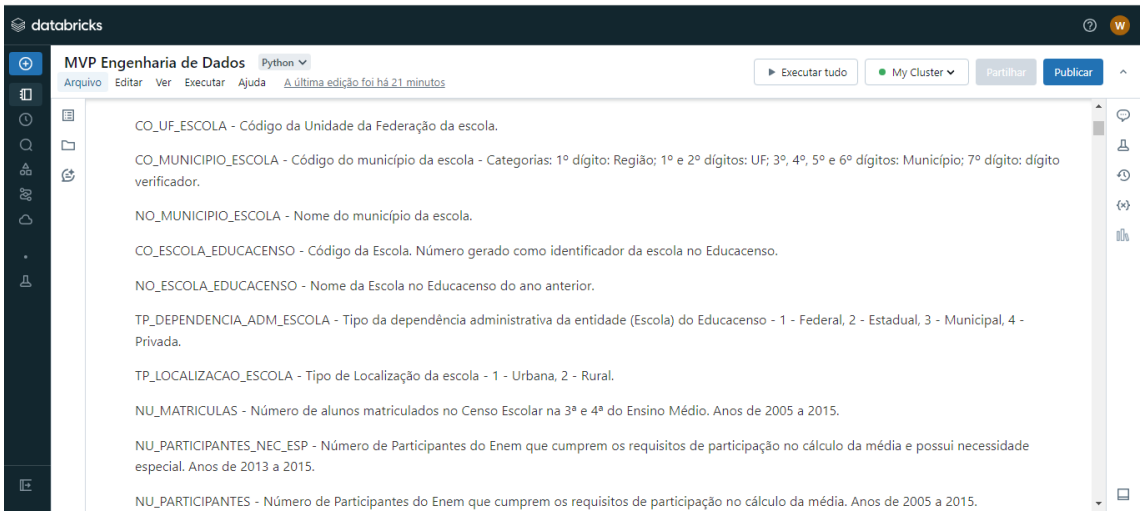
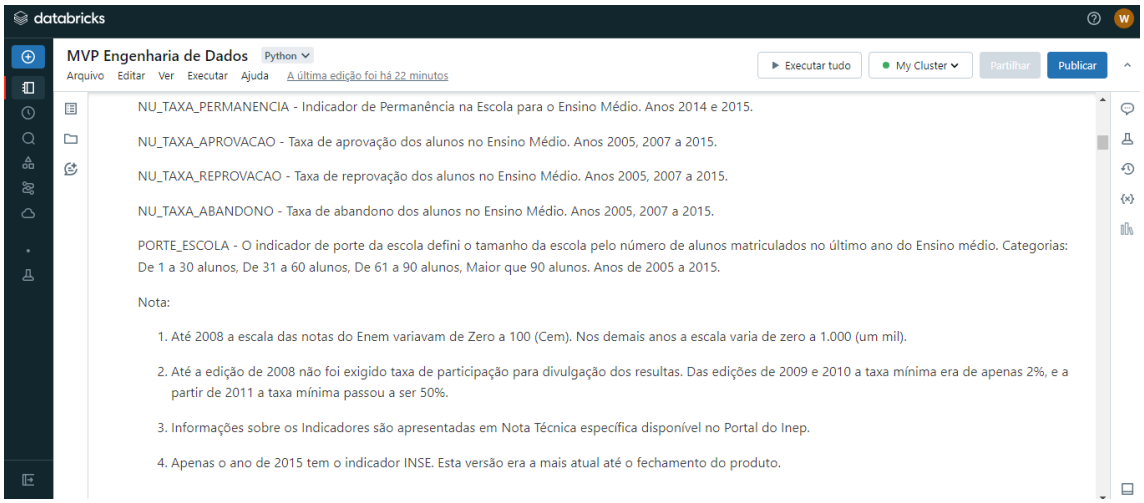
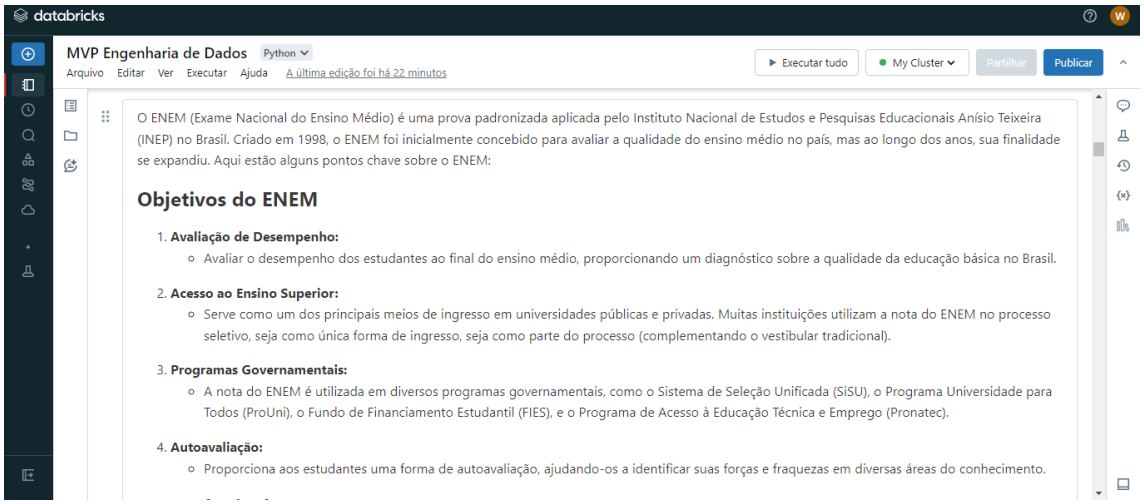
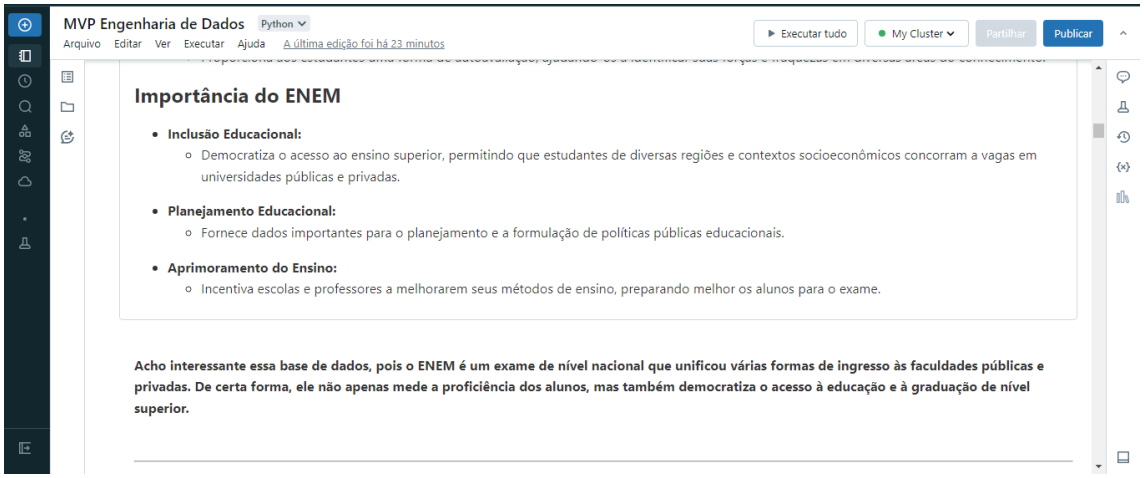

databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 21 minutos
Executar tudo
My Cluster
Partilhar
Publicar
PUC RJ - MPV - Sprint: Engenharia de Dados
Wallace Euzébio dos Santos
Objetivo: Analisar o desempenho de diferentes escolas na prova do ENEM de acordo com dados históricos.
O dataset deste trabalho foi retirado do site pertencente ao Governo Federal, com informações do intervalo de anos de 2005 a 2015, material disponível no caminho: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem-por-escola
Catálogo de dados:
NU_ANO - Ano da edição do ENEM por Escola.
CO_UF_ESCOLA - Código da Unidade da Federação da escola.
CO_MUNICIPIO_ESCOLA - Código do município da escola - Categorias: 1º dígito: Região; 1º e 2º dígitos: UF; 3º, 4º, 5º e 6º dígitos: Município; 7º dígito: dígito

