplinar na área de tecnologia, feito tanto por instituições da área da educação quanto por empresas, para obter novas ideias e revolucionar o mercado. Suas únicas regras são trabalhar por um tempo limitado somente em projetos novos e criativos que não estejam relacionados ao dia a dia da empresa ou da instituição. Muitas coisas que já estão presentes no nosso cotidiano foram criadas em hackathon. Em 2013 o Facebook realizou hackathons para a melhoria de sua plataforma e consequentemente da empresa, funções como álbuns compartilhados, figurinhas de like e climas nos eventos foram uma das ideias implementadas deste hackathon. Até mesmo o botão "curtir" saíram destas maratonas. Seus benefícios são incontáveis.

O hackathon da TripleTen tem como objetivo apresentar dados pertinentes relacionados ao tema de inteligência artificial - seus prós e contras. A organização Sons of Node defende o avanço da IA (explicar melhor a opinião do grupo) e tem como objetivo aprofundar o tema em relação a educação, combatendo o problema de formação de alunos do ensino médio sem de fato ter o conhecimento esperado da Base Nacional Comum Curricular.

Nosso projeto consiste em fazer uma análise exploratória de dados sobre o desempenho dos alunos em todo o Brasil, usando como métrica o Exame Nacional do Ensino Médio (ENEM). Desta forma, a área de web irá criar um site com a apresentação dos dados e com uma prova geral (desenvolvida por IA) sobre a matéria escolhida, gerando ao final dela um feedback personalizado ao aluno descrevendo pontos que ele precisa evoluir para ter o total compreendimento da matéria, para desenvolver a educação brasileira como um todo, podendo assim ser facilmente escalável e tendo tanto escolas como o próprio governo como possíveis clientes

Como dito anteriormente, a Sons of Node utiliza dados do ENEM, disponibilizado pelo próprio governo brasileiro. Tendo em si uma vasta gama de dados sabendo todo o background do aluno. Desde o sexo, idade, raça, qual a situação de conclusão do ensino médio, estado civil, se estudou em escola pública ou privada e etc. (opinião da descrição)

Começamos a análise com uma 'limpeza' dos dados, pois muita coisa encontrada era desnecessária, além de deixar o dataframe pesado e lento ao trabalhar com ele. Tinham colunas, por exemplo, que havia mais de 70% de dados ausentes, então resolvemos removê-las para facilitar a análise já que não eram colunas necessárias. Após a limpeza começamos a AED, análise exploratória de dados. Dentre as várias descobertas encontradas podemos confirmar que matemática é de fato a matéria com o menor desempenho dentre todos os alunos, e que o conhecimento se defasa com o tempo.

O relatório é dividido em sessões. A primeira sessão se chama Materiais e Métodos, mais focada na metodologia de pesquisa, pensamentos, e esclarecimento dos dados de uma forma mais aprofundada. A segunda sessão é Análise Exploratória de Dados, já é sobre a análise em si,

com nossas descobertas e visualizações. Partindo então para a nossa terceira sessão onde compartilhamos os resultados e conclusões gerais.

Sessão 1: Materiais e Métodos

O pensamento geral desde o princípio, sempre foi defender o avanço da inteligência artificial. O campo escolhido foi a educação, que é um campo muito amplo e cheio de oportunidades, mas precisávamos definir um foco, e esse foco foi o desempenho dos estudantes do ensino médio brasileiro. Nossa hipótese é definida por dar mais conhecimento e também direção aos jovens brasileiros para que tenham o entendimento mínimo necessário ao sair do ensino médio, combatendo a formação de alunos sem a devida base da Base Nacional Comum Curricular. Resultando então em um aumento do índice de educação brasileira, fazendo com que o país se desenvolva cada vez mais.

Para então medir o desempenho dos alunos no ensino médio, nada melhor do que o Enem, que também recolhe dados de pessoas já formadas. Nossos dados foram coletados pelo governo brasileiro e disponibilizados no próprio site2. Contendo informações sobre o Enem realizado no ano de 2022.

O dataframe original vem com mais de 3 milhões de linhas e muitas colunas, havia também o dicionário original contendo mais duas páginas sobre outros dados que não foram utilizados neste hackathon. Após a limpeza dos dados criamos nosso próprio dicionário.

Dicionário dos Dados

id: Número de inscrição

gp_idade: Faixa etária

1: Menor de 17 anos

2: 17 anos

3: 18 anos

4: 19 anos

5: 20 anos

6: 21 anos

7: 22 anos

8: 23 anos

9: 24 anos

10: 25 anos

11: Entre 26 e 30 anos

12: Entre 31 e 35 anos

13: Entre 36 e 40 anos

14: Entre 41 e 45 anos

15: Entre 46 e 50 anos

16: Entre 51 e 55 anos

17: Entre 56 e 60 anos

18: Entre 61 e 65 anos

19: Entre 66 e 70 anos

20: Maior de 70 anos

***sexo**:* Sexo

M: Masculino

F: Feminino

***est_civil**:* Estado Civil

0: Não informado

1: Solteiro(a)

2: Casado(a)/Mora com companheiro(a)

3: Divorciado(a)/Desquitado(a)/Separado(a)

4: Viúvo(a)

raca: Cor/raça

0: Não declarado

1: Branca

2: Preta

3: Parda

4: Amarela

5: Indígena

6: Não dispõe da informação

sit_ens_med: Situação de conclusão do Ensino Médio

1: Já concluí o Ensino Médio

2: Estou cursando e concluirei o Ensino Médio em 2022

3: Estou cursando e concluirei o Ensino Médio após 2022

4: Não concluí e não estou cursando o Ensino Médio

ano_conclusao: Ano de Conclusão do Ensino Médio

0: Não informado

1: 2021

2: 2020

3: 2019

4: 2018

5: 2017

6: 2016

7: 2015

8: 2014

9: 2013

10: 2012

11: 2011

12: 2010

13: 2009

14: 2008

15: 2007

16: Antes de 2007

tp_ensino: Tipo de instituição que concluiu ou concluirá o Ensino Médio

0: Não informado

1: Ensino Regular

2: Educação Especial - Modalidade Substitutiva

tp_escola: Tipo de escola do Ensino Médio

1: Não Respondeu

2: Pública

3: Privada

treineiro: Indica se o inscrito fez a prova com intuito de apenas treinar seus conhecimentos

1: Sim

0: Não

present_cien: Presença na prova objetiva de Ciências da Natureza

0: Faltou à prova

1: Presente na prova

2: Eliminado na prova

present_hum: Presença na prova objetiva de Ciências Humanas

0: Faltou à prova

1: Presente na prova

2: Eliminado na prova

present_port: Presença na prova objetiva de Linguagens e Códigos

0: Faltou à prova

1: Presente na prova

2: Eliminado na prova

present_mat: Presença na prova objetiva de Matemática

0: Faltou à prova

1: Presente na prova

2: Eliminado na prova

nota_cien: Nota em Ciências da Natureza

nota_hum: Nota em Ciências Humanas

nota_port: Nota em Linguagens e Códigos

nota_mat: Nota em Matemática

esp_ing: Tipo de prova de língua estrangeira

0: Inglês

1: Espanhol

nota_redac: Nota da Redação

media_final: Média final de todas as notas

Ao finalizar o pré-processamento de dados, pensamos em enriquecer ele com a adição da coluna media_final, desse modo facilitando a comparação do desempenho com outros fatores presente no dataframe. O agrupamento de dados será feito na próxima sessão junto com a visualização do mesmo.

