

Documentação Projeto Final – Equipe 04

Tema: Meio Ambiente

Escola: Soulcode Academy

Curso: Bootcamp Analista de Dados – Martech – AD2

Professores: Franciane Rodrigues, Douglas Ribeiro e Jonathas Carneiro

Integrantes: Camila Barcellos, Camylla Oliveira, Ester Beatriz, Heloisa Gasques, Maria Eduarda Klug e Vanessa Monteiro.

**Sumário:**

Problema...................................................................................................................02

Metodologia..............................................................................................................02

Fluxo de Trabalho.....................................................................................................02

Estrutura do código...................................................................................................04

Código ETL..............................................................................................................15

MongoDB.................................................................................................................17

Análises em Pyspark.................................................................................................18

Consultas na Big Query............................................................................................20

Dashboards...............................................................................................................22

Conclusões e Sugestões............................................................................................23

**Problema:**

Este projeto final do curso de Análise de Dados da SoulCode Academy utiliza o processo de ETL (Extract, Transform and Load) em dois Datasets diferentes, especificamente focados em Efeito estufa e Desmatamento no Brasil. Problema este que vem gerando grande impacto ao meio ambiente, deixando rastros e afetando os biomas e a comunidade à sua volta. Os arquivos foram tratados, analisados e visualizados com ajuda das ferramentas: Google Cloud Platform, Python, Pandas, PySpark, Cloud Storage Big Query e MongoDB.

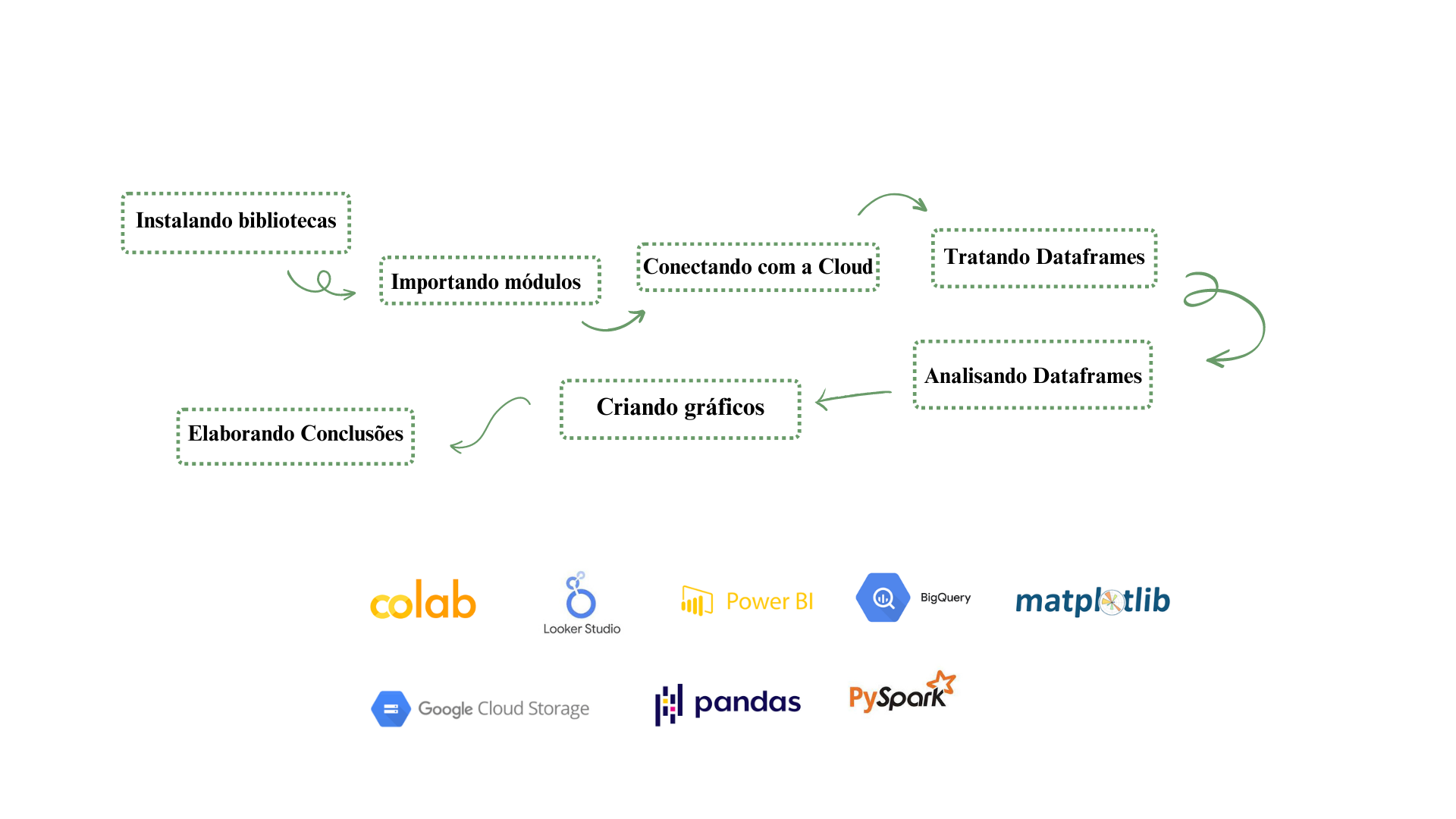
Foram escolhidas estas duas bases de dados a fim de responder as seguintes perguntas:

“Quais são os principais fatores determinantes de desmatamento no Brasil, ou em regiões do Brasil?” e “Quais são os Estados ou regiões que mais são prejudicados pelo desmatamento?”

**Metodologia:**

Nosso projeto foi baseado na metodologia KDD **- (**Knowledge Discovery in Databases) seguindo as seguintes etapas: seleção dos dados; pré-processamento dos dados; transformação dos dados; mineração de dados e interpretação e avaliação dos resultados. Já em questão a organização, optamos pela metodologia Kanban, separando os tópicos a fazer, fazendo e feito.

**Fluxo de Trabalho:**



**Estrutura do código:**

Bases coletadas para a análise:

* Df\_Desmatamento

Link de origem: <https://plataforma.alerta.mapbiomas.org/mapa?monthRange[0]=2019-01&monthRange[1]=2023-06&sources[0]=All&territoryType=all&authorization=all&embargoed=all&locationType=alert_code&activeBaseMap=7>

Dicionário da base:

* Ano: O ano em que os dados do desmatamento foram registrados ou coletados.
* Amazonia\_legal\_area: A área total da região da Amazônia Legal em estudo, que abrange nove estados do Brasil.
* Areaha: A quantidade de área desmatada em hectares (ha) no ano especificado.
* Causador: A causa ou origem do desmatamento.
* Nome\_bioma: O nome do bioma em que ocorreu o desmatamento
* Area\_bioma: A área total do bioma em estudo.
* Territorios\_indigenas: A presença ou impacto do desmatamento em territórios indígenas, que são áreas habitadas por comunidades indígenas.
* Amazonia\_legal: Uma indicação se a área do desmatamento está dentro da região da Amazônia Legal.
* Amazonia\_legal\_area: A área total da Amazônia Legal.
* Bacia\_hidrografica\_nome: O nome da bacia hidrográfica na qual ocorreu o desmatamento. As bacias hidrográficas são áreas de drenagem onde a água flui para um rio principal ou corpo de água.
* Bacia\_hidrografica\_area: A área total da bacia hidrográfica em estudo.
* Municipio: O município onde ocorreu o desmatamento.
* Municipio\_area: A área total do município em estudo.
* Contagem\_prop\_rurais: O número de propriedades rurais envolvidas no desmatamento.
* Estados: Os estados do Brasil nos quais ocorreu o desmatamento.
* Estados\_area: A área total dos estados em estudo.
* Df\_Estufa:

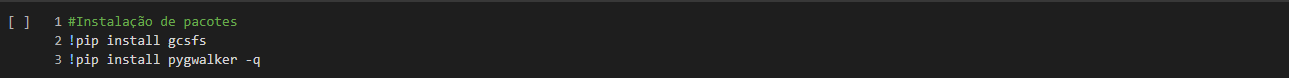
Link de origem: <https://dados.gov.br/dados/conjuntos-dados/inpe-em>;

Dicionário da base:

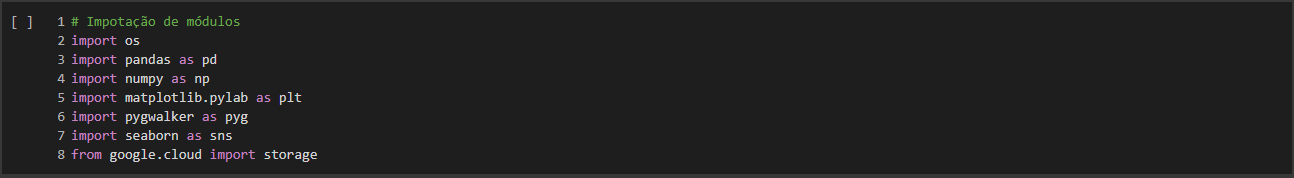
* Ano: Anos referentes às estimativas de emissão de gases de efeito estufa (1960-2020).
* Area\_desmat\_acum: Área Desmatada Acumulada.
* Area\_desmat\_ano:Área desmatada no ano (no desmatamento, a vegetação suprimida).
* Area\_degrad\_ano: Área degradada no ano (na área degradada, a vegetação permanece, em diferentes estados de degeneração).
* Emissao\_CO2\_1: Estimativas de 1ª Ordem (que supõe de modo simplificado que 100% das emissões ocorram no momento da mudança de uso/cobertura).
* Emissao CO2\_2: Estimativas de 2ª Ordem (que buscam representar o processo gradativo de liberação e absorção do carbono como ocorre de fato).
* Emissao\_corte\_veg\_sec: Emissão de CO2 por corte de vegetação secundária (veg. secundária é a resultante de um processo natural de regeneração da vegetação).
* Absorcao\_recresc\_sec: Absorção de CO2 por recrescimento de vegetação secundária (resultante de um processo natural de regeneração da vegetação).
* Emissao\_degrad\_floresta: Emissão de CO2 por degradação da floresta (na área degradada, a vegetação permanece, em diferentes estados de degeneração).
* Absorcao\_recup\_area\_degrad: Absorção de CO2 por recuperação da área degradada.
* Balanco\_emissoes\_1\_ordem: Balanço considerando emissões comprometidas (balanço de primeira ordem).
* Balanco\_segunda\_ordem: Balanço considerando o processo (balanço de segunda ordem).

**Código ETL:**

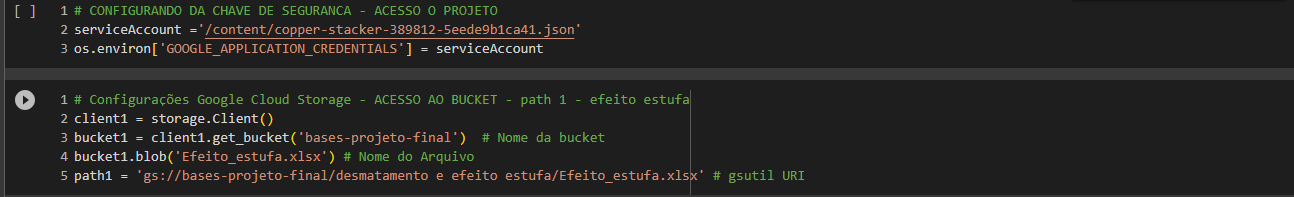
Instalação de Bibliotecas

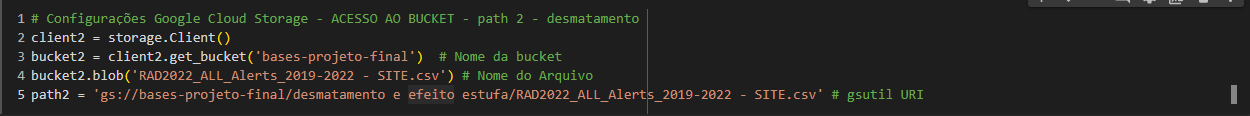


Importação de Módulos



Conector para Cloud Storage (Bucket)

Usando a chave de serviço e acessando a Bucket na GCP

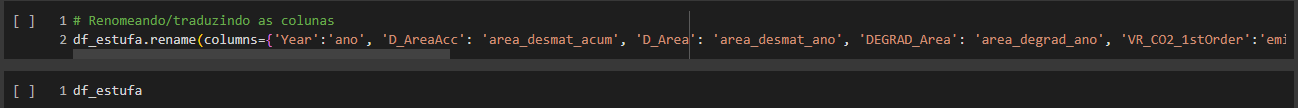


Tratamento Dataset Estufa

Analisando o Dataframe de Estufa



Renomeando/traduzindo colunas e verificando as mudanças feitas



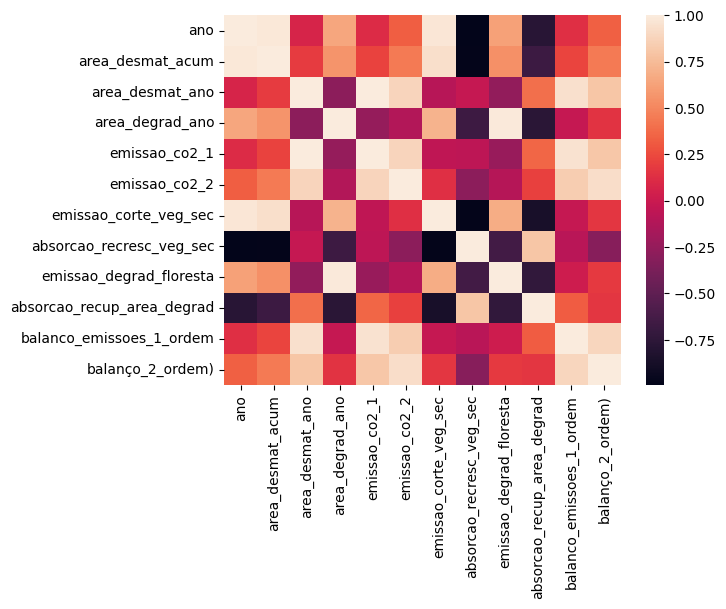
Retirada de colunas que não serão utilizadas na análise