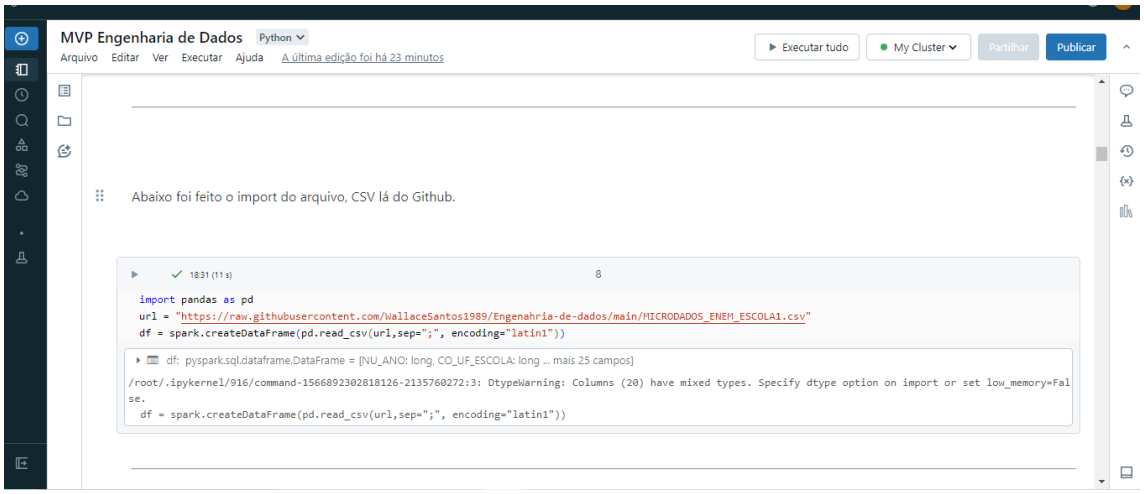

databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 21 minutos
Executar tudo
My Cluster
Partilhar
Publicar
CO_UF_ESCOLA - Código da Unidade da Federação da escola.
CO_MUNICIPIO_ESCOLA - Código do município da escola - Categorias: 1º dígito: Região; 1º e 2º dígitos: UF; 3º, 4º, 5º e 6º dígitos: Município; 7º dígito: dígito verificador.
NO_MUNICIPIO_ESCOLA - Nome do município da escola.
CO_ESCOLA_EDUCACENSO - Código da Escola. Número gerado como identificador da escola no Educacenso.
NO_ESCOLA_EDUCACENSO - Nome da Escola no Educacenso do ano anterior.
TP_DEPENDENCIA_ADM_ESCOLA - Tipo da dependência administrativa da entidade (Escola) do Educacenso - 1 - Federal, 2 - Estadual, 3 - Municipal, 4 - Privada.
TP_LOCALIZACAO_ESCOLA - Tipo de Localização da escola - 1 - Urbana, 2 - Rural.
NU_MATRICULAS - Número de alunos matriculados no Censo Escolar na 3ª e 4ª do Ensino Médio. Anos de 2005 a 2015.
NU_PARTICIPANTES_NEC_ESP - Número de Participantes do Enem que cumprem os requisitos de participação no cálculo da média e possui necessidade especial. Anos de 2013 a 2015.
NU_PARTICIPANTES - Número de Participantes do Enem que cumprem os requisitos de participação no cálculo da média. Anos de 2005 a 2015.

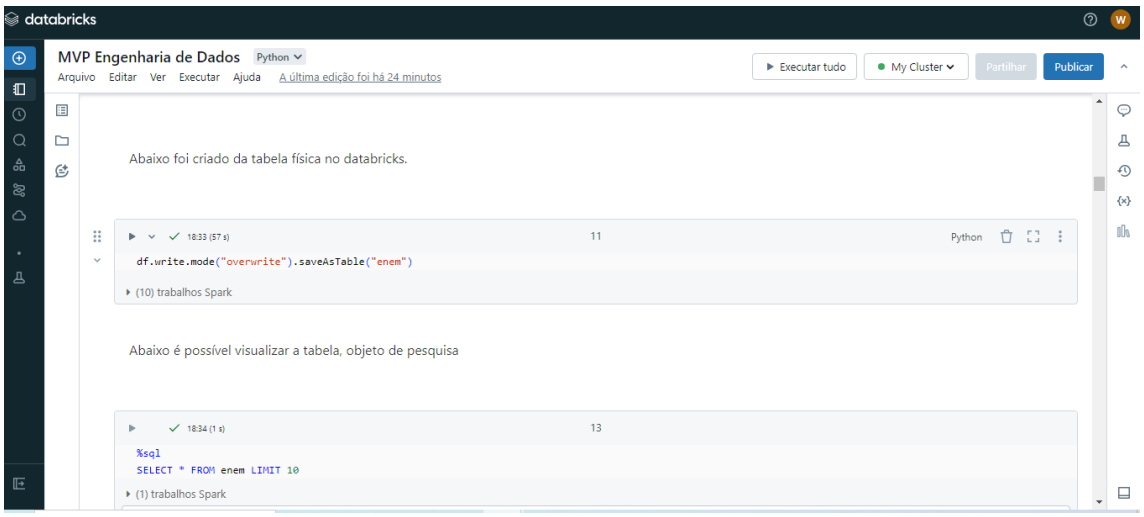

databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 21 minutos
Executar tudo
My Cluster
Partilhar
Publicar
NU_TAXA_PARTICIPACAO - A taxa de participação é o número total de participantes no ENEM da escola dividido pelo número de matriculados nos anos finais do Ensino Médio da escola, informado no Censo Escolar. Anos de 2009 a 2015.
NU_MEDIA_CN - Média das notas de Ciências da Natureza do Ensino Médio Regular. Anos de 2009 a 2015.
NU_MEDIA_CH - Média das notas de Ciências Humanas do Ensino Médio Regular. Anos de 2009 a 2015.
NU_MEDIA_LP - Média das notas de Linguagens e Códigos do Ensino Médio Regular. Anos de 2009 a 2015.
NU_MEDIA_MT - Média das notas de Matemática do Ensino Médio Regular. Anos de 2009 a 2015.
NU_MEDIA_RED - Média das notas de Redação do Ensino Médio Regular. Anos de 2008 a 2015.
NU_MEDIA_OBJ - Média da prova objetiva do Ensino Médio Regular. Ano 2008.
NU_MEDIA_TOT - Média Total (Redação e Prova Objetiva) Ensino Médio Regular corrigida pelo número de participantes. Anos 2005 a 2007.
INSE - Indicador de Nível Socioeconômico da escola - Para melhor descrever o nível socioeconômico das escolas, foram criados seis grupos, de forma que o Grupo 1 congrega as escolas com Inse médio mais baixo e o Grupo 6, com mais alto. (Edição 2015) – Categorias: Grupo 1, Grupo 2, Grupo 3, Grupo 4, Grupo 5, Grupo 6. Ano 2015.
PC_FORMACAO_DOCENTE - Indicador de Adequação da Formação Docente da escola para lecionar no Ensino Médio. Anos 2013 a 2015.


databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 22 minutos
Executar tudo
My Cluster
Partilhar
Publicar
NU_TAXA_PERMANENCIA - Indicador de Permanência na Escola para o Ensino Médio. Anos 2014 e 2015.
NU_TAXA_APROVACAO - Taxa de aprovação dos alunos no Ensino Médio. Anos 2005, 2007 a 2015.
NU_TAXA_REPROVACAO - Taxa de reprovação dos alunos no Ensino Médio. Anos 2005, 2007 a 2015.
NU_TAXA_ABANDONO - Taxa de abandono dos alunos no Ensino Médio. Anos 2005, 2007 a 2015.
PORTE_ESCOLA - O indicador de porte da escola defini o tamanho da escola pelo número de alunos matriculados no último ano do Ensino médio. Categorias: De 1 a 30 alunos, De 31 a 60 alunos, De 61 a 90 alunos, Maior que 90 alunos. Anos de 2005 a 2015.
Nota:
1. Até 2008 a escala das notas do Enem variavam de Zero a 100 (Cem). Nos demais anos a escala varia de zero a 1.000 (um mil).
2. Até a edição de 2008 não foi exigido taxa de participação para divulgação dos resultas. Das edições de 2009 e 2010 a taxa mínima era de apenas 2%, e a partir de 2011 a taxa mínima passou a ser 50%.
3. Informações sobre os Indicadores são apresentadas em Nota Técnica específica disponível no Portal do Inep.
4. Apenas o ano de 2015 tem o indicador INSE. Esta versão era a mais atual até o fechamento do produto.

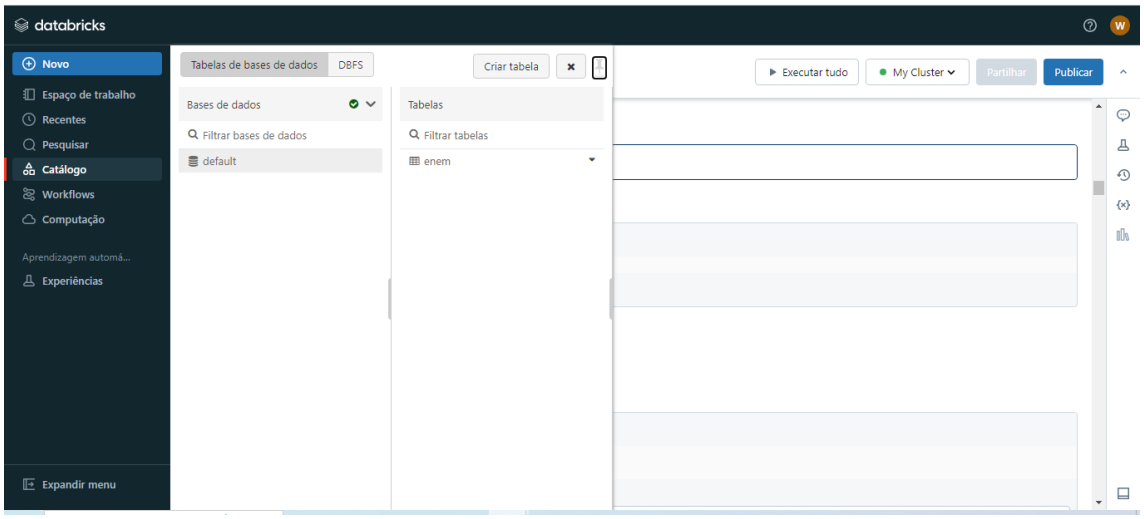

databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 22 minutos
Executar tudo
My Cluster
Partilhar
Publicar
O ENEM (Exame Nacional do Ensino Médio) é uma prova padronizada aplicada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) no Brasil. Criado em 1998, o ENEM foi inicialmente concebido para avaliar a qualidade do ensino médio no país, mas ao longo dos anos, sua finalidade se expandiu. Aqui estão alguns pontos chave sobre o ENEM:
Objetivos do ENEM
1. Avaliação de Desempenho:
o Avaliar o desempenho dos estudantes ao final do ensino médio, proporcionando um diagnóstico sobre a qualidade da educação básica no Brasil.
2. Acesso ao Ensino Superior:
o Serve como um dos principais meios de ingresso em universidades públicas e privadas. Muitas instituições utilizam a nota do ENEM no processo seletivo, seja como única forma de ingresso, seja como parte do processo (complementando o vestibular tradicional).
3. Programas Governamentais:
o A nota do ENEM é utilizada em diversos programas governamentais, como o Sistema de Seleção Unificada (SiSU), o Programa Universidade para Todos (ProUni), o Fundo de Financiamento Estudantil (FIES), e o Programa de Acesso à Educação Técnica e Emprego (Pronatec).
4. Autoavaliação:
o Proporciona aos estudantes uma forma de autoavaliação, ajudando-os a identificar suas forças e fraquezas em diversas áreas do conhecimento.