Sessão 2: Análise exploratória dos dados

- Aqui podemos observar a diferença gritante nas médias finais de quem vai para uma escola particular (570) em relação a quem vai para uma pública (370).

- Observando este gráfico podemos perceber muito mais provas aparentemente completadas no ENEM em pessoas vindo de escola particular do que pessoas de escola

pública, já que as notas baixas são consideradas outliars em escolas particulares, mas em públicas estão dentro do q1, sendo muito mais comum provas zeradas.

- Infelizmente a diferença é tanta que a mediana da particular é maior que o top75% da pública.

- Mulheres participam em 61.1% das cadeiras do ENEM em dias de prova. Percebe-se uma dominância em relação a inscrição.

- Porém a dominância acontece apenas em quantidade mas não em qualidade, pois este gráfico mostra a igualdade de conhecimento entre generos.

Outra opcao para esses graficos

- Você sabia que as mulheres são verdadeiras protagonistas quando se trata de participação nas provas do ENEM? Representando 61.1% das inscrições. Apesar dessa dominância numérica, os dados revelam igualdade de conhecimento entre os gêneros.

- Como era de se esperar, os dados revelam uma diferença significativa nas médias entre o ensino regular e o ensino especial, com uma margem impressionante de mais de 100 pontos de diferença.
- Acompanhe-nos nesta jornada de descobertas e entenda como podemos transformar a educação, tornando-a acessível a todos. Vamos construir um futuro onde todos tenham a oportunidade de brilhar!

- Há uma clara relação entre o aumento da nota da redação e a média final, porém a correlação da prova de matemática e a média final é maior do que a da redação com a média final, então ao ficar na dúvida entre números ou letras, vá nos números a maioria das vezes.

- Em razão para descobrir se matemática é mais importante ou a redação no quesito de influênciar a média final, no ENEM de 2022 a nota de matemática foi mais influente, tendo uma correlação de `0.827, enquanto a nota da redação teve uma correlação de 0.819`

- De forma surpreendente quem é do grupo 1 de idade (menos de 17 anos) tem uma vantagem de média das notas em relação a quem tem 17 anos. Claramente conforme o tempo passa o conhecimento é dissipado, atingindo seu menor ponto no grupo 12 (entre 31 e 35 anos).
- Uma percepção no mínimo curiosa que podemos tirar deste gráfico também, é que quem é do grupo 18 (entre 61 e 65) tem uma média maior do que quem é do grupo 6 (21 anos de idade) em diante, aparentemente uma explicação lógica para este dado seria dizer que quem tem uma idade mais avançada esta mais comprometido em estudar em comparação aos que saíram do ensino médio, mas acreditam ainda ter conhecimento o suficiente para a prova.

- Este gráfico representa um sample aleatório de 20.000 médias finais em relação ao tempo de conclusão do ensino médio. Partindo como parâmetro o grupo 1 de quem se formou em 2021, tem uma média final de 402.7. Após 3 anos essa média cai para aproximadamente 335.2 resultando em uma queda de aproximadamente 16,7%. Após 10 anos por exemplo essa média cai para 285.4, resultando em uma queda de aproximadamente 29%.

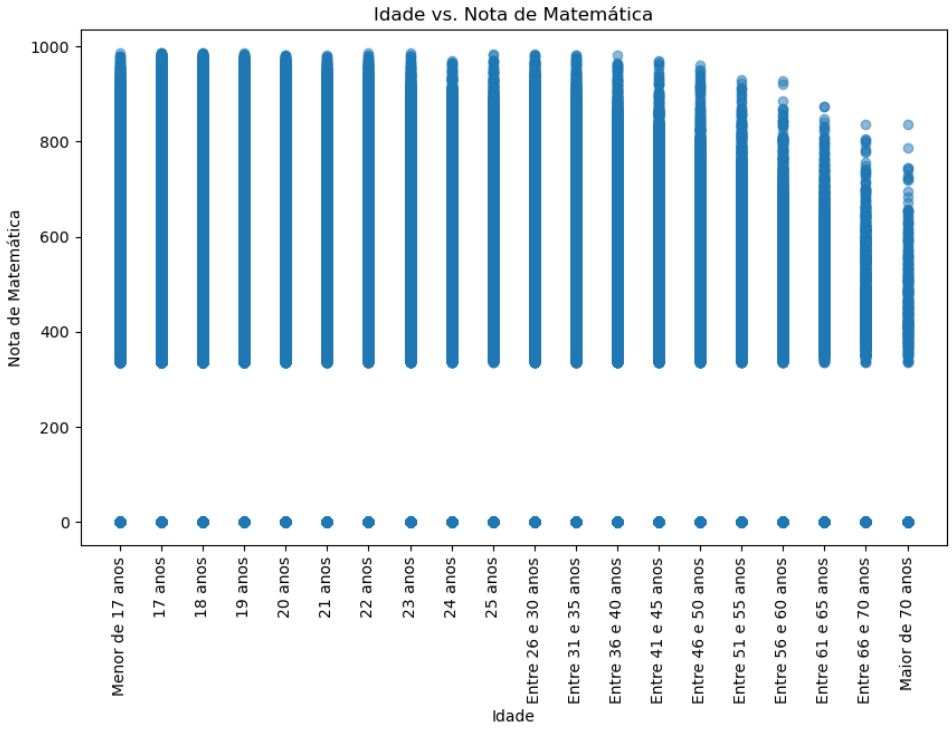
Outra opcao de texto

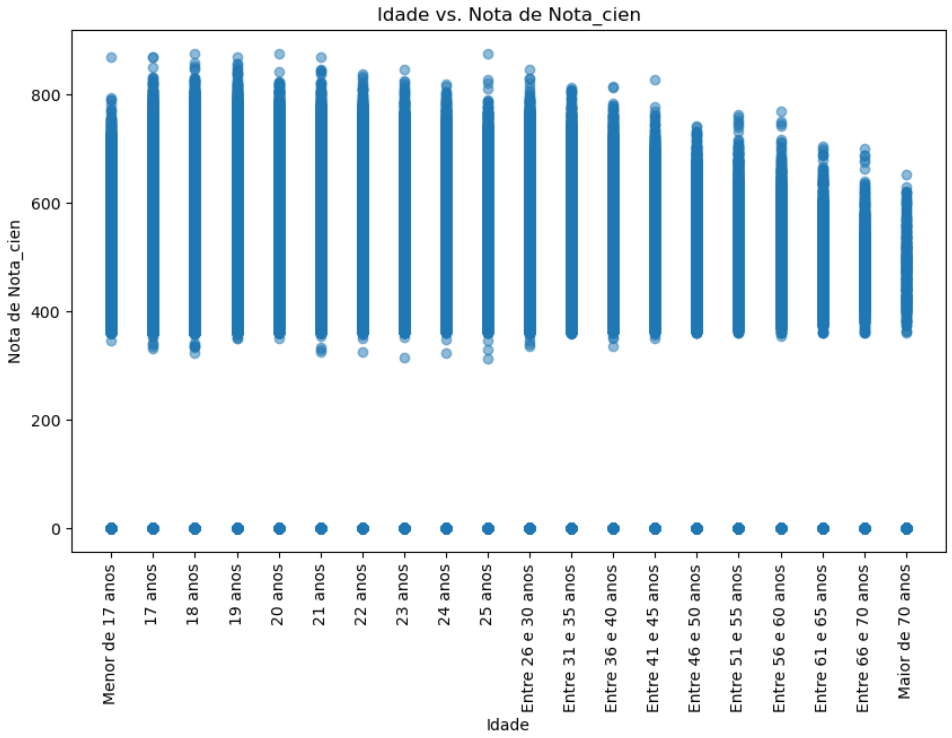
- Neste gráfico, desvendamos as histórias por trás de 20.000 médias finais, traçando um caminho desde a formatura no ensino médio. Vamos começar com o Grupo 1, os formandos de 2021, que possuem uma média sólida de 402.7.