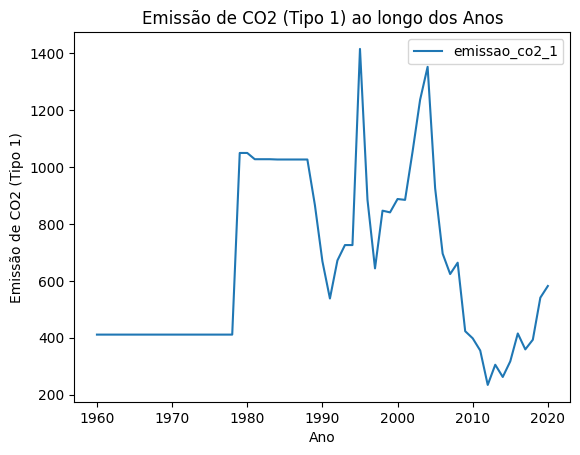


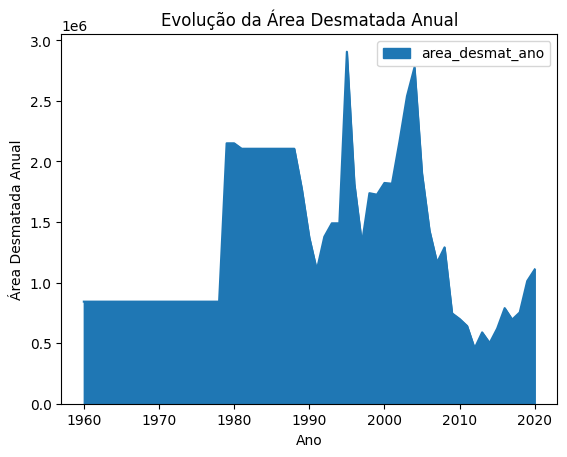
Verificando o Dataframe

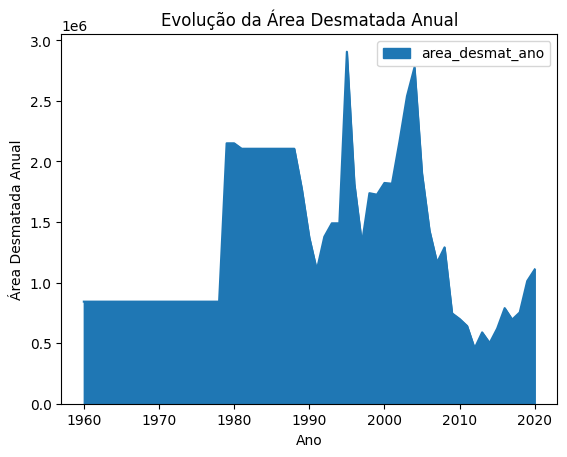


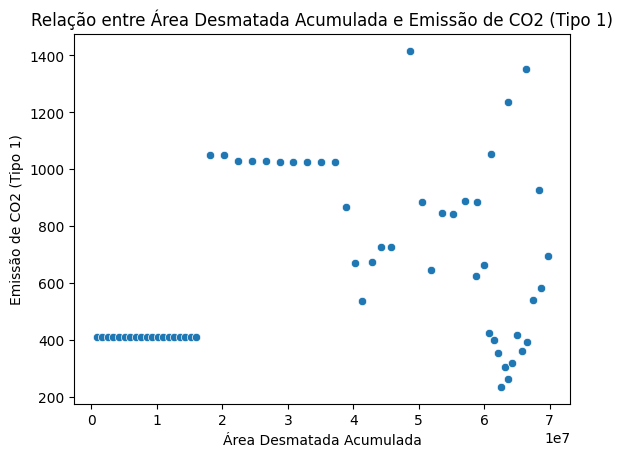
Análise Dataset Estufa (Gráficos)











Tratamento Dataset Desmatamento

Analisando a base de dados



Retirada de colunas não úteis



Verificando valores únicos 

Analisando o Dataframe de Desmatamento

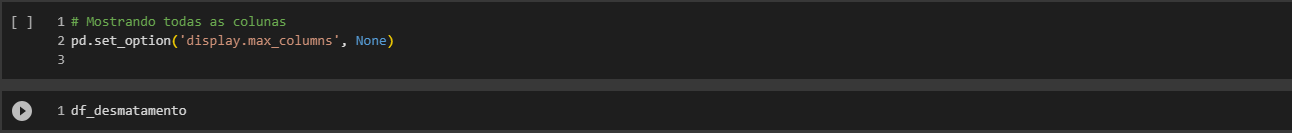


Verificando valores distintos

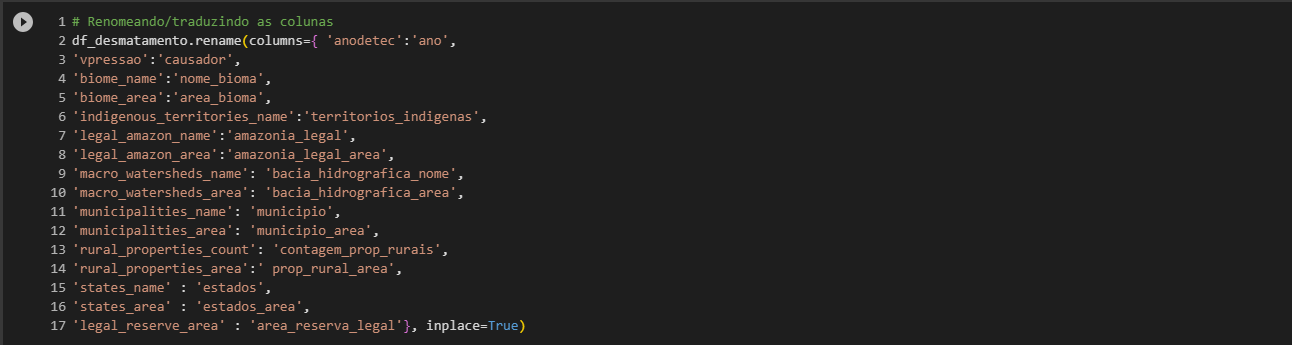


Analisando colunas do Dataframe

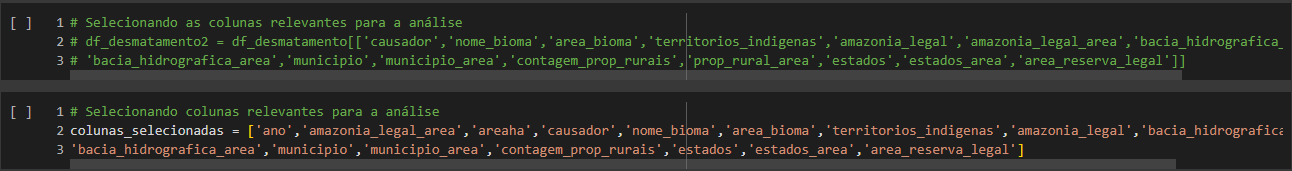
Observando todas as colunas e vendo o Dataframe



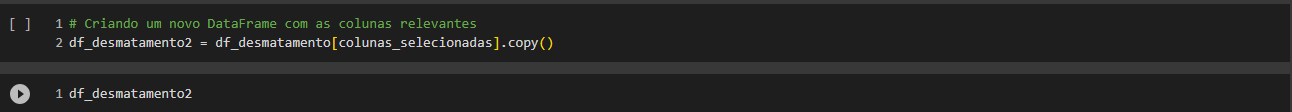
Renomeando/Traduzindo colunas



Selecionando colunas relevantes para a análise



Criando um novo Dataframe com as colunas relevantes e verificando o mesmo



Visualizando valores únicos da coluna “Causador”