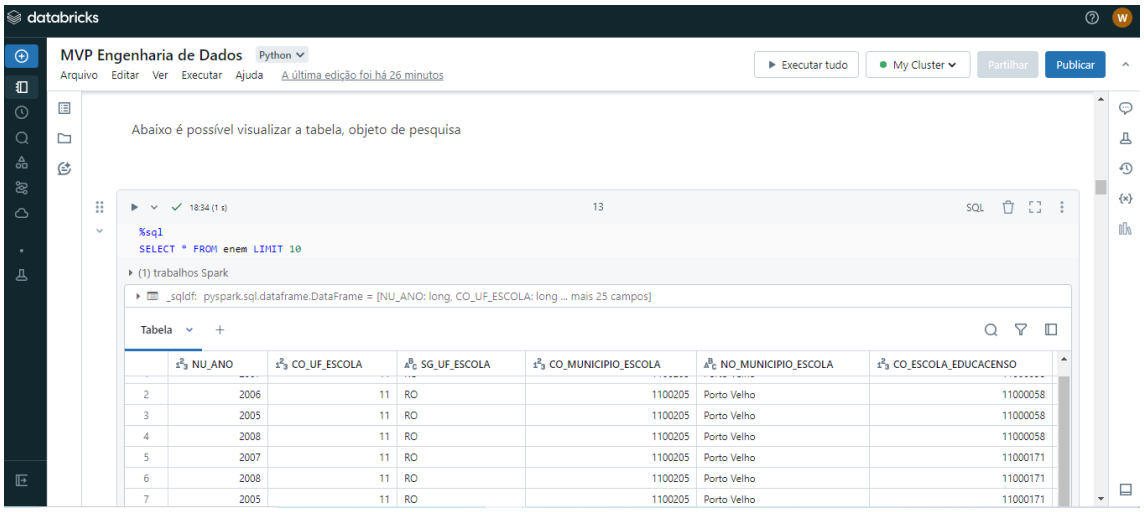

MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 23 minutos
Executar tudo
My Cluster
Partilhar
Publicar
Importância do ENEM
• Inclusão Educacional:
o Democratiza o acesso ao ensino superior, permitindo que estudantes de diversas regiões e contextos socioeconômicos concorram a vagas em universidades públicas e privadas.
• Planejamento Educacional:
o Fornece dados importantes para o planejamento e a formulação de políticas públicas educacionais.
• Aprimoramento do Ensino:
o Incentiva escolas e professores a melhorarem seus métodos de ensino, preparando melhor os alunos para o exame.
Acho interessante essa base de dados, pois o ENEM é um exame de nível nacional que unificou várias formas de ingresso às faculdades públicas e privadas. De certa forma, ele não apenas mede a proficiência dos alunos, mas também democratiza o acesso à educação e à graduação de nível superior.

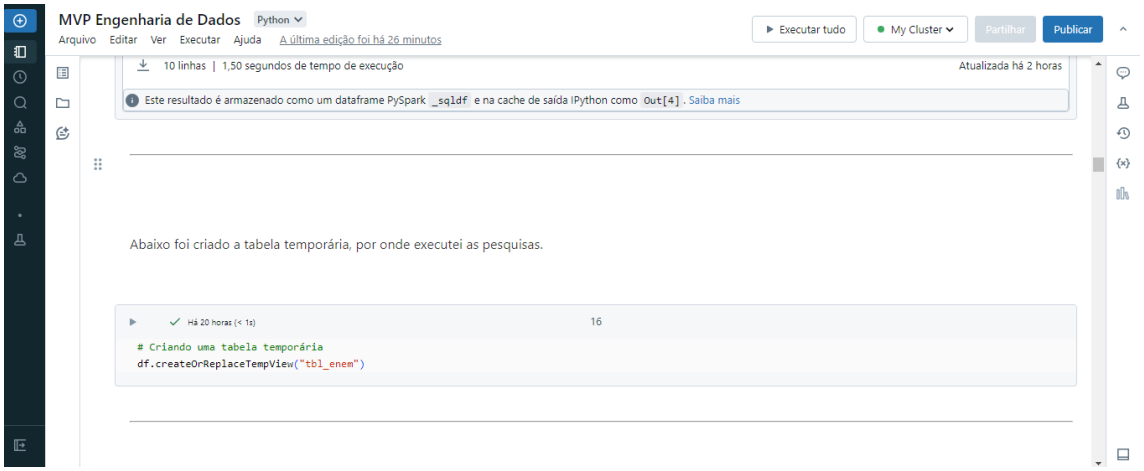




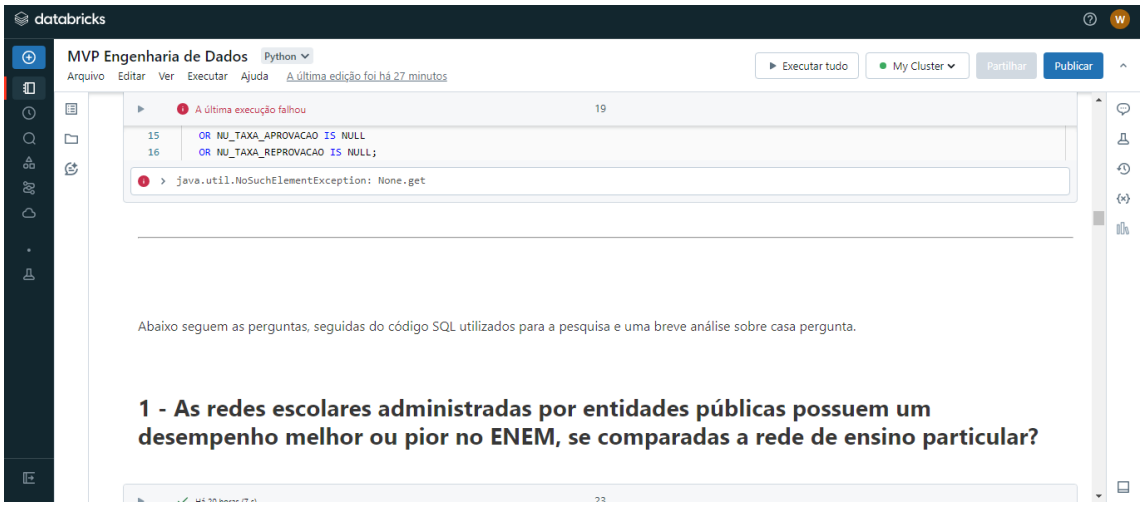
Criação da tabela física:




databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 26 minutos
Executar tudo
My Cluster
Partilhar
Publicar
Abaixo é possível visualizar a tabela, objeto de pesquisa
18:34 (1 s)
13
SQL
%sql
SELECT * FROM enem LIMIT 10
(1) trabalhos Spark
_sqldf: pyspark.sql.dataframe.DataFrame = [NU_ANO: long, CO_UF_ESCOLA: long ... mais 25 campos]
Tabela
NU_ANO
CO_UF_ESCOLA
SG_UF_ESCOLA
CO_MUNICIPIO_ESCOLA
NO_MUNICIPIO_ESCOLA
CO_ESCOLA_EDUCACENSO
2 2006 11 RO 1100205 Porto Velho 11000058
3 2005 11 RO 1100205 Porto Velho 11000058
4 2008 11 RO 1100205 Porto Velho 11000058
5 2007 11 RO 1100205 Porto Velho 11000171
6 2008 11 RO 1100205 Porto Velho 11000171
7 2005 11 RO 1100205 Porto Velho 11000171

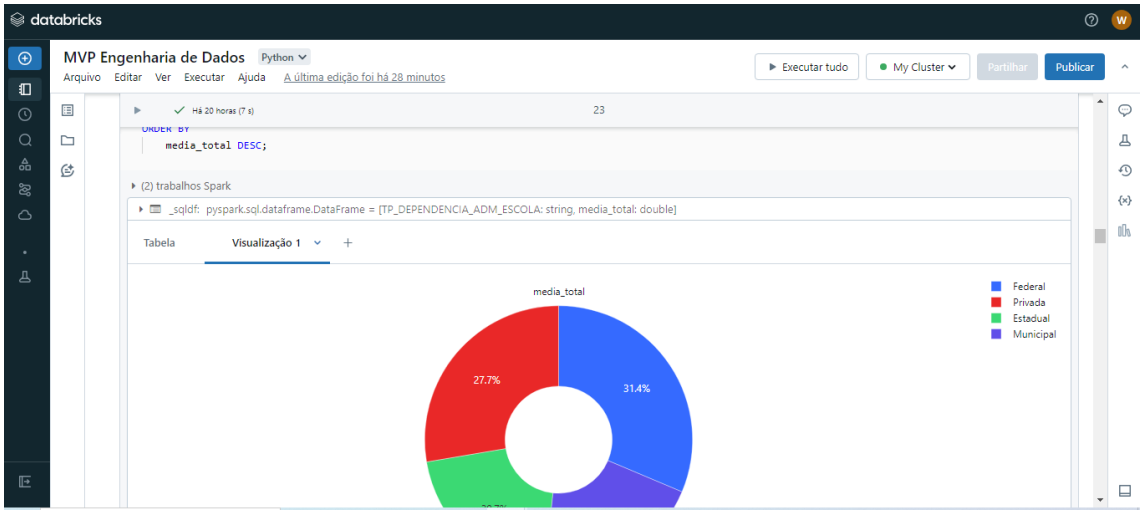

MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 26 minutos
Executar tudo
My Cluster
Partilhar
Publicar
10 linhas | 1,50 segundos de tempo de execução
Atualizada há 2 horas
Este resultado é armazenado como um dataframe PySpark _sqldf e na cache de saída IPython como Out[4] . Saiba mais
Abaixo foi criado a tabela temporária, por onde executei as pesquisas.
Há 20 horas (< 1s)
16
# Criando uma tabela temporária
df.createOrReplaceTempView("tbl_enem")