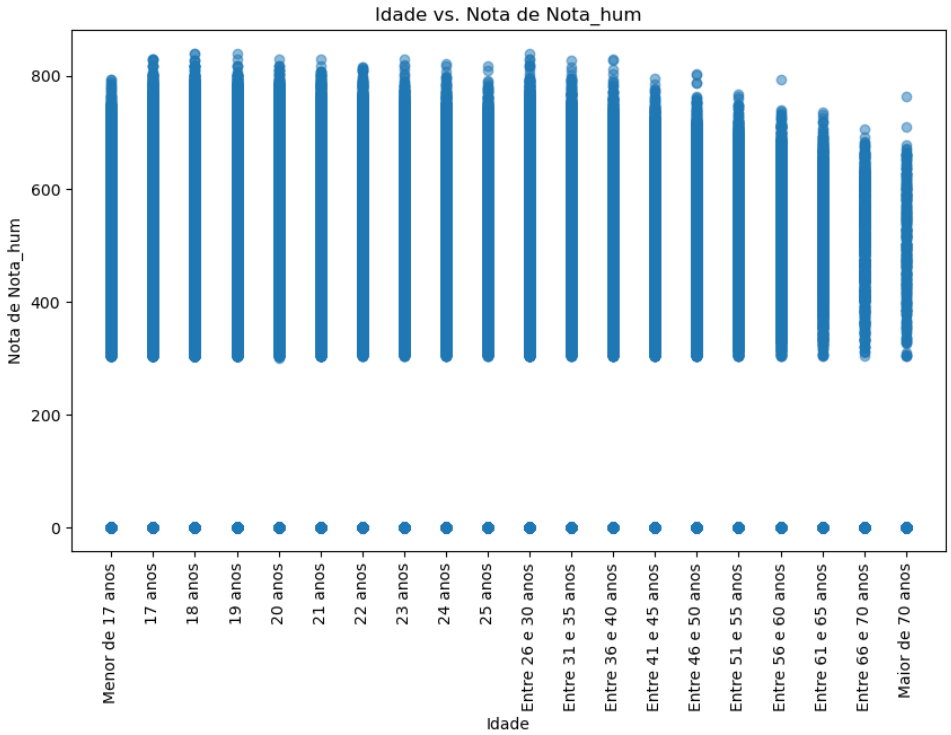
- Após apenas 3 anos, essa média despenca para cerca de 335.2, uma queda de cerca de 16,7%. Após uma década, a média desce para 285.4, resultando em uma incrível queda de cerca de 29%.

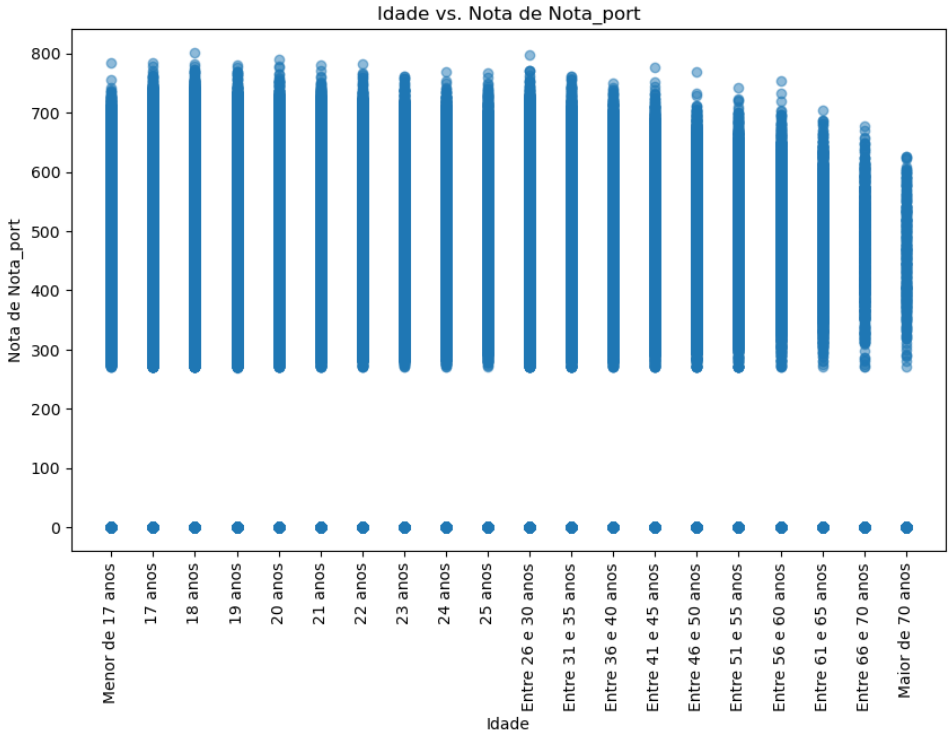
LISTA DE OUTRAS VIZUALIZAÇÕES GERADAS

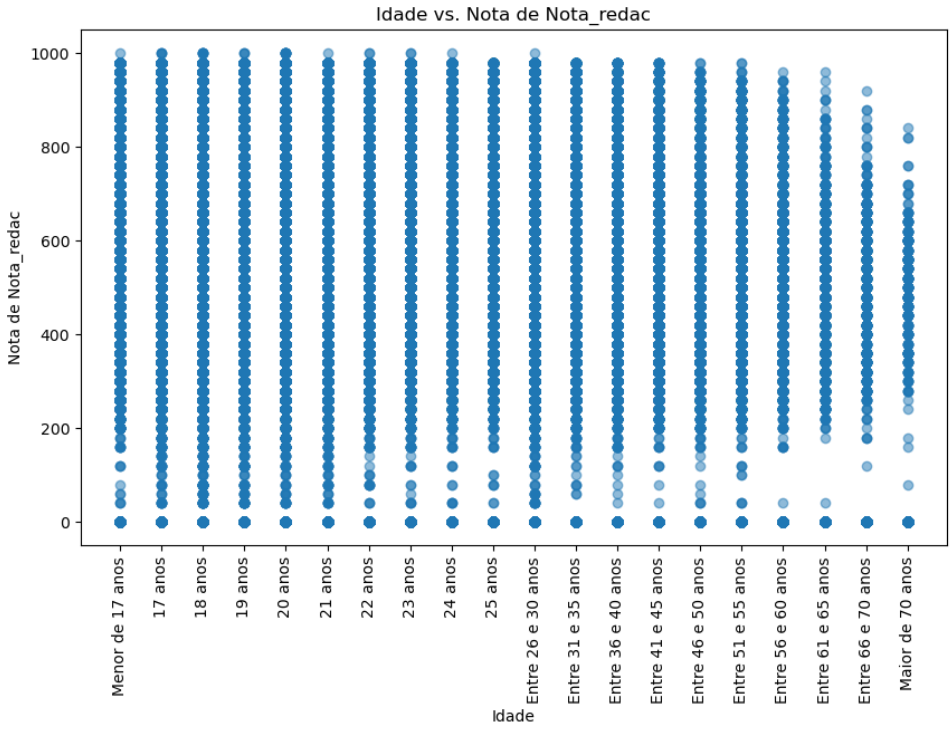
- Primeiro grupo de visualização


Idade vs. Nota de Matemática
Nota de Matemática
0
200
400
600
800
1000
Menor de 17 anos
17 anos
18 anos
19 anos
20 anos
21 anos
22 anos
23 anos
24 anos
25 anos
Entre 26 e 30 anos
Entre 31 e 35 anos
Entre 36 e 40 anos
Entre 41 e 45 anos
Entre 46 e 50 anos
Entre 51 e 55 anos
Entre 56 e 60 anos
Entre 61 e 65 anos
Entre 66 e 70 anos
Maior de 70 anos
Idade