Substituindo os valores da coluna 'causador' utilizando o método replace

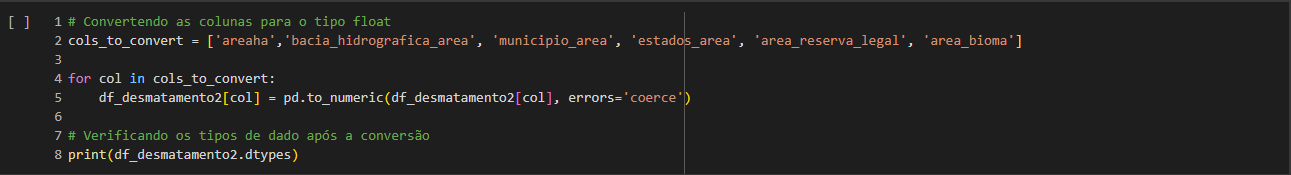
Substituindo valores nulos por 0



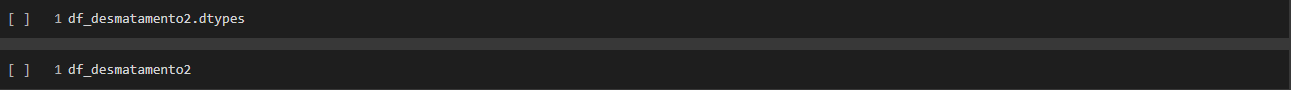
Substituindo o tipo de coluna (de float para int)



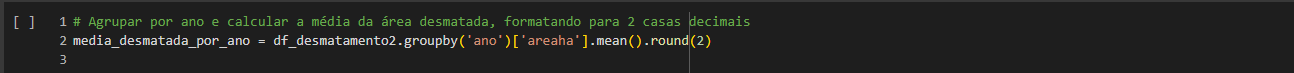
Convertendo colunas para o tipo float



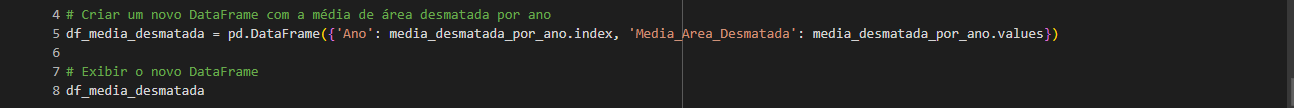
Verificando mudanças feitas



Agrupando por ano, calculando a média da área desmatada com duas casas decimais



Criando um novo dataframe com a média da área desmatada e verificando-o



Análise Dataset Desmatamento (Gráficos)

Visualizações em Gráficos de barras

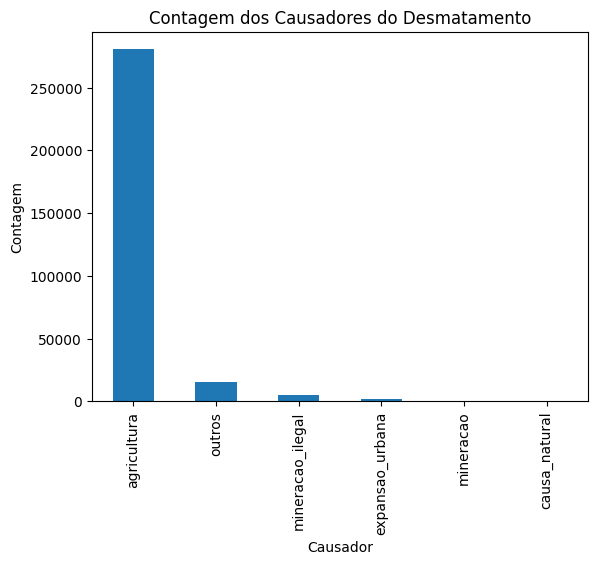
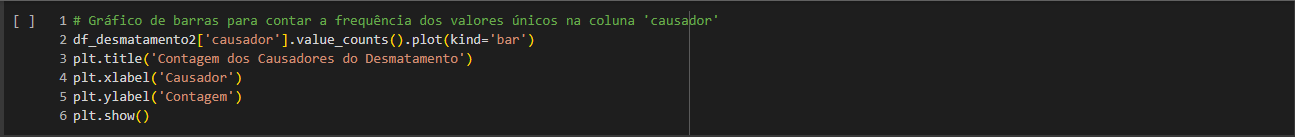
Gráfico de barras mostrando a frequência dos valores únicos na coluna 'causador'

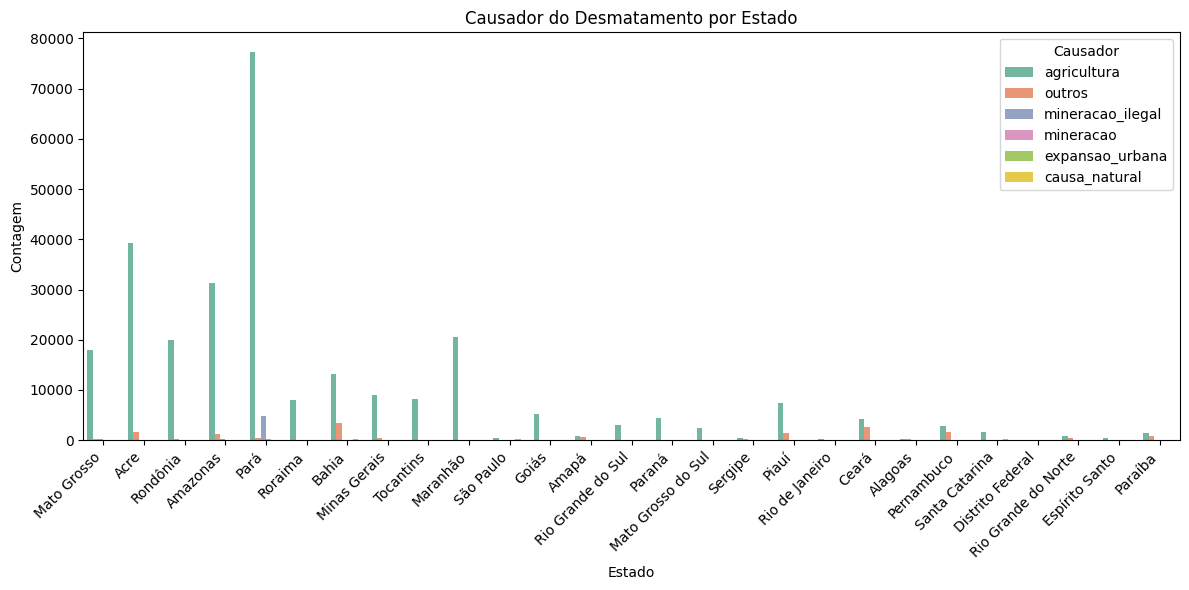
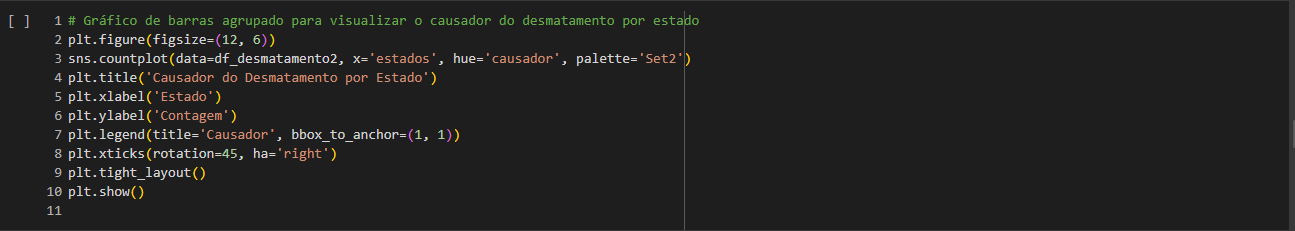
Gráfico de barras agrupado para visualizar o causador do desmatamento por estado