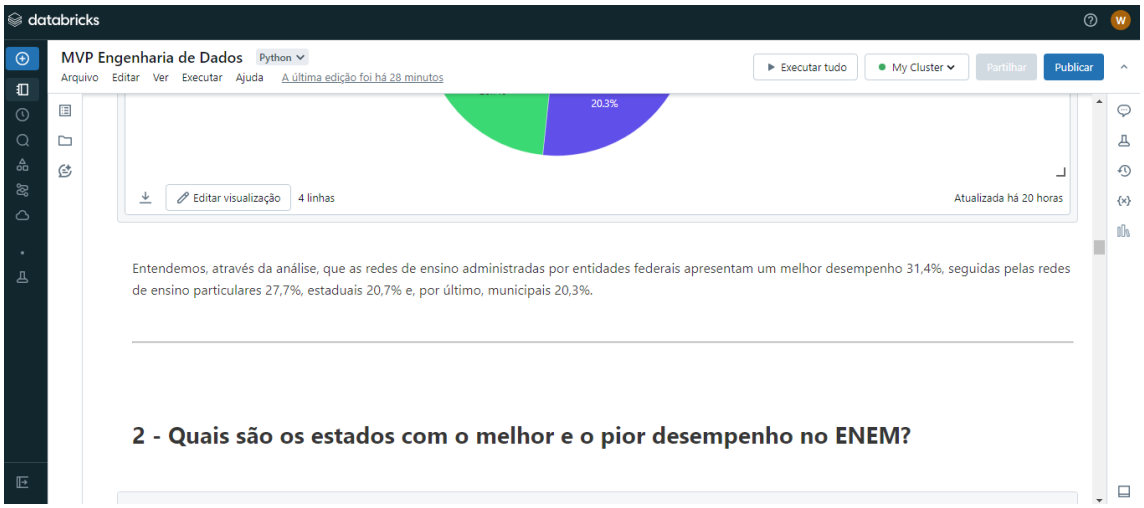

MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 27 minutos
Executar tudo
My Cluster
Partilhar
Publicar
Tentei normalizar os dados no código abaixo retirando as linhas nulas, no entanto alguns dos atributos não possuem dados históricos, como é o caso da média total no Enem que só possui informações nos anos de 2005, 2006 e 2007, então retirar as linhas nulas prejuducaria a conteúdo do dataset de forma geral.
A última execução falhou
19
%sql
DELETE FROM tbl_enem
WHERE NU_PARTICIPANTES_NEC_ESP IS NULL
OR NU_TAXA_PARTICIPACAO IS NULL
OR NU_MEDIA_CN IS NULL
OR NU_MEDIA_CH IS NULL
OR NU_MEDIA_LP IS NULL
OR NU_MEDIA_MT IS NULL
OR NU_MEDIA_RED IS NULL
OR NU_MEDIA_OBJ IS NULL
OR NU_MEDIA_TOT IS NULL



Abaixo seguem as perguntas, seguidas do código SQL utilizados para a pesquisa e uma breve análise sobre casa pergunta.

1 - As redes escolares administradas por entidades públicas possuem um desempenho melhor ou pior no ENEM, se comparadas a rede de ensino particular?

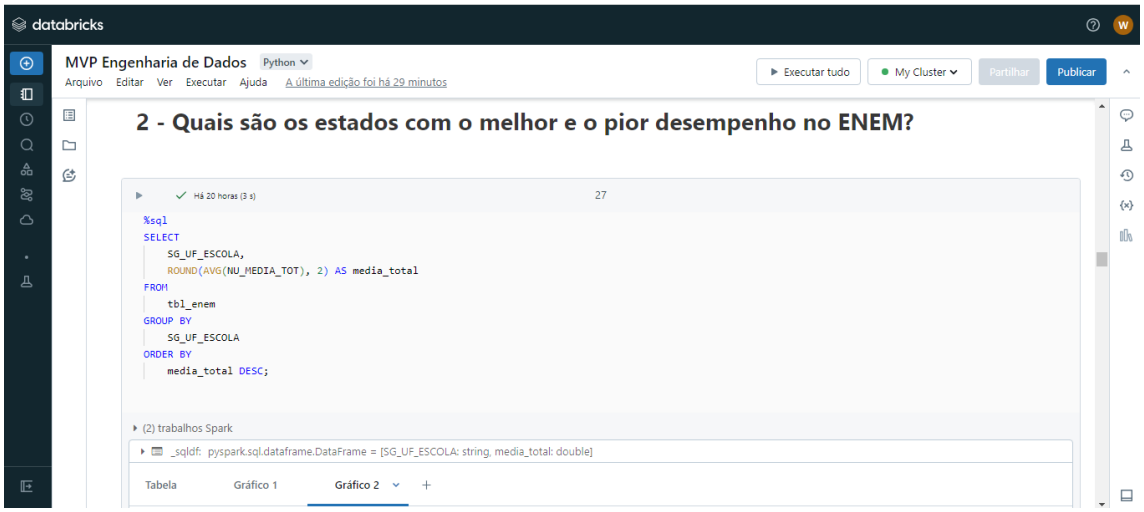




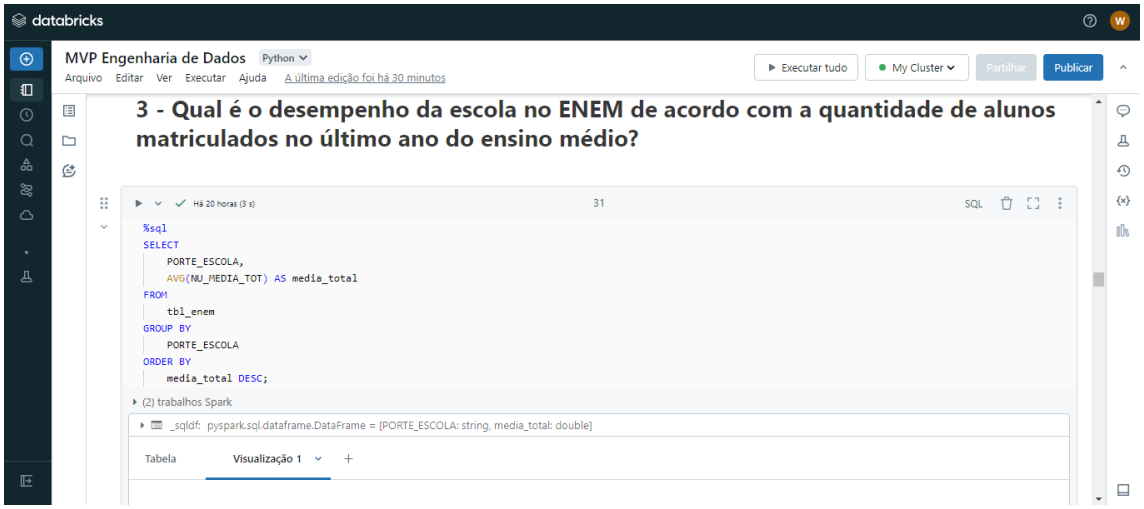
Entendemos, através da análise, que as redes de ensino administradas por entidades federais apresentam um melhor desempenho 31,4%, seguidas pelas redes de ensino particulares 27,7%, estaduais 20,7% e, por último, municipais 20,3%.

---

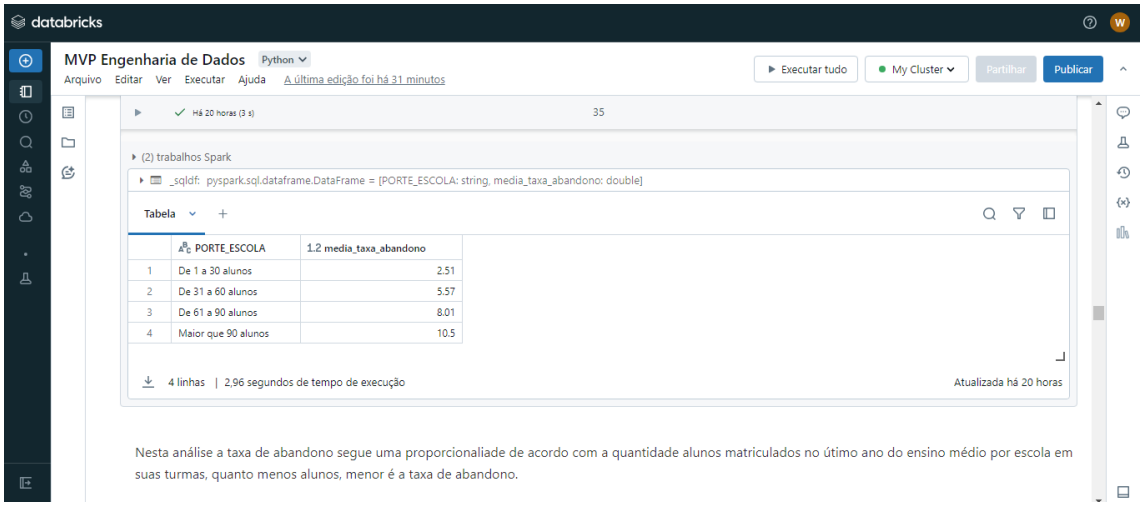
2 - Quais são os estados com o melhor e o pior desempenho no ENEM?