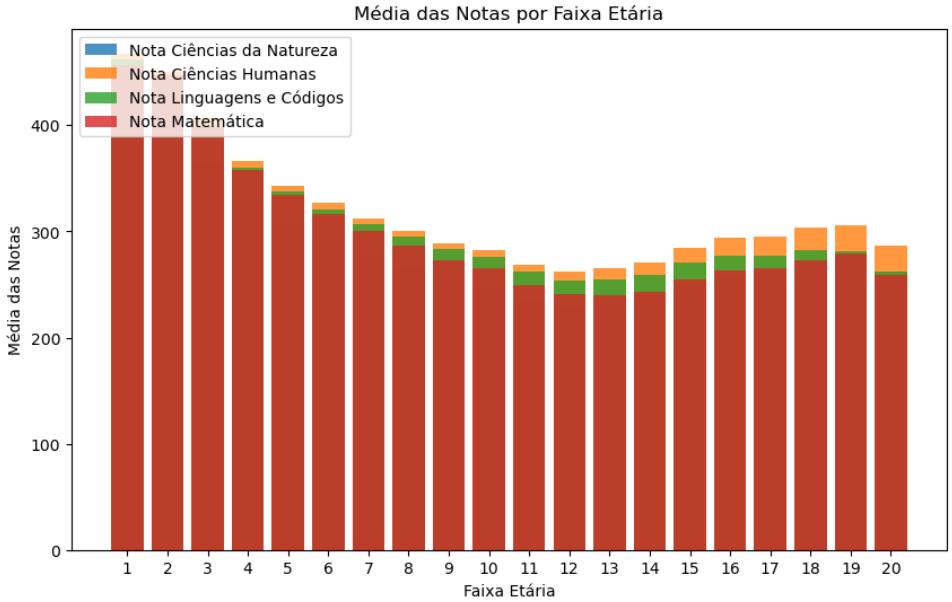

Idade vs. Nota de Nota_cien
Nota de Nota_cien
0
200
400
600
800
Menor de 17 anos
17 anos
18 anos
19 anos
20 anos
21 anos
22 anos
23 anos
24 anos
25 anos
Entre 26 e 30 anos
Entre 31 e 35 anos
Entre 36 e 40 anos
Entre 41 e 45 anos
Entre 46 e 50 anos
Entre 51 e 55 anos
Entre 56 e 60 anos
Entre 61 e 65 anos
Entre 66 e 70 anos
Maior de 70 anos
Idade

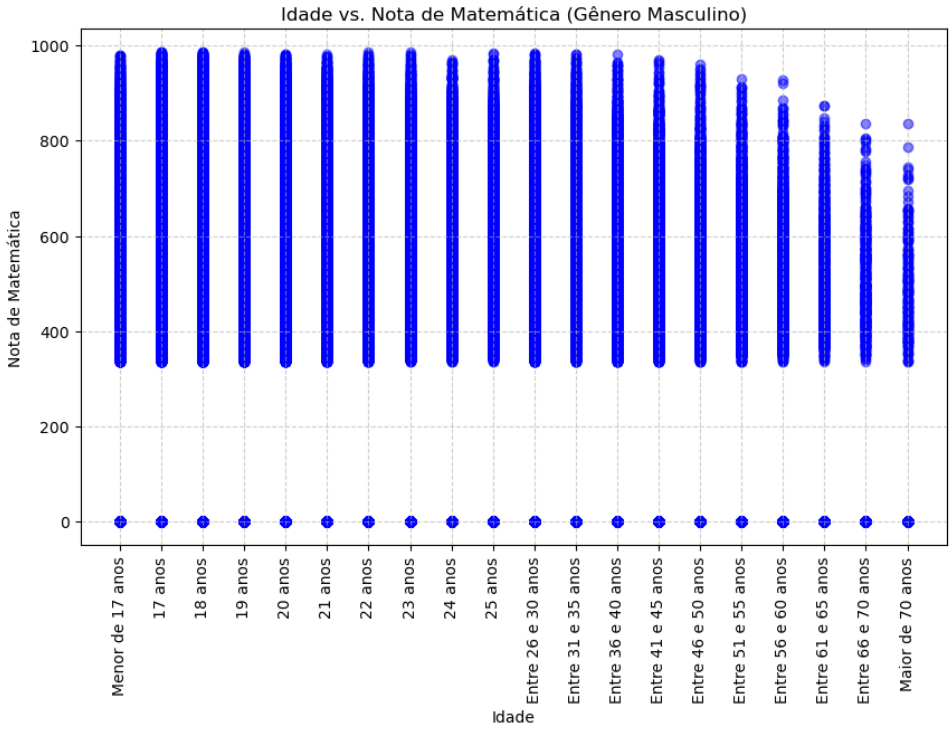

Idade vs. Nota de Nota_hum
Nota de Nota_hum
0
200
400
600
800
Menor de 17 anos
17 anos
18 anos
19 anos
20 anos
21 anos
22 anos
23 anos
24 anos
25 anos
Entre 26 e 30 anos
Entre 31 e 35 anos
Entre 36 e 40 anos
Entre 41 e 45 anos
Entre 46 e 50 anos
Entre 51 e 55 anos
Entre 56 e 60 anos
Entre 61 e 65 anos
Entre 66 e 70 anos
Maior de 70 anos
Idade