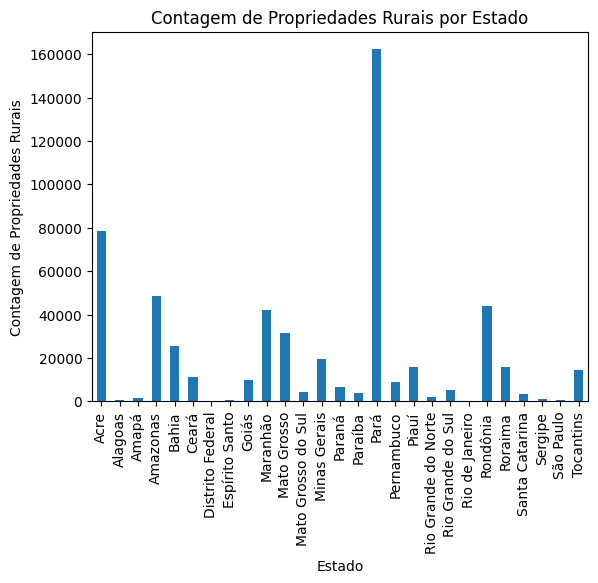
Gráfico de barras para ’estados’ e ’contagem\_prop\_rurais’

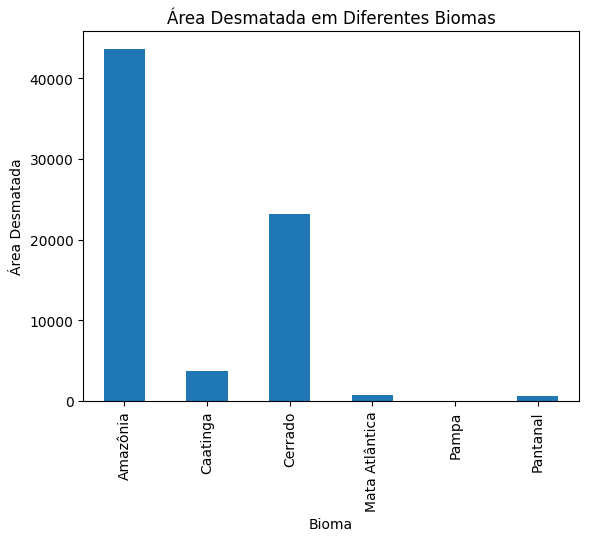
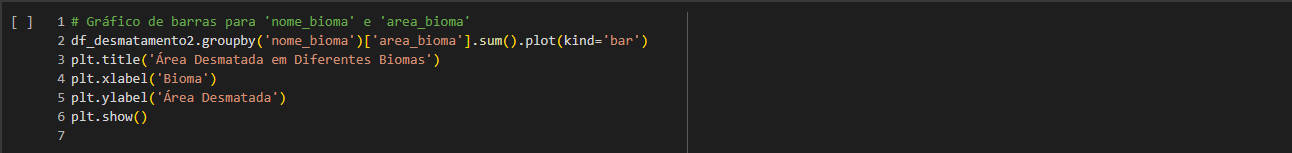
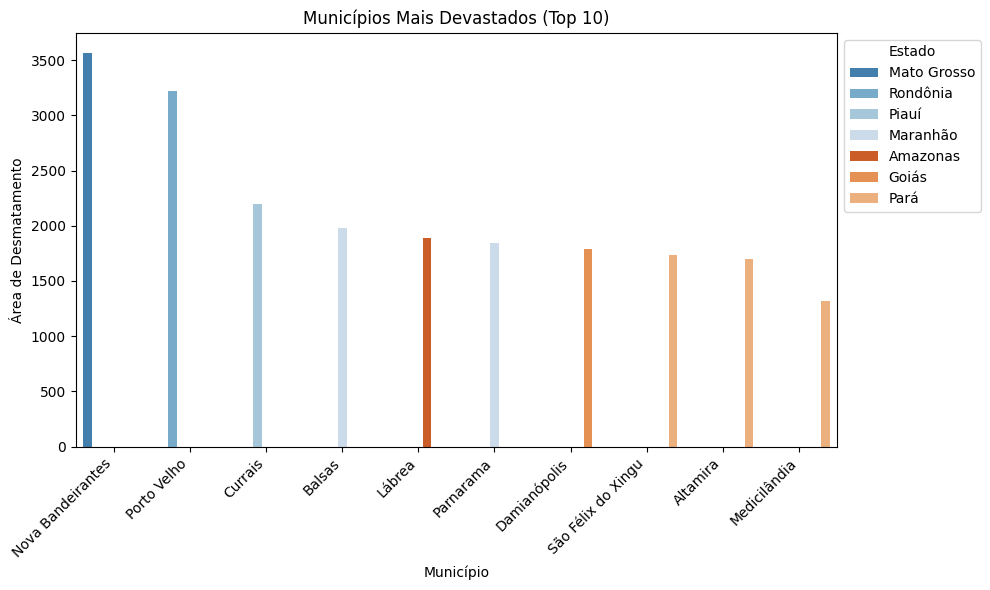
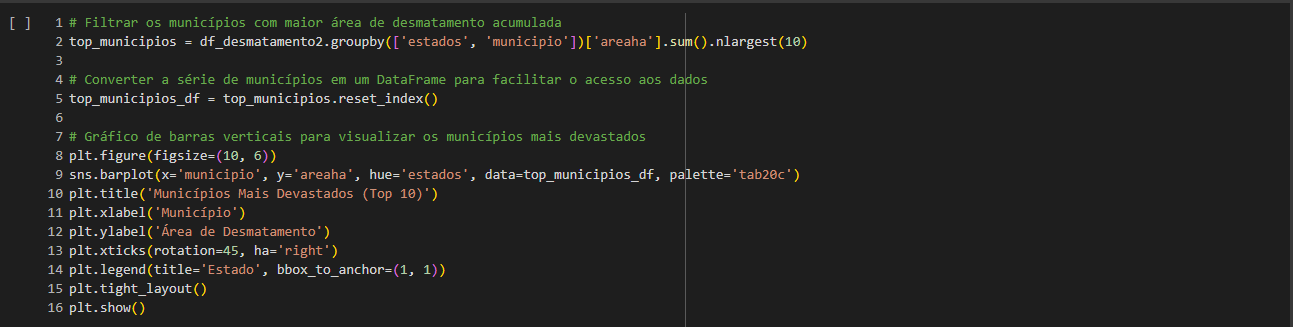
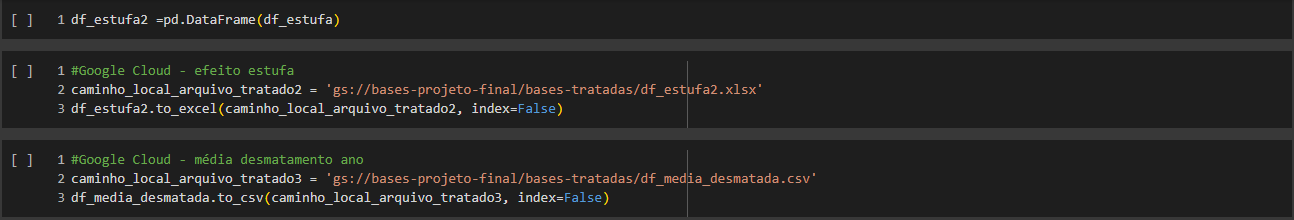
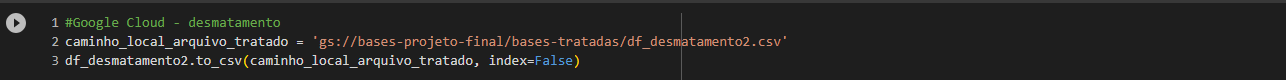
Gráfico de barras para ‘nome\_bioma’ e ‘area\_bioma’