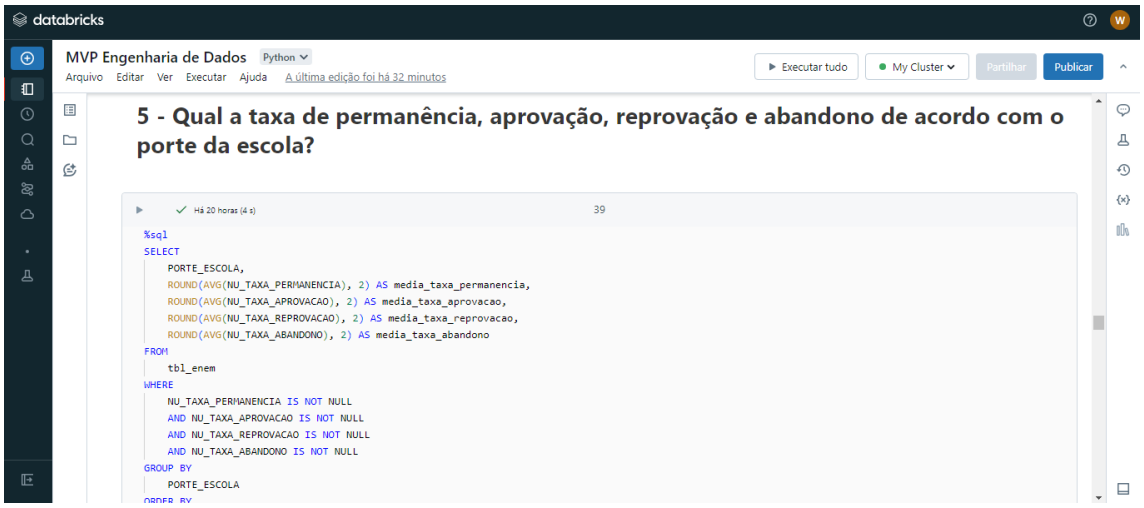
Em ordem de pela média total de proficiência, temos Distrito Federal, Rio Grande do Sul, Espírito Santo, Santa Catarina e Minas Gerais. O que significa que não necessariamente possa existir uma correlação geográfica com o desempenho, mas com políticas públicas e incentivos específicos de cada estado e o Distrito Federal que ocupa a primeira posição nesta análise.

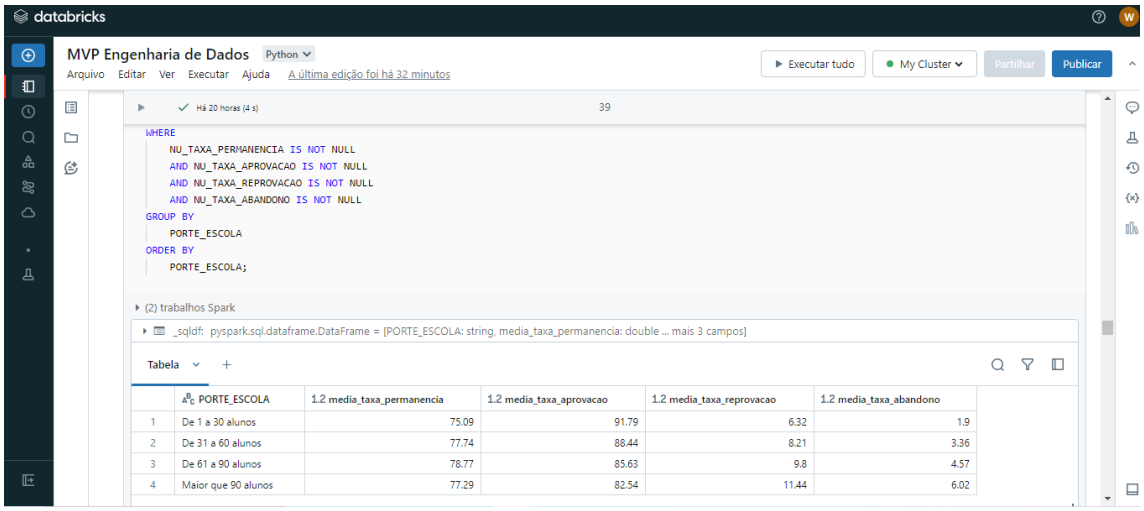


Pode existir uma relação entre quantidade de alunos por turma e a qualidade do desempenho, segundo a análise feita, quanto menor são as turmas, melhor é o desempenho no ENEM.



Nesta análise a taxa de abandono segue uma proporcionaliade de acordo com a quantidade alunos matriculados no útimo ano do ensino médio por escola em suas turmas, quanto menos alunos, menor é a taxa de abandono.





	PORTE_ESCOLA	media_taxa_permanencia	media_taxa_aprovacao	media_taxa_reprovacao	media_taxa_abandono
1	De 1 a 30 alunos	75.09	91.79	6.32	1.9
2	De 31 a 60 alunos	77.74	88.44	8.21	3.36
3	De 61 a 90 alunos	78.77	85.63	9.8	4.57
4	Maior que 90 alunos	77.29	82.54	11.44	6.02

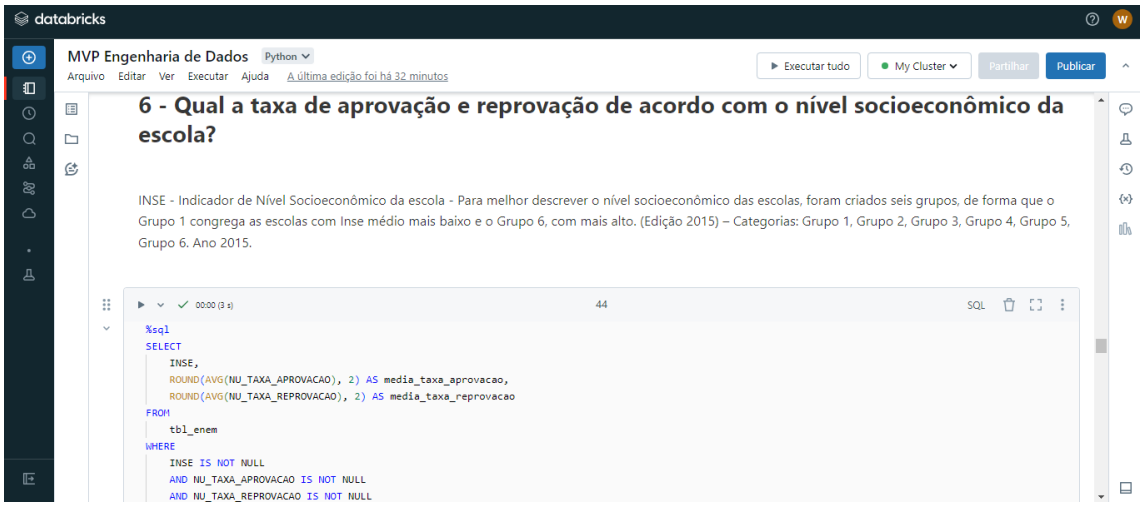
	PORTE_ESCOLA	media_taxa_permanencia	media_taxa_aprovacao	media_taxa_reprovacao	media_taxa_abandono
1	De 1 a 30 alunos	75.09	91.79	6.32	1.9
2	De 31 a 60 alunos	77.74	88.44	8.21	3.36
3	De 61 a 90 alunos	78.77	85.63	9.8	4.57
4	Maior que 90 alunos	77.29	82.54	11.44	6.02

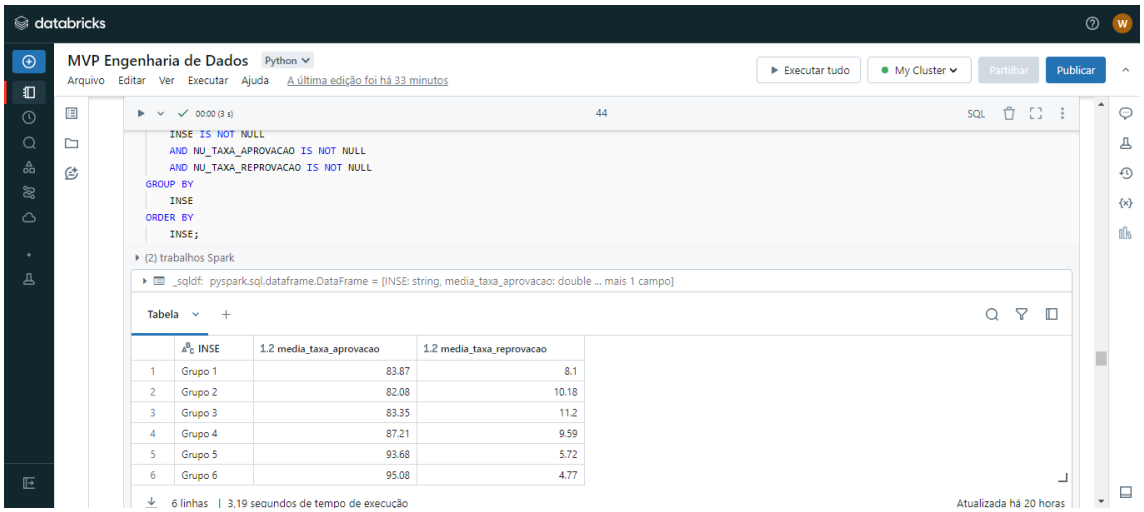
4 linhas | 4,34 segundos de tempo de execução Atualizada há 20 horas

É possível fazer análises importantes comparando estes dados, turmas com menos alunos possuem uma taxa de permanêcia menor se comparado aos demais, no entanto a taxa de aprovação é maior, a taxa de reprovação menor eataxa de abandono também é a menor de todos os dados analisados.