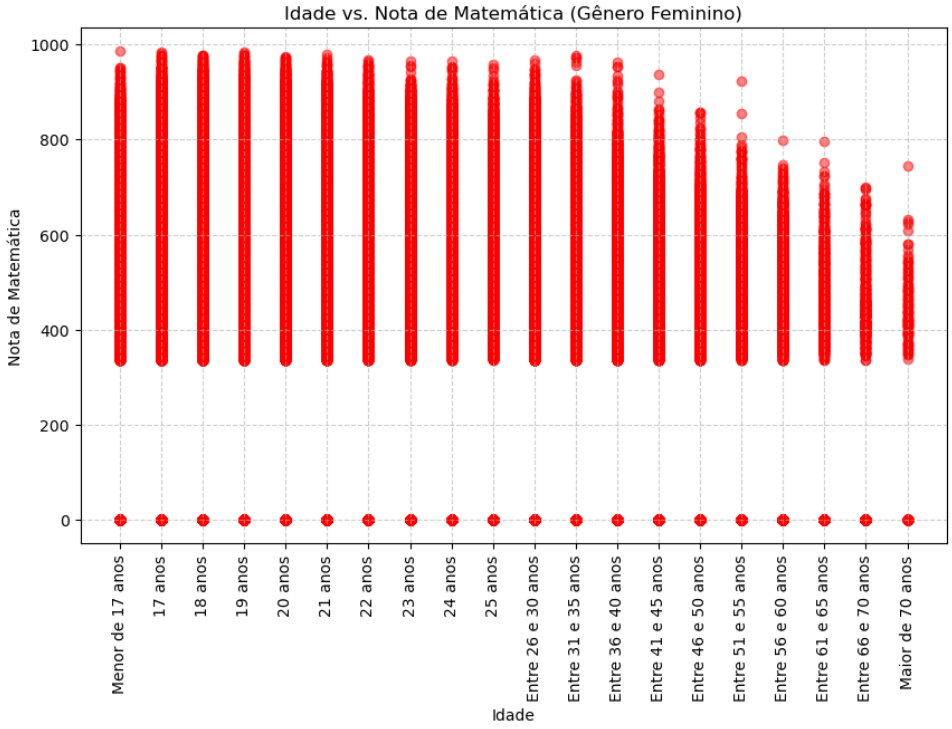

Idade vs. Nota de Nota_port
Nota de Nota_port
0
100
200
300
400
500
600
700
800
Menor de 17 anos
17 anos
18 anos
19 anos
20 anos
21 anos
22 anos
23 anos
24 anos
25 anos
Entre 26 e 30 anos
Entre 31 e 35 anos
Entre 36 e 40 anos
Entre 41 e 45 anos
Entre 46 e 50 anos
Entre 51 e 55 anos
Entre 56 e 60 anos
Entre 61 e 65 anos
Entre 66 e 70 anos
Maior de 70 anos
Idade

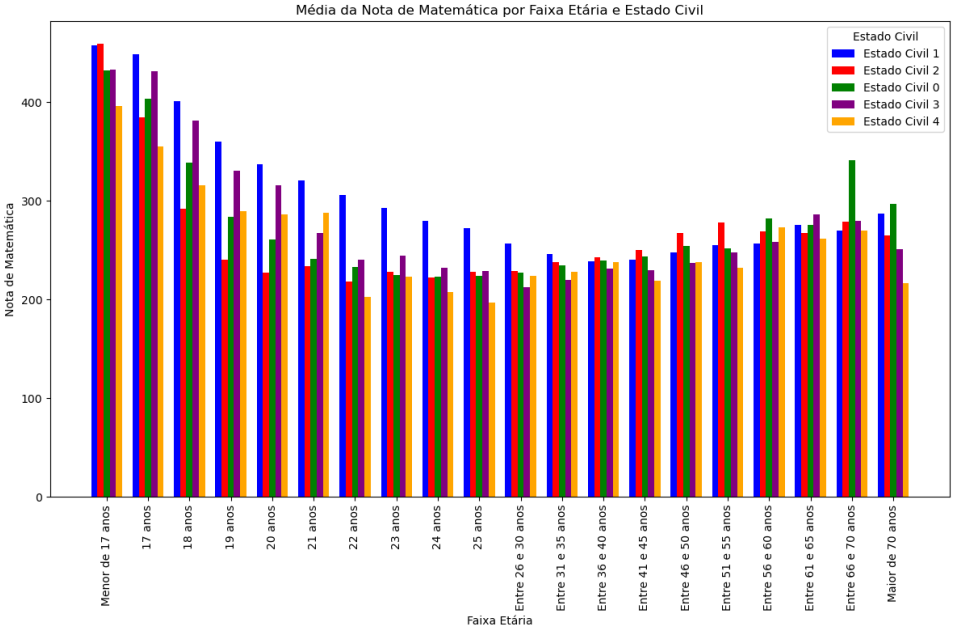


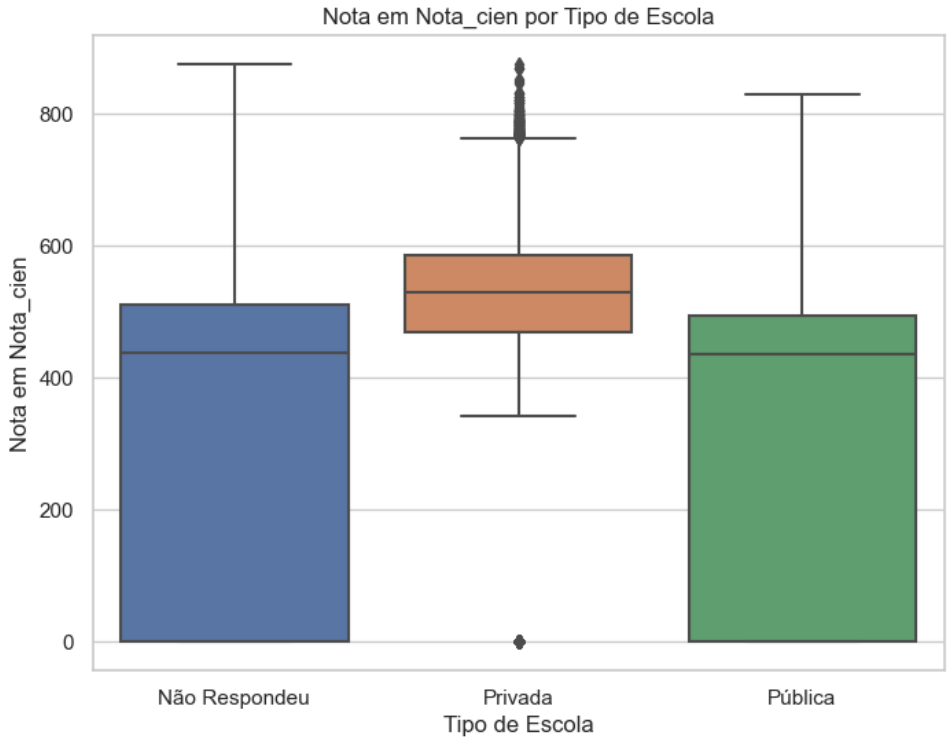
faixa_etaria	nota_cien	nota_hum	nota_port	nota_mat
1	406.050483	467.561186	462.019906	456.424195
2	403.489563	450.677897	443.549694	446.594407
3	364.067894	406.324298	399.393946	398.626376
4	327.923537	366.372597	360.295439	357.483394
5	306.897324	343.094103	337.246694	333.959275
6	291.615083	326.659415	320.695673	316.359706
7	277.202442	312.482504	306.859014	300.273913
8	265.452275	300.796026	294.950454	286.520541
9	253.759309	288.761717	283.170194	272.914088
10	247.165829	281.867980	276.411939	264.997999
11	233.972676	268.547210	262.405557	249.629451
12	228.273680	262.354406	253.954452	241.172146
13	230.184123	265.284198	254.595699	239.456795
14	236.644492	271.028606	258.818211	243.210073
15	249.079592	284.588791	270.183320	255.167741
16	256.576227	293.618617	276.926984	263.203601
17	261.287748	295.129088	276.604635	264.924092
18	268.151454	303.472491	282.426382	273.173330
19	277.857342	305.427442	281.544053	279.133023
20	262.014360	286.660554	262.119031	259.462976