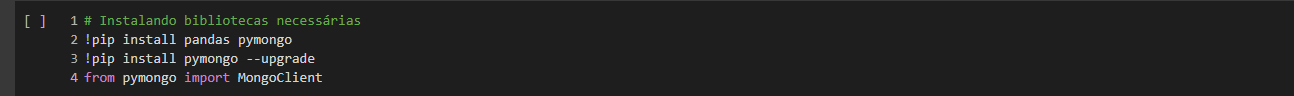
Gráfico de barras verticais para visualizar os municípios mais devastados

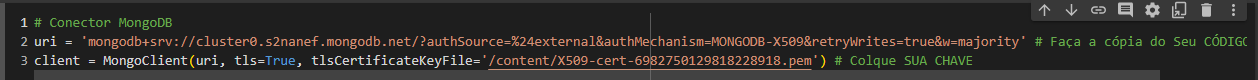


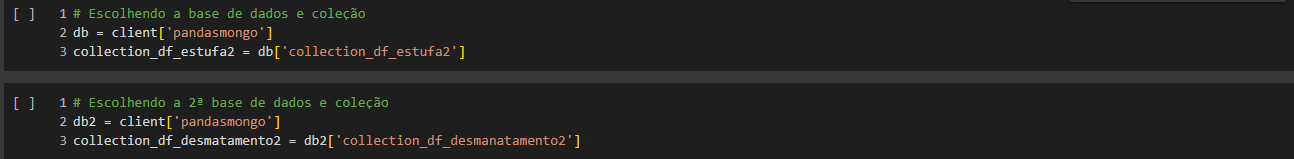
Carregamento das bases tratadas para uma nova pasta na bucket

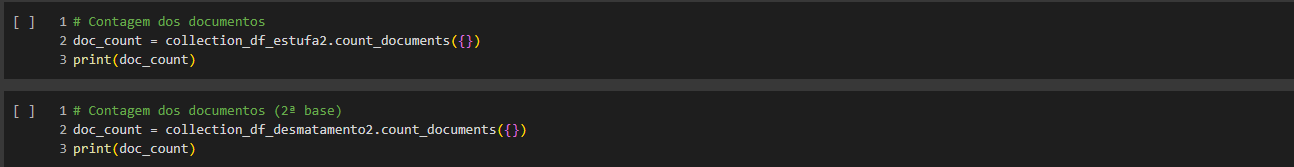
Mongo DB

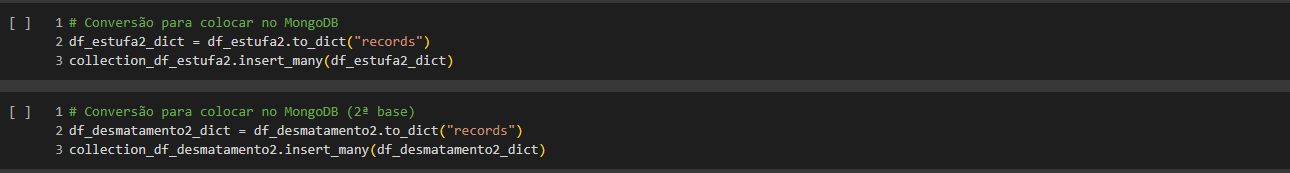
Key de acesso ao cluster: uri="mongodb+srv://cluster0.s2nanef.mongodb.net/?authSource=%24external&authMechanism=MONGODB-X509&retryWrites=true&w=majority"

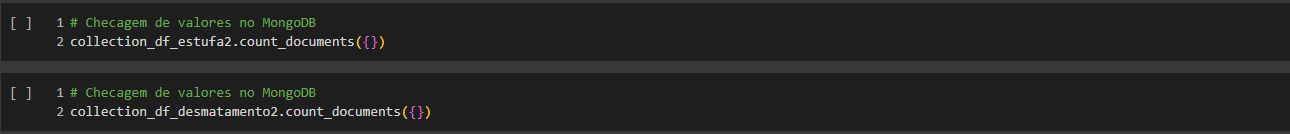
Instalando bibliotecas necessárias

Conector 

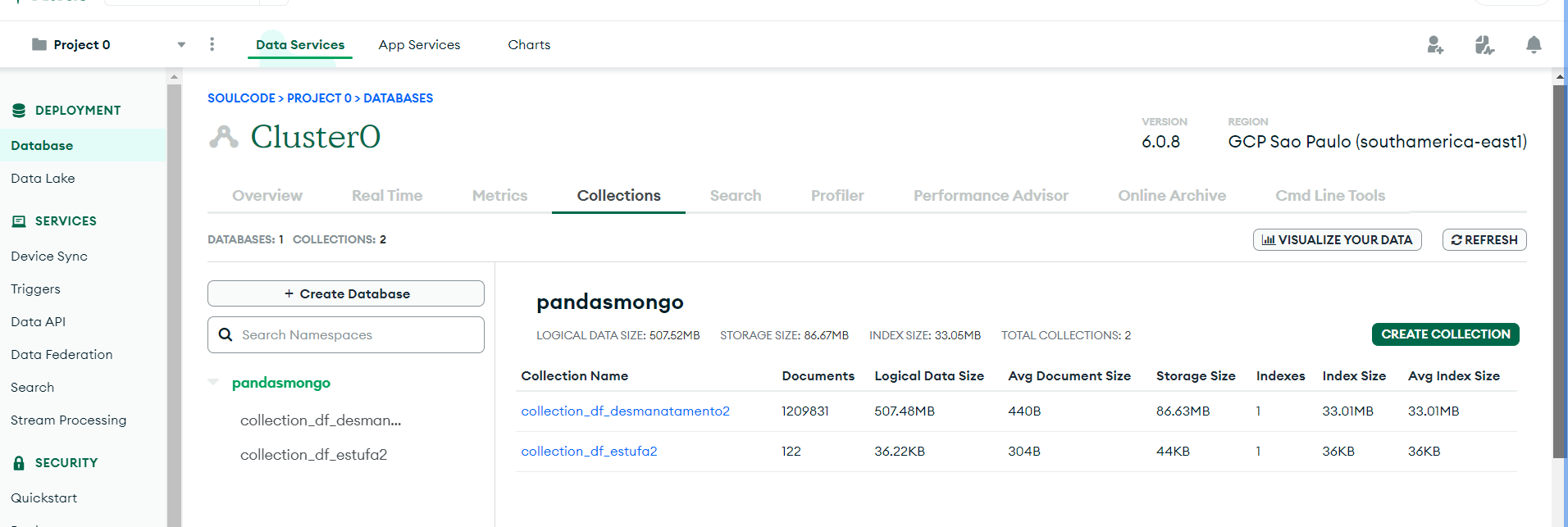
Escolhendo as bases de dados e as coleções