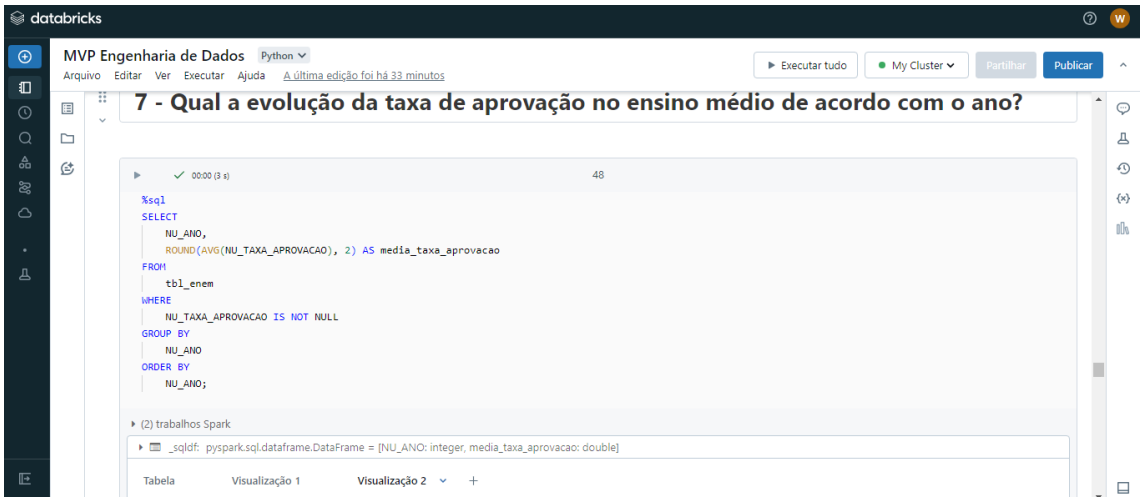
A taxa de permanência varia pouco de acordo com a quantidade de alunos.

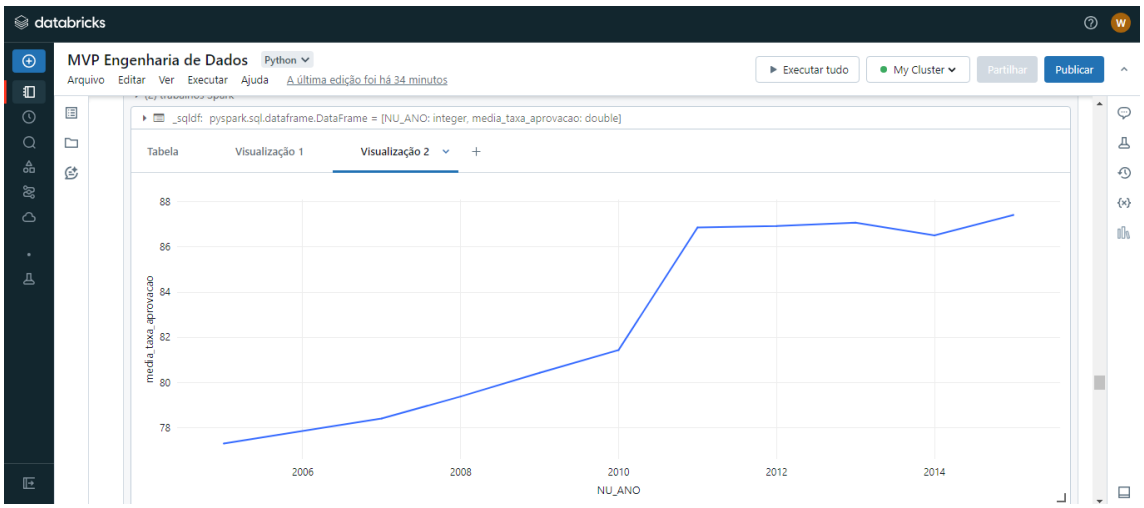
A taxa de reprovação e abandono é maior quando a quantidade de alunos é maior.


databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 32 minutos
Executar tudo
My Cluster
Partilhar
Publicar
6 - Qual a taxa de aprovação e reprovação de acordo com o nível socioeconômico da escola?
INSE - Indicador de Nível Socioeconômico da escola - Para melhor descrever o nível socioeconômico das escolas, foram criados seis grupos, de forma que o Grupo 1 congrega as escolas com Inse médio mais baixo e o Grupo 6, com mais alto. (Edição 2015) – Categorias: Grupo 1, Grupo 2, Grupo 3, Grupo 4, Grupo 5, Grupo 6. Ano 2015.
00:00 (3 s)
44
SQL
%sql
SELECT
INSE,
ROUND(AVG(NU_TAXA_APROVACAO), 2) AS media_taxa_aprovacao,
ROUND(AVG(NU_TAXA_REPROVACAO), 2) AS media_taxa_reprovacao
FROM
tbl_enem
WHERE
INSE IS NOT NULL
AND NU_TAXA_APROVACAO IS NOT NULL
AND NU_TAXA_REPROVACAO IS NOT NULL

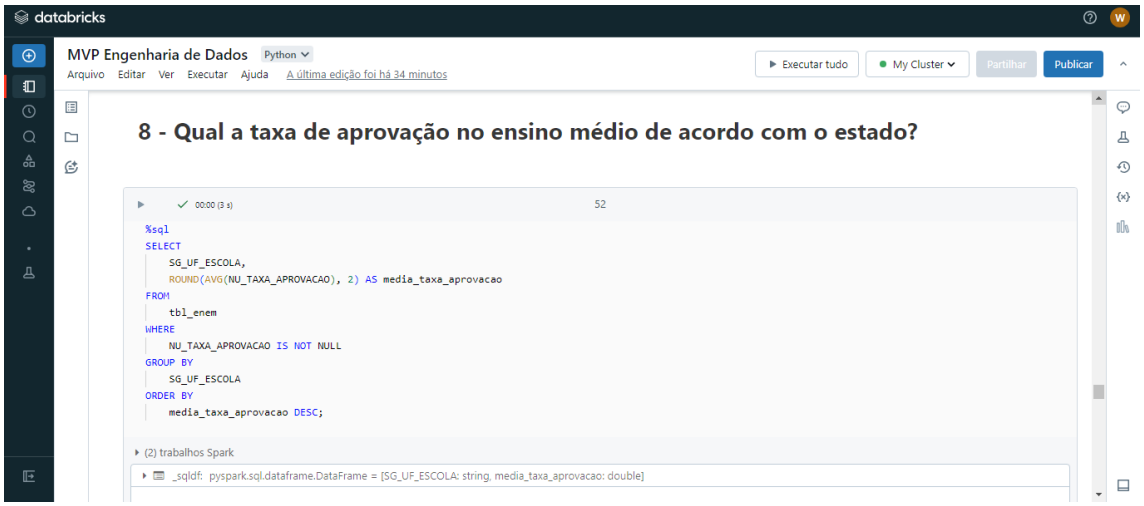

databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 33 minutos
Executar tudo
My Cluster
Partilhar
Publicar
00:00 (3 s)
44
SQL
INSE IS NOT NULL
AND NU_TAXA_APROVACAO IS NOT NULL
AND NU_TAXA_REPROVACAO IS NOT NULL
GROUP BY
INSE
ORDER BY
INSE;
(2) trabalhos Spark
_sqldf: pyspark.sql.dataframe.DataFrame = [INSE: string, media_taxa_aprovacao: double ... mais 1 campo]
Tabela
INSE
1.2 media_taxa_aprovacao
1.2 media_taxa_reprovacao
1 Grupo 1 83.87 8.1
2 Grupo 2 82.08 10.18
3 Grupo 3 83.35 11.2
4 Grupo 4 87.21 9.59
5 Grupo 5 93.68 5.72
6 Grupo 6 95.08 4.77
6 linhas | 3,19 segundos de tempo de execução
Atualizada há 20 horas

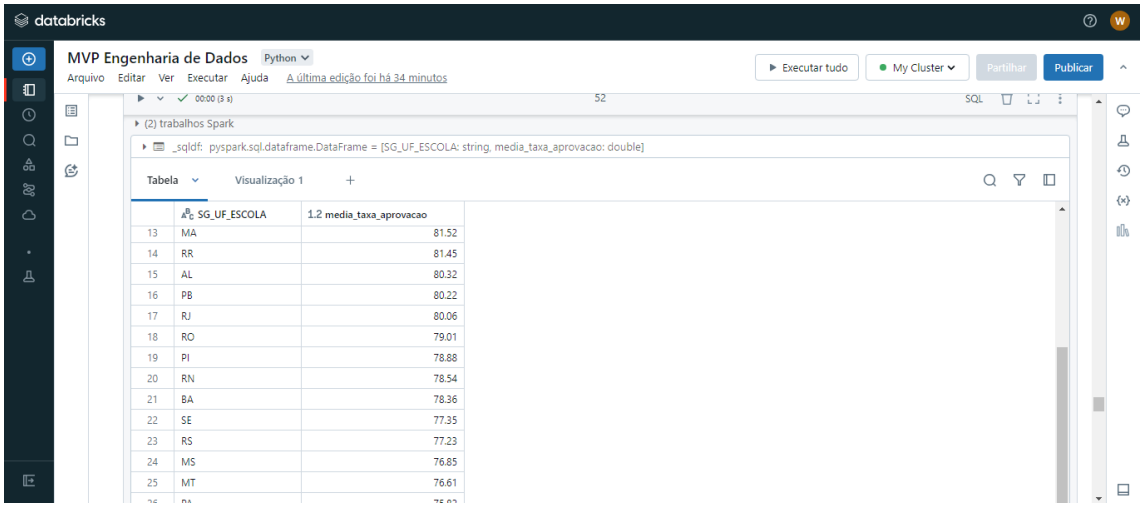

databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 33 minutos
Executar tudo
My Cluster
Partilhar
Publicar
1 Grupo 1 83.87 8.1
2 Grupo 2 82.08 10.18
3 Grupo 3 83.35 11.2
4 Grupo 4 87.21 9.59
5 Grupo 5 93.68 5.72
6 Grupo 6 95.08 4.77
6 linhas | 3,19 segundos de tempo de execução
Atualizada há 20 horas
Podemos analisar de acordo com o nível socioeconômico maior o resultados são na maior parte das vezes mais favoráveis, isso pode indicar um acesso mais amplo desses grupos sociais a um ensino com melhor qualidade.