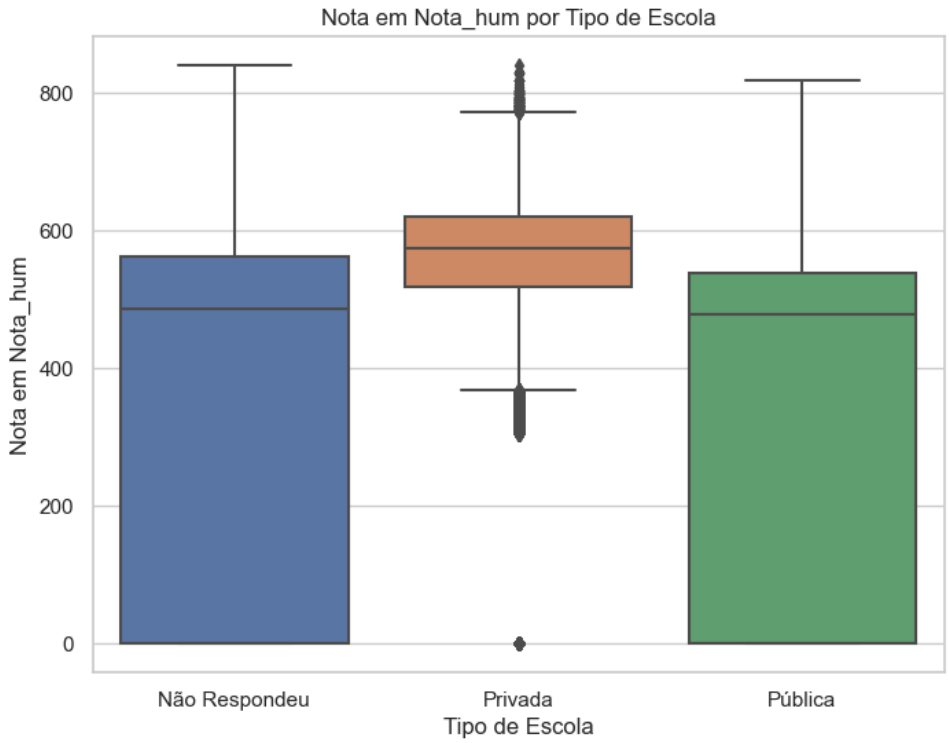

Média das Notas por Faixa Etária
Nota Ciências da Natureza
Nota Ciências Humanas
Nota Linguagens e Códigos
Nota Matemática
Média das Notas
0
100
200
300
400
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Faixa Etária

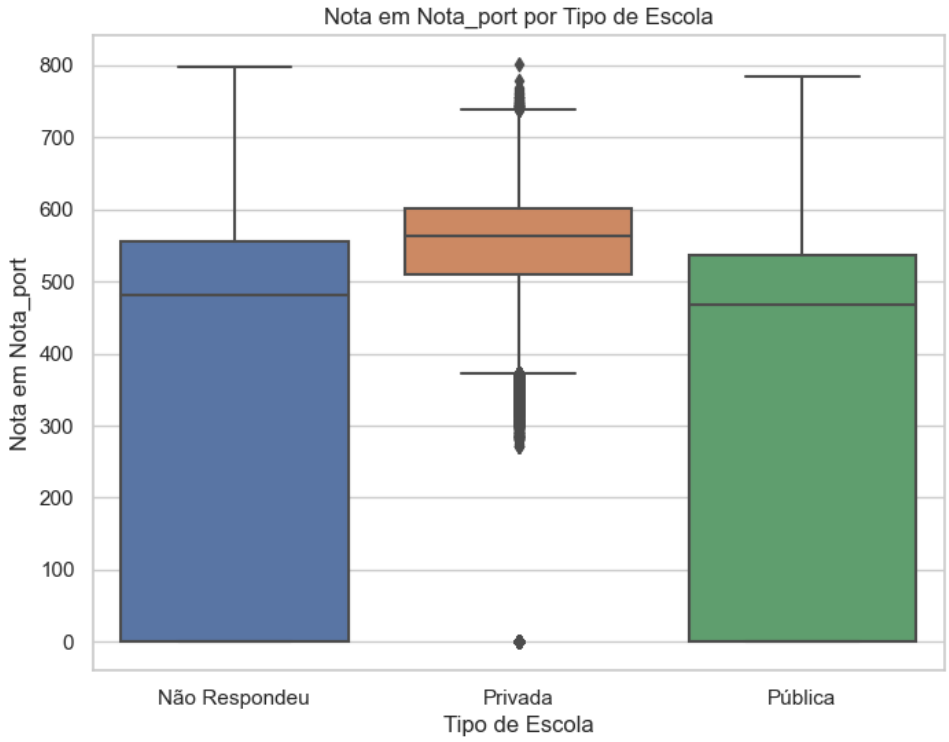

Idade vs. Nota de Matemática (Gênero Masculino)
Nota de Matemática
0
200
400
600
800
1000
Menor de 17 anos
17 anos
18 anos
19 anos
20 anos
21 anos
22 anos
23 anos
24 anos
25 anos
Entre 26 e 30 anos
Entre 31 e 35 anos
Entre 36 e 40 anos
Entre 41 e 45 anos
Entre 46 e 50 anos
Entre 51 e 55 anos
Entre 56 e 60 anos
Entre 61 e 65 anos
Entre 66 e 70 anos
Maior de 70 anos
Idade

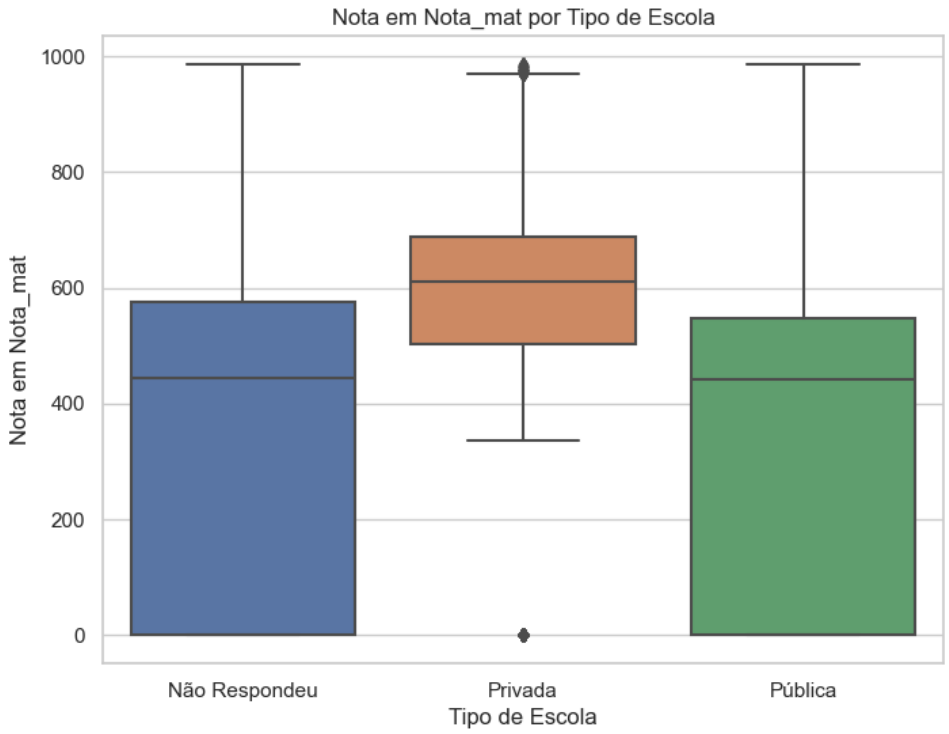

Idade vs. Nota de Matemática (Gênero Feminino)
Nota de Matemática
1000
800
600
400
200
0
Menor de 17 anos
17 anos
18 anos
19 anos
20 anos
21 anos
22 anos
23 anos
24 anos
25 anos
Entre 26 e 30 anos
Entre 31 e 35 anos
Entre 36 e 40 anos
Entre 41 e 45 anos
Entre 46 e 50 anos
Entre 51 e 55 anos
Entre 56 e 60 anos
Entre 61 e 65 anos
Entre 66 e 70 anos
Maior de 70 anos
Idade