Contagem dos documentos

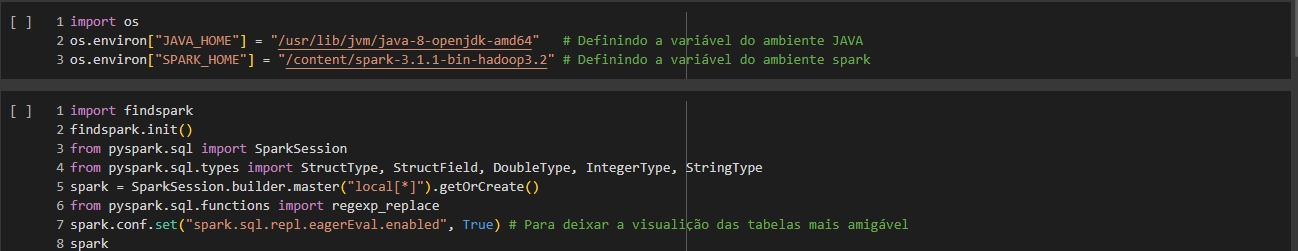
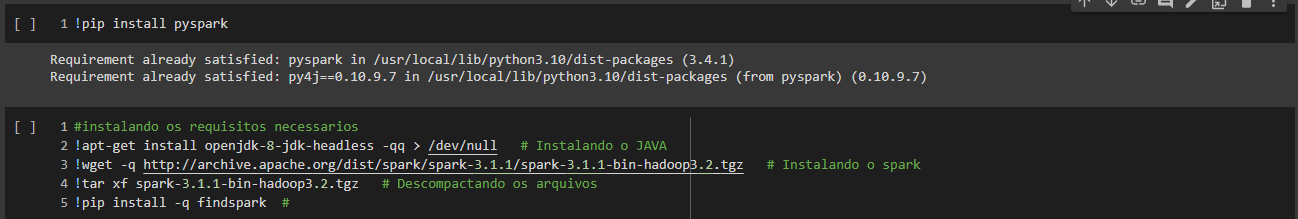
Conversão para MongoDB

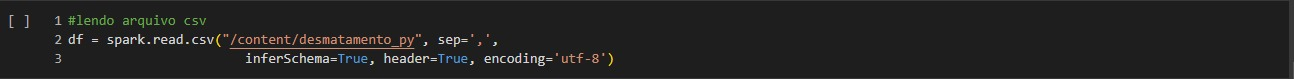
Checagem de valores

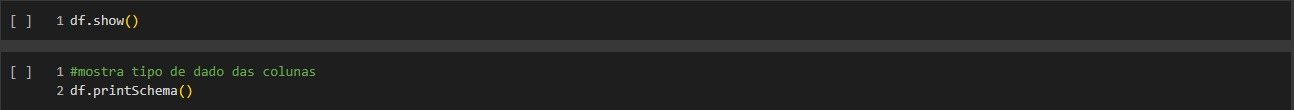
Visualização das coleções no MongoDB



Análises em Pyspark

Instalando a biblioteca e os requisitos necessários

Importando e lendo um arquivo csv

Visualizando o Dataframe e mostrando os tipos de dados das colunas

Medidas descritivas

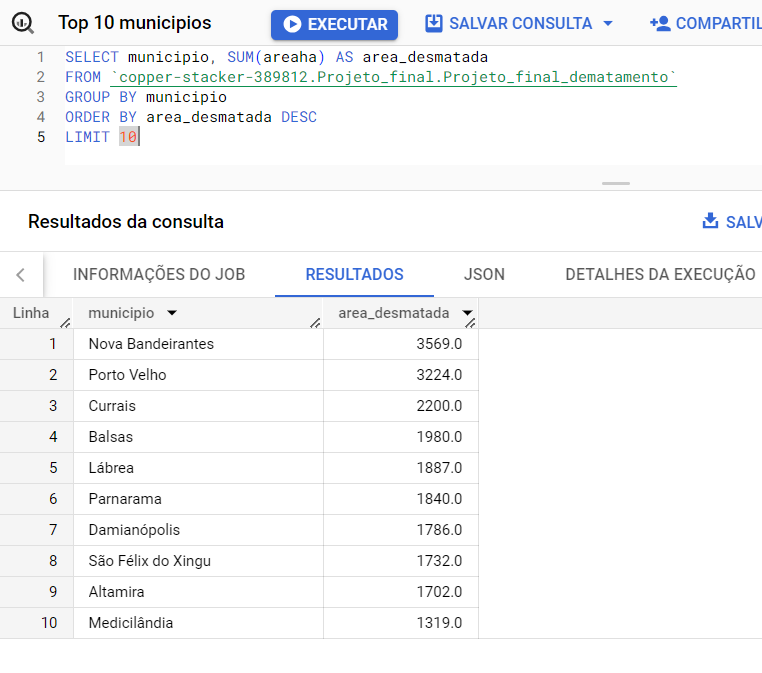
Ordenando Dataframe por coluna específica (causador)

Visualizando o tamanho do Dataframe

Lendo a primeira linha do Dataframe

**Consultas na Big Query:**

Top 10 munícipios mais devastados



Contagem de causadores



Total de área desmatada:



Área desmatada acumulada ao longo dos anos

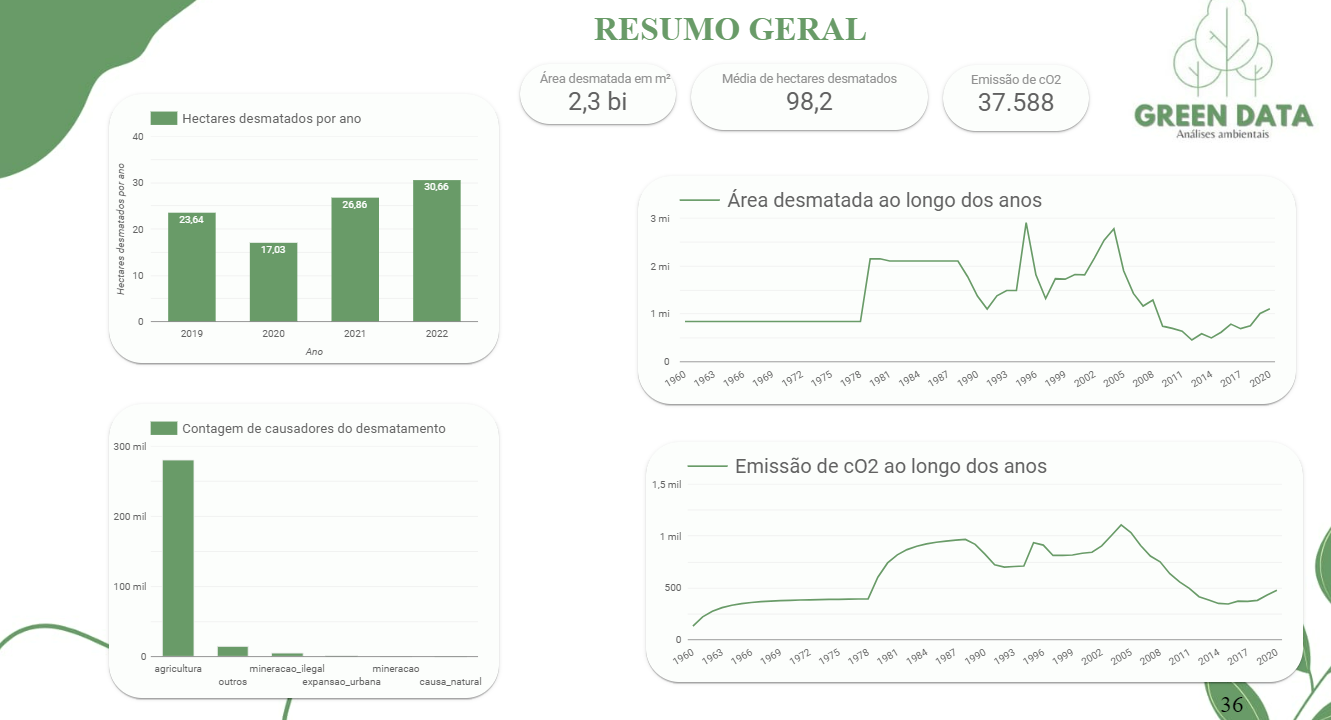


**Dashboards:**

**Looker Studio:**

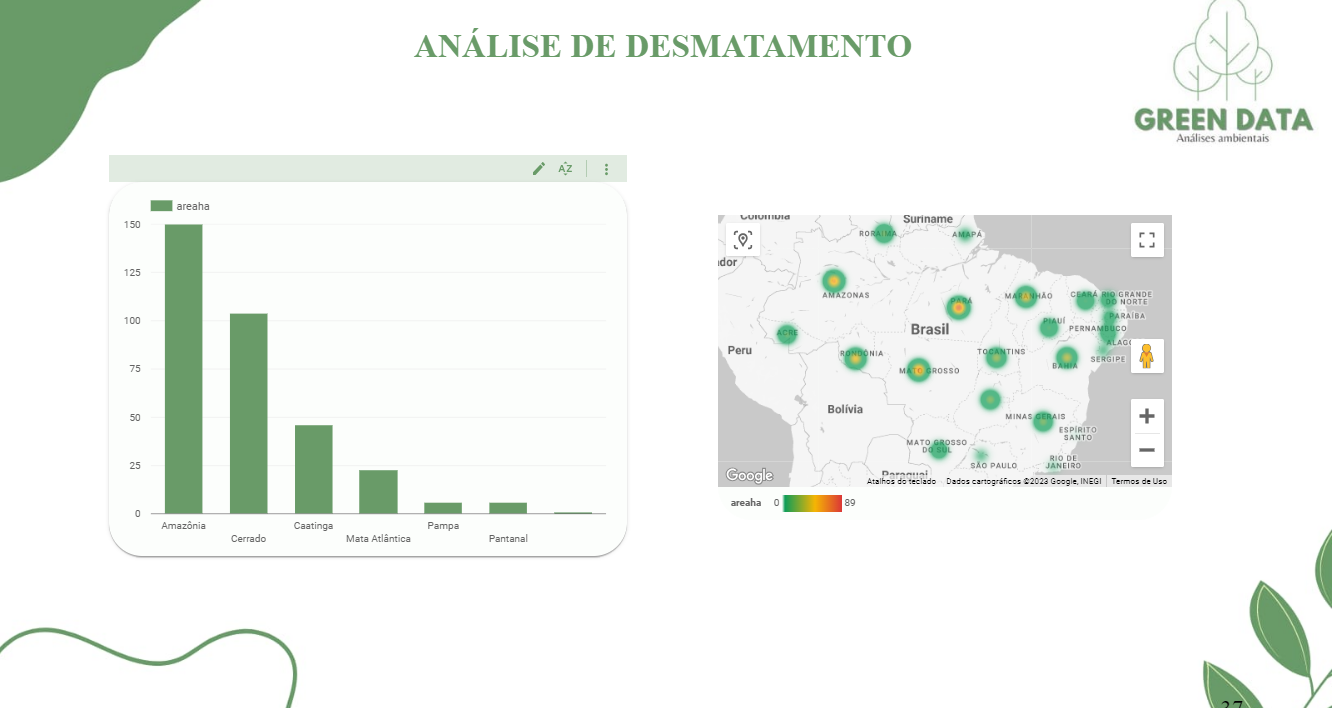
**Dashboard 1:**

No dashboard abaixo apresentamos inicialmente três cards, o primeiro indicando o total de metros² desmatados em 4 anos, o segundo informando a média de hectares desmatados em 60 anos e o terceiro a quantidade de quilos de Gás carbônico emitidos no Brasil desde 1960. Ao lado esquerdo apresentamos dois gráficos de colunas, sendo eles em relação a quantidade de hectares desmatados por ano e abaixo a comparação da quantidade de causadores do desmatamento. Ao lado direito temos então dois gráficos de linhas representando o crescimento da área desmatada ao longo dos anos e o crescimento da emissão de cO2 no Brasil.



**Dashboard 2:**

Neste dashboard apresentamos um gráfico de colunas indicando os biomas mais devastados, sendo eles a Amazônia e o Cerrado, e ao lado, um mapa de calor indicando os estados com maior área desmatada, podendo interpretá-los através do tamanho do centro amarelo, indicando então o estado do Pará e do Mato Grosso como os mais preocupantes.



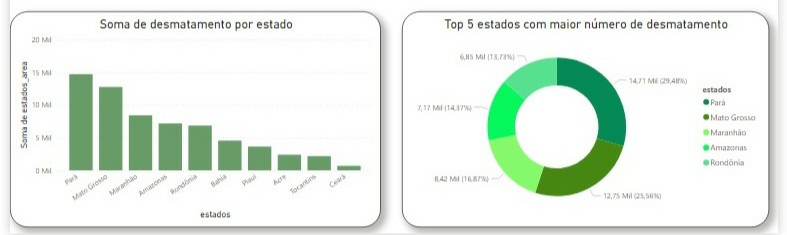
**Power BI**

**Dashboard 1:**

O primeiro gráfico nos mostra o quanto cada causador desmatou por ano, e embaixo, novamente os biomas mais desmatados.  


**Dashboard 2:**

Aqui então apresentamos um gráfico de colunas apontando os estados com maior índice de desmatamento e ao lado, em um gráfico de rosca, a área desmatada em cada um deles.



**Conclusões:**

O Pará é o estado que mais desmata do país, somando quase 15 mil metros quadrados, o que equivale a quase 30% da área do país todo, sendo que o maior causador de desmatamento pode-se dizer que é a agricultura, (que ultrapassa os 90% dos níveis de desmatamento). A quantidade de hectares que vem sendo desmatados vem aumentando nos últimos anos, tendo tido uma breve queda em 2020 e voltando a subir no ano seguinte.

A emissão de CO2 é algo de extrema preocupação e importância, já que tem uma tendência a aumentar cada vez mais, porém, de forma surpreendente, as taxas vêm diminuindo desde 2004. Há 19 anos, os números passavam de 1000 e hoje os níveis de emissão se encontram abaixo de 500

**Sugestões de melhoria:**

Como medidas de soluções, pensamos em a SoulCode se afiliar a uma ong próxima, apoiando e participando de projetos que ela promove e também passar a fazer reciclagem e a devida separação do lixo que seus funcionários produzem na empresa. Além disso, o atual governo já deixou públicas suas metas e é nosso dever como cidadãos apoiar tais medidas, nos disponibilizar a conhecer as metas propostas e fazer a devida cobrança ao governo no tempo proposto das mesmas. Quanto à produção de CO2, pode-se armazená-lo em salmouras subterrâneas, o que pode levar a gerar energia geotérmica e também metano como combustíveis, caso se interesse. A empresa também pode se tornar carbon free, o que se traduz como carbono zero e significa calcular o total das suas emissões, reduzi-las conforme suas possibilidades e, ainda, balancear o restante das emissões por meio da compensação. Em nossas pesquisas, encontramos algumas ongs que podem servir de interesse da SoulCode, sendo elas a SOS Amazônia (criada em uma universidade), Amazônia Legal (criada pelo governo), Amazônia Protege (projeto do Ministério Público Federal).