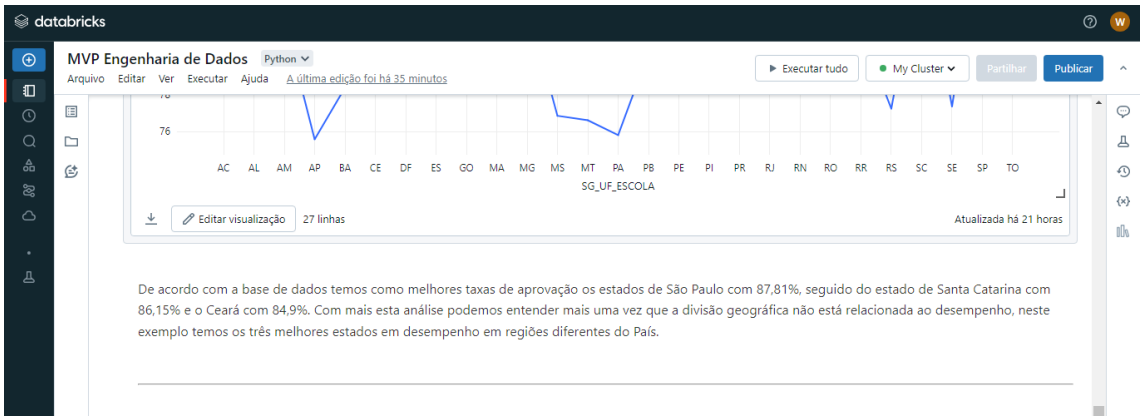


Podemos compreender de acordo com os anos históricos de aplicação do ENEM que a taxa de aprovação no ensino médio subiu considerávelmente.

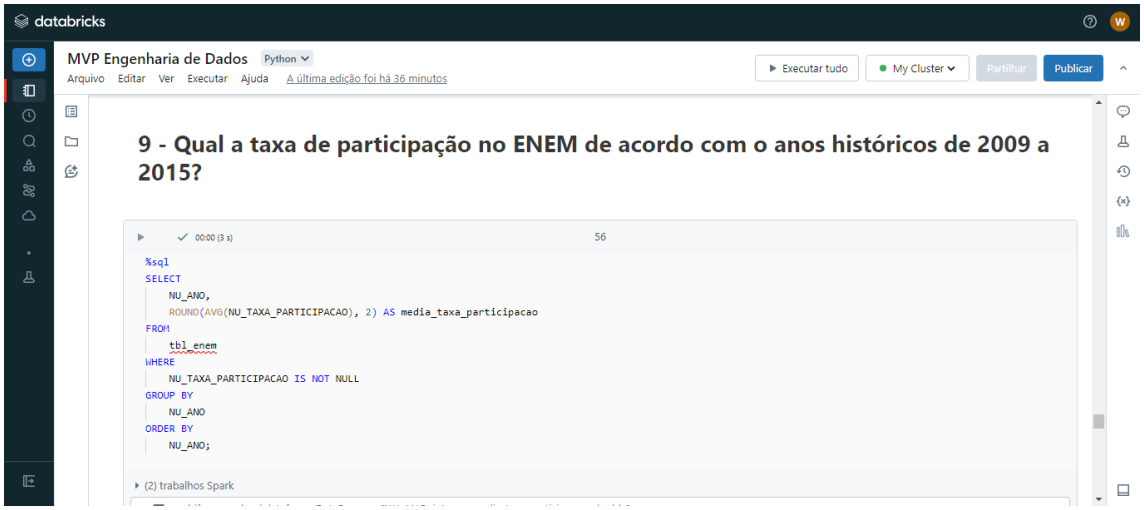

databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 34 minutos
Executar tudo
My Cluster
Partilhar
Publicar
8 - Qual a taxa de aprovação no ensino médio de acordo com o estado?
00:00 (3 s)
52
%sql
SELECT
SG_UF_ESCOLA,
ROUND(AVG(NU_TAXA_APROVACAO), 2) AS media_taxa_aprovacao
FROM
tbl_enem
WHERE
NU_TAXA_APROVACAO IS NOT NULL
GROUP BY
SG_UF_ESCOLA
ORDER BY
media_taxa_aprovacao DESC;
(2) trabalhos Spark
_sqldf: pyspark.sql.dataframe.DataFrame = [SG_UF_ESCOLA: string, media_taxa_aprovacao: double]


databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 34 minutos
Executar tudo
My Cluster
Partilhar
Publicar
00:00 (3 s)
52
SQL
(2) trabalhos Spark
_sqldf: pyspark.sql.dataframe.DataFrame = [SG_UF_ESCOLA: string, media_taxa_aprovacao: double]
Tabela
Visualização 1
SG_UF_ESCOLA
1.2 media_taxa_aprovacao
13 MA 81.52
14 RR 81.45
15 AL 80.32
16 PB 80.22
17 RJ 80.06
18 RO 79.01
19 PI 78.88
20 RN 78.54
21 BA 78.36
22 SE 77.35
23 RS 77.23
24 MS 76.85
25 MT 76.61


databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 35 minutos
Executar tudo
My Cluster
Partilhar
Publicar
Tabela
Visualização 1
88
86
84
82
80
78
76
media_taxa_aprovacao
AC AL AM AP BA CE DF ES GO MA MG MS MT PA PB PE PI PR RJ RN RO RR RS SC SE SP TO
SG_UF_ESCOLA
Editar visualização
27 linhas
Atualizada há 21 horas



De acordo com a base de dados temos como melhores taxas de aprovação os estados de São Paulo com 87,81%, seguido do estado de Santa Catarina com 86,15% e o Ceará com 84,9%. Com mais esta análise podemos entender mais uma vez que a divisão geográfica não está relacionada ao desempenho, neste exemplo temos os três melhores estados em desempenho em regiões diferentes do País.

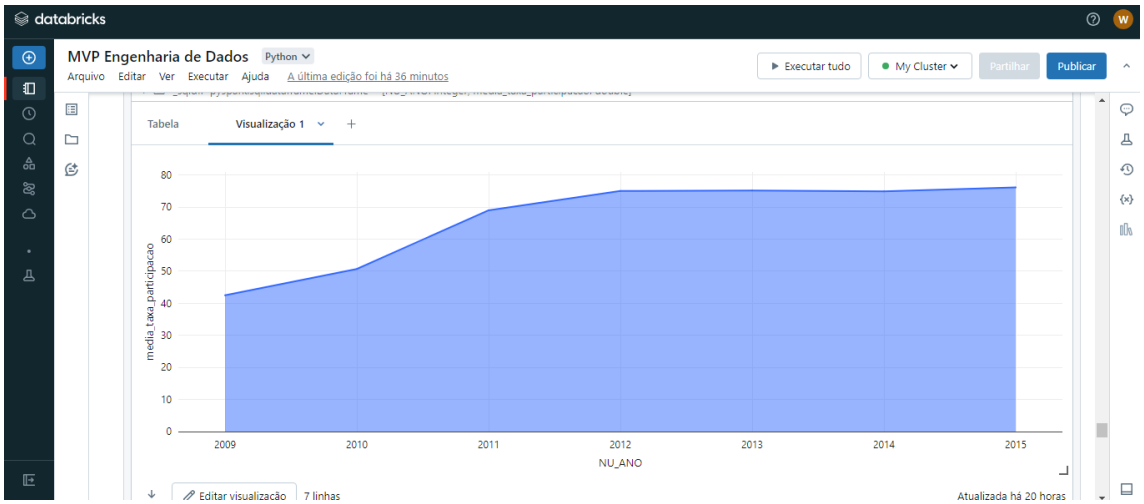


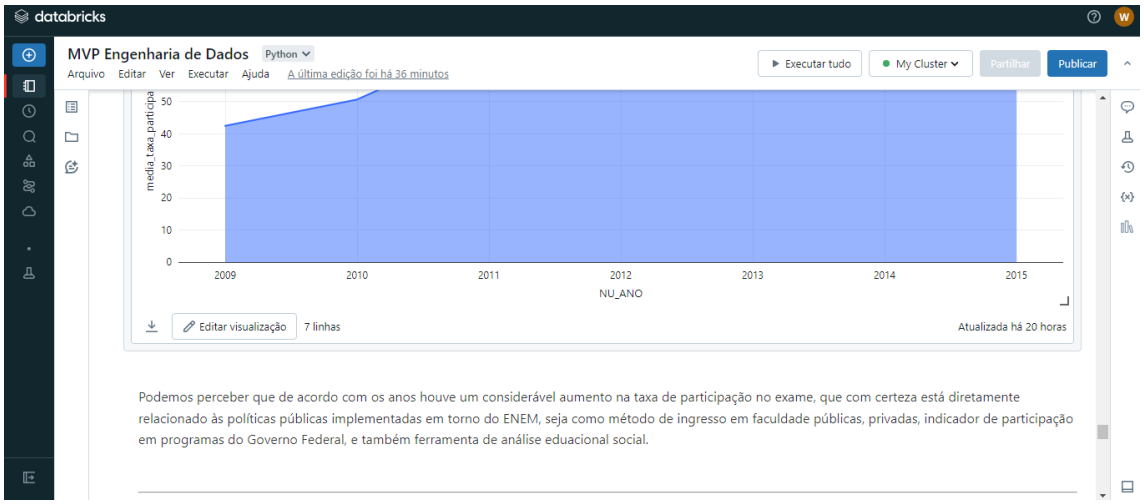
9 - Qual a taxa de participação no ENEM de acordo com o anos históricos de 2009 a 2015?