Média da Nota de Matemática por Faixa Etária e Estado Civil
Estado Civil
Estado Civil 1
Estado Civil 2
Estado Civil 0
Estado Civil 3
Estado Civil 4
Nota de Matemática
400
300
200
100
0
Menor de 17 anos
17 anos
18 anos
19 anos
20 anos
21 anos
22 anos
23 anos
24 anos
25 anos
Entre 26 e 30 anos
Entre 31 e 35 anos
Entre 36 e 40 anos
Entre 41 e 45 anos
Entre 46 e 50 anos
Entre 51 e 55 anos
Entre 56 e 60 anos
Entre 61 e 65 anos
Entre 66 e 70 anos
Maior de 70 anos
Faixa Etária


Nota em Nota_cien por Tipo de Escola
Nota em Nota_cien
800
600
400
200
0
Não Respondeu
Privada
Pública
Tipo de Escola


Nota em Nota_hum por Tipo de Escola
Nota em Nota_hum
800
600
400
200
0
Não Respondeu
Privada
Pública
Tipo de Escola


Nota em Nota_port por Tipo de Escola
Nota em Nota_port
800
700
600
500
400
300
200
100
0
Não Respondeu
Privada
Pública
Tipo de Escola


Nota em Nota_mat por Tipo de Escola
Nota em Nota_mat
1000
800
600
400
200
0
Não Respondeu
Privada
Pública
Tipo de Escola

- Desempenho por Faixa Etária:

Em geral, parece haver uma tendência de queda nas notas à medida que a faixa etária aumenta. As notas são mais altas para os grupos mais jovens (faixas etárias 1 a 5) e tendem a diminuir conforme a idade aumenta.

- Variação nas Disciplinas:

As médias das notas variam entre as disciplinas. Por exemplo, as notas em Ciências Humanas e Linguagens e Códigos são geralmente mais altas em comparação com Ciências da Natureza e Matemática.

- Estabilidade das Notas:

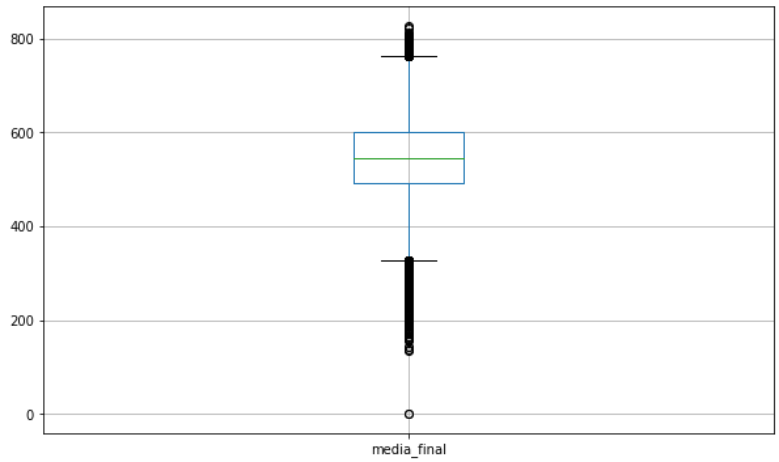
Observa-se uma relativa estabilidade nas notas de Ciências Humanas e Linguagens e Códigos em comparação com Ciências da Natureza e Matemática, que têm variações mais pronunciadas.

- Possível Correlação entre Idade e Desempenho:

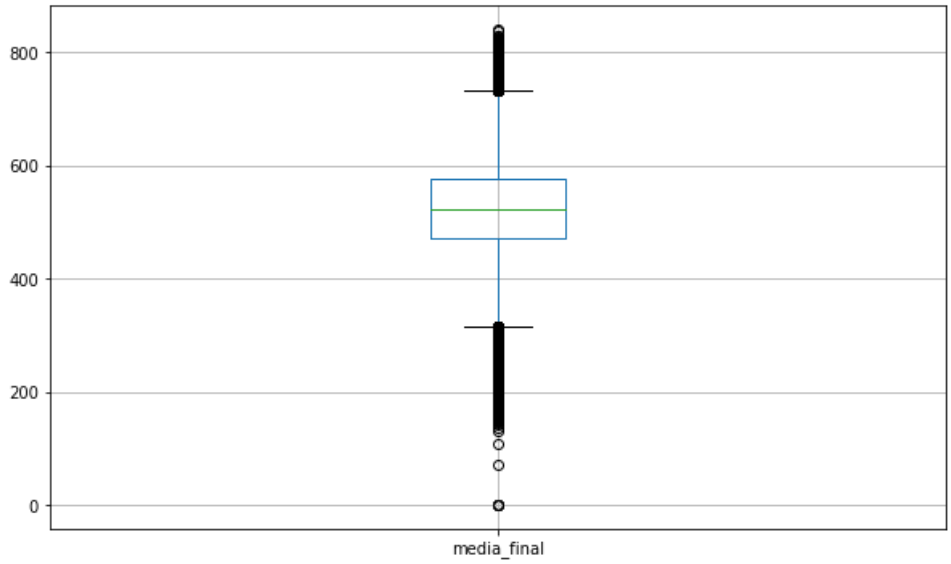
Como mencionado anteriormente, há uma tendência de queda nas notas com o aumento da idade. Isso pode ser influenciado por diversos fatores, como o tempo desde a conclusão do Ensino Médio, experiência prévia em áreas específicas, entre outros.

Essas são apenas algumas observações iniciais. É importante lembrar que essas são tendências gerais e podem haver exceções individuais. Para uma análise mais aprofundada, considerando outros fatores como histórico educacional, contexto socioeconômico, entre outros, seria necessário um estudo mais detalhado.