00:00 (3 s) 56

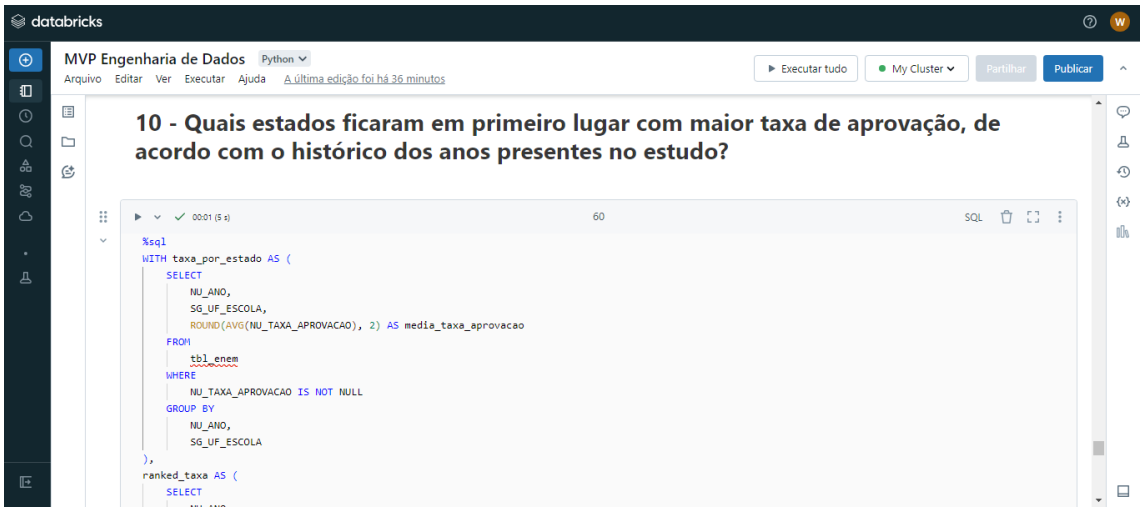
```
%sql
SELECT
    NU_ANO,
    ROUND(AVG(NU_TAXA_PARTICIPACAO), 2) AS media_taxa_participacao
FROM
    tbl_enem
WHERE
    NU_TAXA_PARTICIPACAO IS NOT NULL
GROUP BY
    NU_ANO
ORDER BY
    NU_ANO;
```

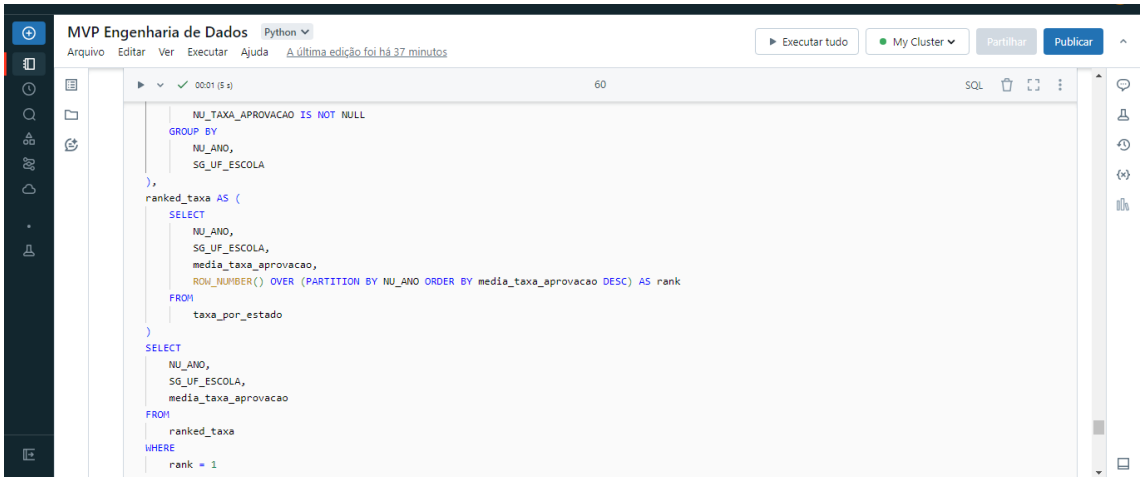
(2) trabalhos Spark

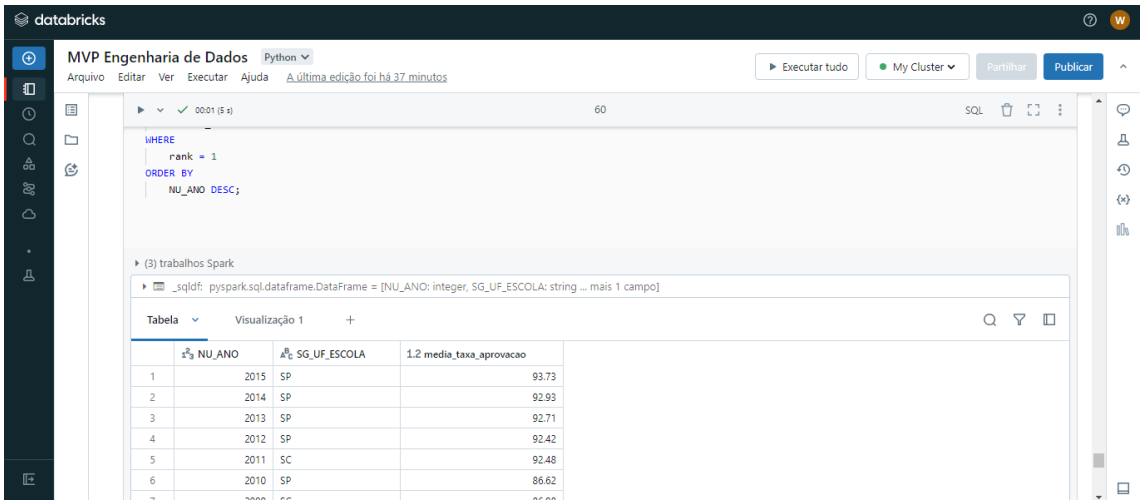




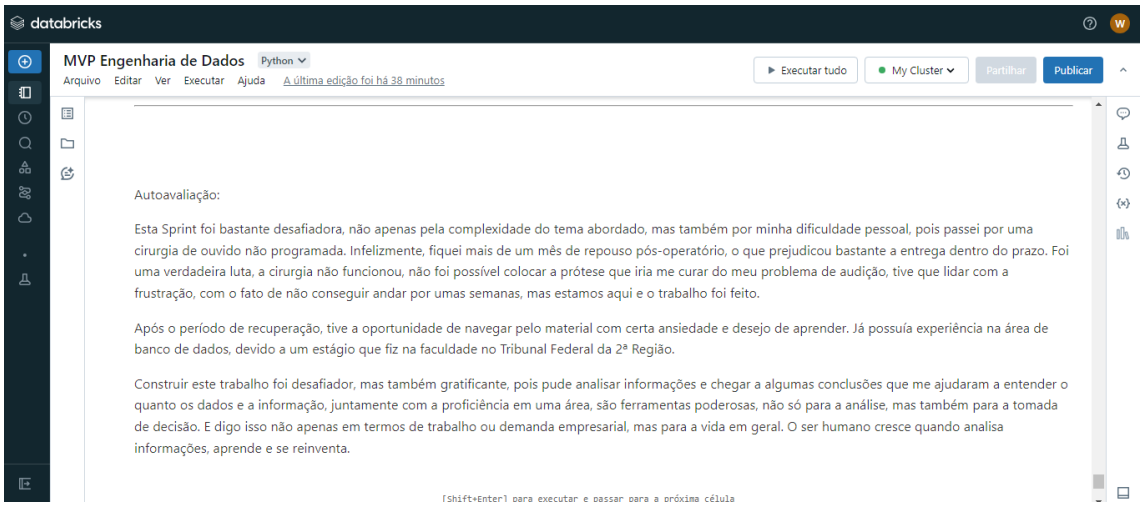
Podemos perceber que de acordo com os anos houve um considerável aumento na taxa de participação no exame, que com certeza está diretamente relacionado às políticas públicas implementadas em torno do ENEM, seja como método de ingresso em faculdade públicas, privadas, indicador de participação em programas do Governo Federal, e também ferramenta de análise eduacional social.







Os estados de São Paulo e Santa Catarina sempre estiveram em primeiro lugar em relação a taxa e aprovação de acordo com os anos e todos os dados históricos.


databricks
MVP Engenharia de Dados
Python
Arquivo Editar Ver Executar Ajuda A última edição foi há 38 minutos
Executar tudo
My Cluster
Partilhar
Publicar
Autoavaliação:
Esta Sprint foi bastante desafiadora, não apenas pela complexidade do tema abordado, mas também por minha dificuldade pessoal, pois passei por uma cirurgia de ouvido não programada. Infelizmente, fiquei mais de um mês de repouso pós-operatório, o que prejudicou bastante a entrega dentro do prazo. Foi uma verdadeira luta, a cirurgia não funcionou, não foi possível colocar a prótese que iria me curar do meu problema de audição, tive que lidar com a frustração, com o fato de não conseguir andar por umas semanas, mas estamos aqui e o trabalho foi feito.
Após o período de recuperação, tive a oportunidade de navegar pelo material com certa ansiedade e desejo de aprender. Já possuía experiência na área de banco de dados, devido a um estágio que fiz na faculdade no Tribunal Federal da 2ª Região.
Construir este trabalho foi desafiador, mas também gratificante, pois pude analisar informações e chegar a algumas conclusões que me ajudaram a entender o quanto os dados e a informação, juntamente com a proficiência em uma área, são ferramentas poderosas, não só para a análise, mas também para a tomada de decisão. E digo isso não apenas em termos de trabalho ou demanda empresarial, mas para a vida em geral. O ser humano cresce quando analisa informações, aprende e se reinventa.
[Shift+Enter] para executar e passar para a próxima célula