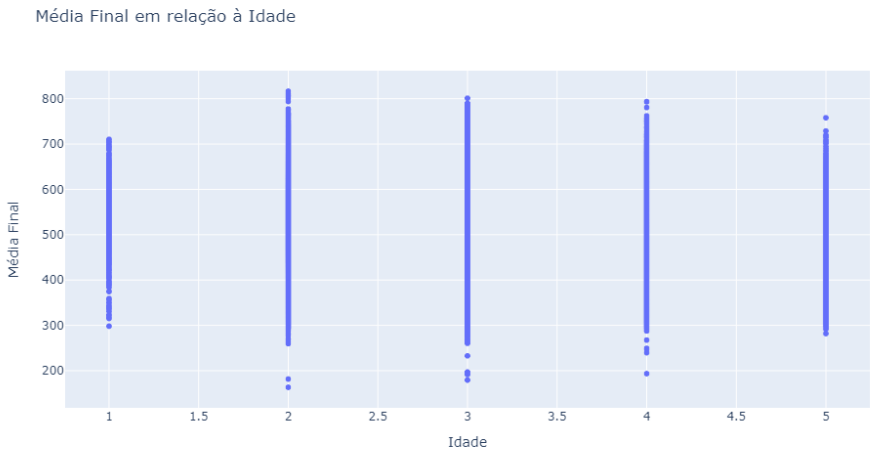
Segundo Grupo



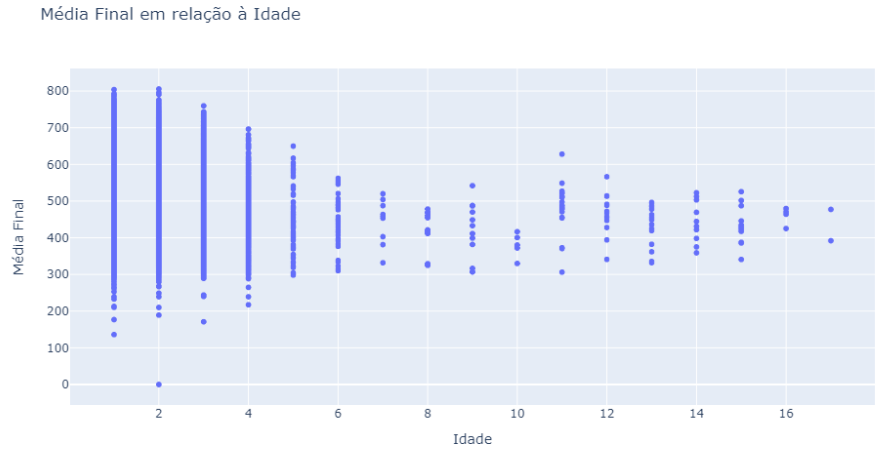
Com esse diagrama vemos que a media final fica em torno dos 540, o que seria uma nota mediana no enem. Vemos que temos valores acima dos 700, ultrapassando ate mesmo 800, o que podemos considerar como notas altas, ja que o enem vale 1000. Por outro lado, vemos muitas notas abaixo de 400. Sendo assim, conseguimos perceber uma grande diferença no desempenho dos alunos e nosso produto serviria exatamente para tentar nivelar o conhecimento de todos os alunos.



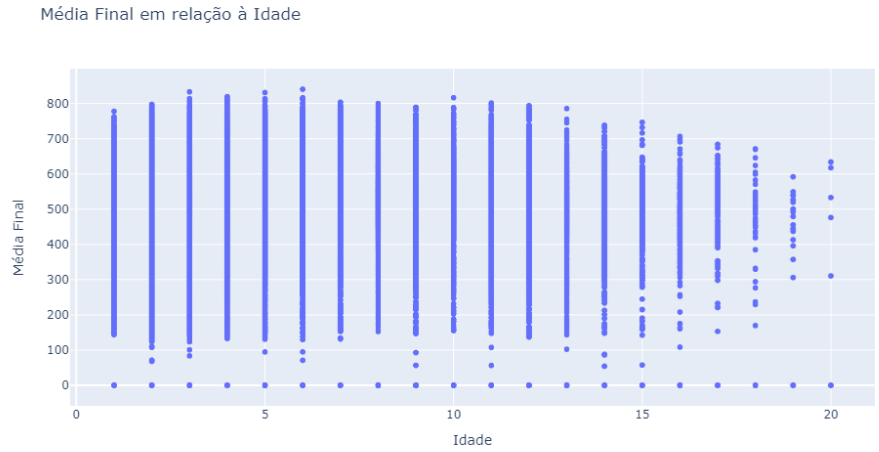
Vemos um comportamento bem semelhante nesse diagrama, mas a media das notas esta entre os 520, o que e normal ja que o numero de alunos e maior



Nesse grafico interativo vemos o comportamento das medias finais com a faixa etaria de 17 a 20 anos, que seriam os alunos que acabaram de se formar e que podem estar fazendo cursinho.

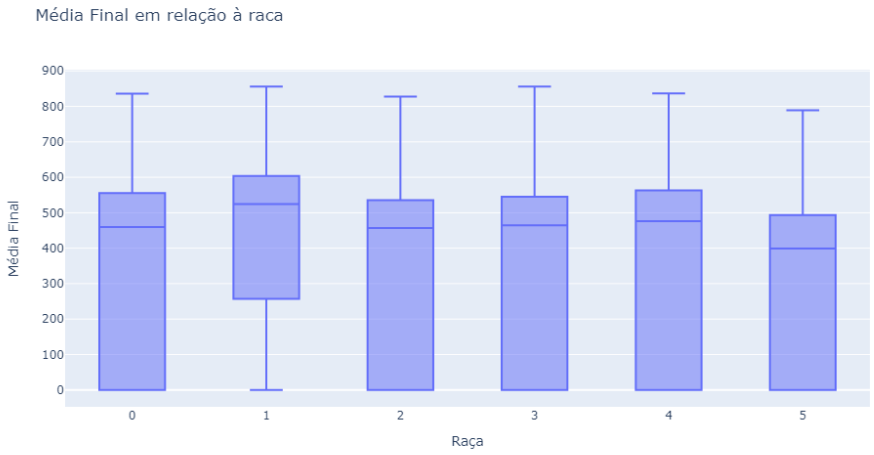


Aqui vemos o comportamento da media final dos treineiros, que selecionamos os menores de 17 anos, que seria a idade ideal de concluir a escola.



Vemos o comportamento das notas no geral, com todas as faixa etárias. Podemos perceber que a partir do grupo 15, as medias finais vão ficando mais concentradas em um intervalo e menos frequentes, o que acredito que indique menos pessoas realizando a prova.

Terceiro Grupo



Vemos uma clara desigualdade de notas em relação a raça branca dentre as demais, além de ter as maiores médias, o q1 da raça branca é 256 e o das demais é 0, o que significa que o top 25% de todas as outras raças tem uma média de 0, enquanto o da branca é 256. O valor mais alto da raça preta ta mais longe do que o valor mais alto da raça branca do que a média da raça branca ta em relação ao top 75% da raça preta.

- Uma conclusão de pelo menos 50 palavras, indicando quais outras explorações podem ser feitas no conjunto de dados

Acreditamos que muitas das exploraçoes possiveis foram feitas, algo que pode acontecer para ter uma nocao social melhor é explorar qual a quantidade de alunos dentro dos tipos de escola pela raça de cada um deles, tendo uma noção do tipo de raça da população de cada escola

Sessão 3: Resultados e Conclusões

Confirmamos nossa hipótese que o conhecimento se dissipa com o tempo e temos parâmetros que medem isso. Podemos afirmar que após uma década de um conhecimento em desuso seu aproveitamento nele cai aproximadamente 30%.
Provamos que matemática é a matéria que mais tem correlação com a média final. Além de ser a matéria com o pior desempenho. Então se está em dúvida entre números e letras vá para números a maioria das vezes.
Vemos um baixa clara na média final nas idades de 21 anos até aproximadamente 39, onde aparentemente a pessoa ainda acredita reter muito do conteúdo estudado, mas na verdade não, ou claro por muitos outros motivos também acabam indo mal. Algo surpreendente se mostrou nessa análise, em que pessoas idosas tem uma média da nota final maior do que pessoas de 21, 22 anos. Pode ser devido a muitos fatores claro, mas podemos afirmar que quem é experiente na vida também estuda.
Infelizmente a diferença de ensino entre escolas públicas e privadas é gritante, onde apenas estar na média em uma escola privada você já está acima do top75% de escolas públicas. A diferença entre os tipos de ensino também existe claro, o ensino regular tem uma média de aproximadamente 100 pontos a mais sobre o ensino especial.
Sem contar também sobre a diferença de notas entre raças onde o comportamento da relação privada x publica acontece também na relação de notas entre pessoas brancas x pretas.
Após toda essa análise garantimos que nós da Sons of Node temos um produto excelente para mudar a educação do Brasil e quem sabe do mundo.

- **PROXÍMOS PASSOS**

Acreditamos que os proximos passos sejam implementar a analise de dados e grafica com dados sobre o ensino medio em si, por exemplo alguma taxa de aprovação. Enquanto desenvolvemos nossa prova para todas as matérias do Enem. Após isso devemos perseguir algo mais focado nos anos do ensino médio e no primeiro ano após dele.

Referências

1- https://www.techtudo.com.br/noticias/2014/01/sete-ideias-do-facebook-que-viraram-produtos-em-hackathons-de-2013.ghtml
2- https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem