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.

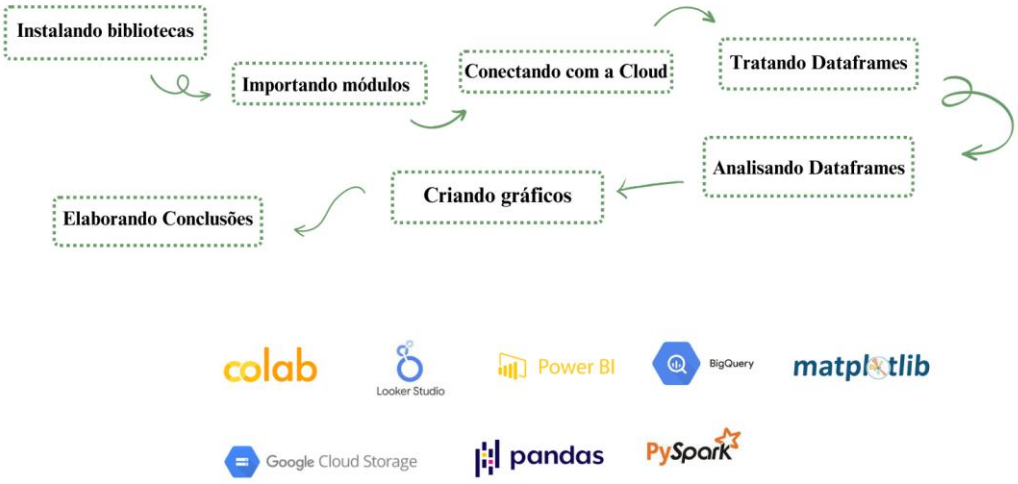
Sumário:

Problema:

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

```
#Instalação de pacotes
!pip install gcsfs
!pip install pygwalker -q
```

Importação de Módulos

```
# Impotação de módulos
import os
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
import pygwalker as pyg
import seaborn as sns
from google.cloud import storage
```

Conector para Cloud Storage (Bucket)

Usando a chave de serviço e acessando a Bucket na GCP

```
# CONFIGURANDO DA CHAVE DE SEGURANCA - ACESSO O PROJETO
serviceAccount ='/content/copper-stacker-389812-5eede9b1ca41.json'
os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = serviceAccount
```

```
# Configurações Google Cloud Storage - ACESSO AO BUCKET - path 1 - efeito estufa
client1 = storage.Client()
bucket1 = client1.get_bucket('bases-projeto-final')  # Nome da bucket
bucket1.blob('Efeito_estufa.xlsx') # Nome do Arquivo
path1 = 'gs://bases-projeto-final/desmatamento e efeito estufa/Efeito_estufa.xlsx' # gsutil URI
```

```
# Configurações Google Cloud Storage - ACESSO AO BUCKET - path 2 - desmatamento
client2 = storage.Client()
bucket2 = client2.get_bucket('bases-projeto-final')  # Nome da bucket
bucket2.blob('RAD2022_ALL_Alerts_2019-2022 - SITE.csv') # Nome do Arquivo
path2 = 'gs://bases-projeto-final/desmatamento e efeito estufa/RAD2022_ALL_Alerts_2019-2022 - SITE.csv' # gsutil URI
```

Tratamento Dataset Estufa

Analisando o Dataframe de Estufa

```
df_estufa
```

```
df_estufa.head()
```

```
df_estufa.dtypes
```

```
df_estufa.describe()
```

Renomeando/traduzindo colunas e verificando as mudanças feitas

```
# Renomeando/traduzindo as colunas
df_estufa.rename(columns={'Year':'ano', 'D_AreaAcc': 'area_desmat_acum', 'D_Area': 'area_desmat_ano', 'DEGRAD_Area': 'area_degrad_ano', 'VR_CO2_1stOrder':'emi
```

```
df_estufa
```

Retirada de colunas que não serão utilizadas na análise

```
#Retirada de colunas não utilidades na análise
df_estufa.drop(['-','-.1','-.2'], axis=1, inplace=True)
```

Verificando o Dataframe

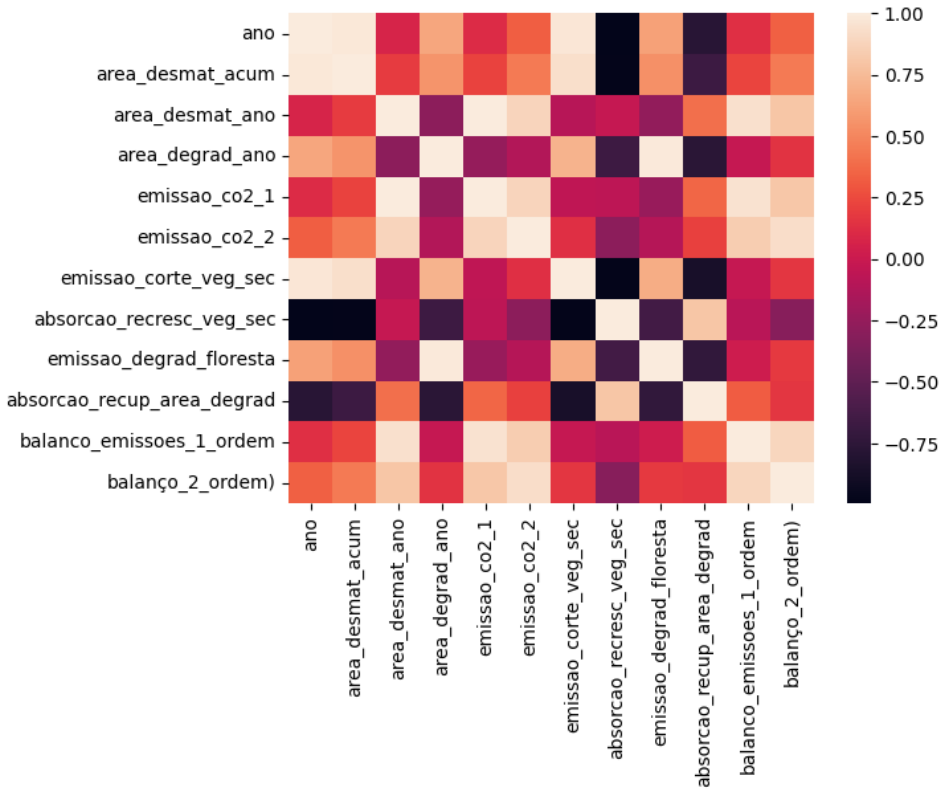
```
df_estufa
```

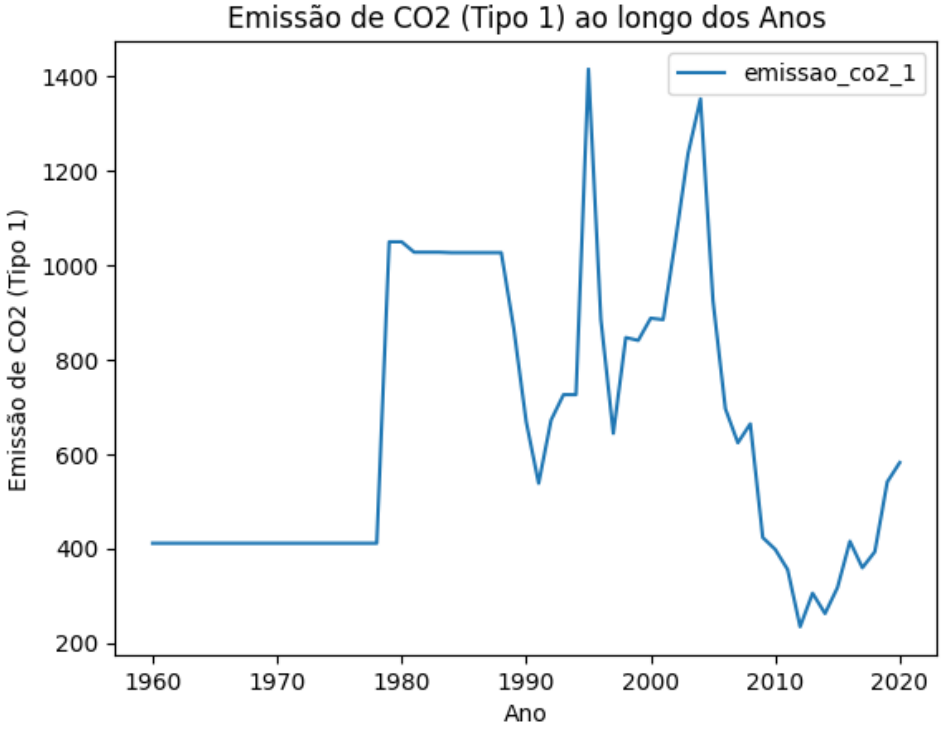
```
df_estufa.shape
```

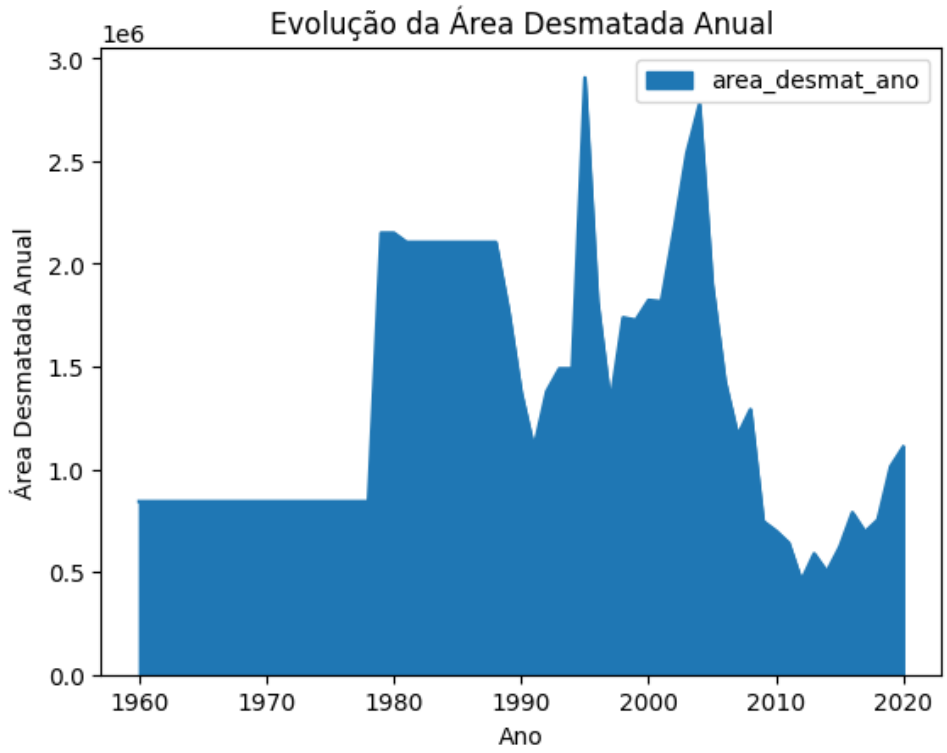
```
df_estufa.dtypes
```

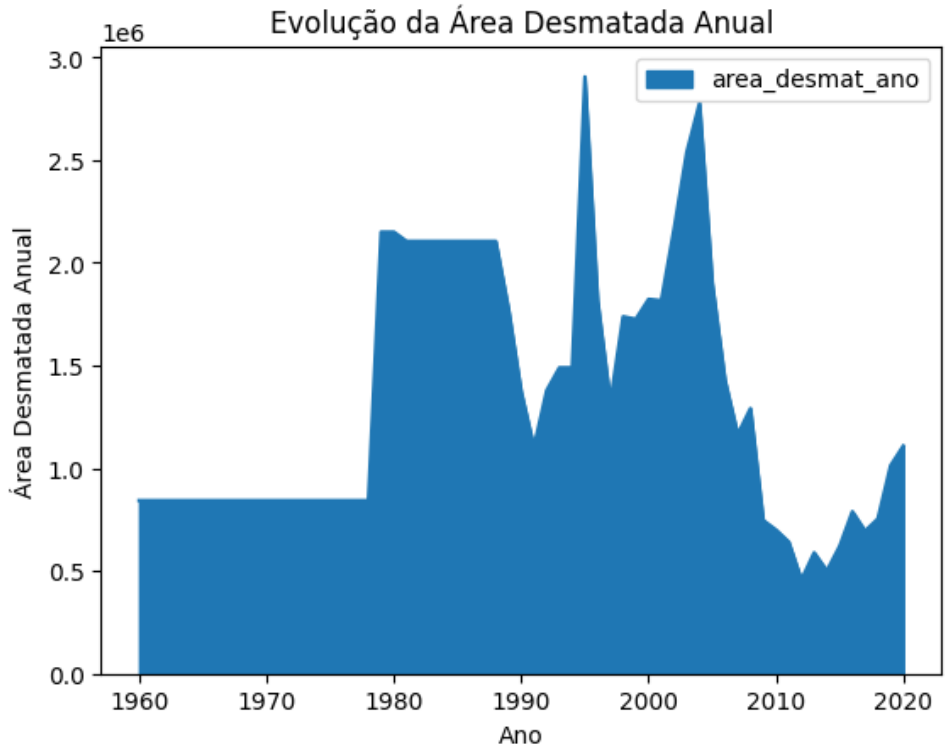
```
df_estufa.info()
```

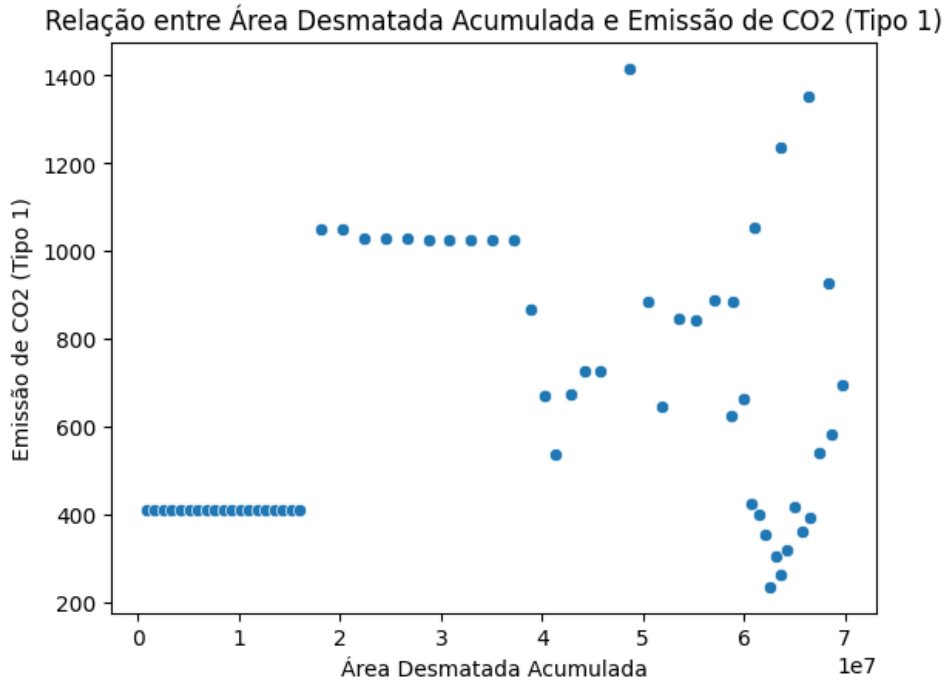
Análise Dataset Estufa (Gráficos)


ano
area_desmat_acum
area_desmat_ano
area_degrad_ano
emissao_co2_1
emissao_co2_2
emissao_corte_veg_sec
absorcao_recresc_veg_sec
emissao_degrad_floresta
absorcao_recup_area_degrad
balanco_emissoes_1_ordem
balanço_2_ordem)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Emissão de CO2 (Tipo 1) ao longo dos Anos
emissao_co2_1
Emissão de CO2 (Tipo 1)
Ano
200
400
600
800
1000
1200
1400
1960
1970
1980
1990
2000
2010
2020


Evolução da Área Desmatada Anual
1e6
area_desmat_ano
Área Desmatada Anual
Ano
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1960
1970
1980
1990
2000
2010
2020


Evolução da Área Desmatada Anual
1e6
area_desmat_ano
Área Desmatada Anual
Ano
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1960
1970
1980
1990
2000
2010
2020



Tratamento Dataset Desmatamento

Analisando a base de dados

```
[ ]  1 df_desmatamento.info()
```

Retirada de colunas não úteis

```
[ ]  1 #Retirada de colunas não utilizadas na análise
     2 df_desmatamento.drop(['Unnamed: 84','Unnamed: 85','Unnamed: 86','Unnamed: 87','Unnamed: 88'], axis=1, inplace=True)
```

Verificando valores únicos

```
[ ]  1 # Verificando se os valores são únicos na df_desmatamento
     2 df_desmatamento.fonte.is_unique
```

Analisando o Dataframe de Desmatamento

```
[ ]  1 df_desmatamento
```

Verificando valores distintos

```
[ ]  1 # Verificando valores distinct na df_desmatamento
     2 print(sorted(pd.unique(df_desmatamento['fonte'])))
```

Analisando colunas do Dataframe

```
[ ]  1 df_desmatamento.columns
```

Observando todas as colunas e vendo o Dataframe

```
[ ]  1 # Mostrando todas as colunas
     2 pd.set_option('display.max_columns', None)
     3

     1 df_desmatamento
```

Renomeando/Traduzindo colunas

```python
# Renomeando/traduzindo as colunas
df_desmatamento.rename(columns={ 'anodetec':'ano',
'vpressao':'causador',
'biome_name':'nome_bioma',
'biome_area':'area_bioma',
'indigenous_territories_name':'territorios_indigenas',
'legal_amazon_name':'amazonia_legal',
'legal_amazon_area':'amazonia_legal_area',
'macro_watersheds_name': 'bacia_hidrografica_nome',
'macro_watersheds_area': 'bacia_hidrografica_area',
'municipalities_name': 'municipio',
'municipalities_area': 'municipio_area',
'rural_properties_count': 'contagem_prop_rurais',
'rural_properties_area':' prop_rural_area',
'states_name' : 'estados',
'states_area' : 'estados_area',
'legal_reserve_area' : 'area_reserva_legal'}, inplace=True)
```

Selecionando colunas relevantes para a análise

```python
# Selecionando as colunas relevantes para a análise
# df_desmatamento2 = df_desmatamento[['causador','nome_bioma','area_bioma','territorios_indigenas','amazonia_legal','amazonia_legal_area','bacia_hidrografica_
# 'bacia_hidrografica_area','municipio','municipio_area','contagem_prop_rurais','prop_rural_area','estados','estados_area','area_reserva_legal']]
```

```python
# Selecionando colunas relevantes para a análise
colunas_selecionadas = ['ano','amazonia_legal_area','areaha','causador','nome_bioma','area_bioma','territorios_indigenas','amazonia_legal','bacia_hidrografica
'bacia_hidrografica_area','municipio','municipio_area','contagem_prop_rurais','estados','estados_area','area_reserva_legal']
```

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

```python
# Criando um novo DataFrame com as colunas relevantes
df_desmatamento2 = df_desmatamento[colunas_selecionadas].copy()
```

```python
df_desmatamento2
```

Visualizando valores únicos da coluna “Causador”

```python
# Visualizando valores únicos da coluna "causador"
print(sorted(pd.unique(df_desmatamento2['causador'])))
```

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

```python
# Substituir os valores da coluna 'causador' usando o método replace()
df_desmatamento2['causador'] = df_desmatamento2['causador'].replace(mapeamento_causador)
```

Substituindo valores nulos por 0

```python
# Substituindo os valores nulos da coluna por 0
df_desmatamento2['contagem_prop_rurais'].fillna(0, inplace=True)
```

Substituindo o tipo de coluna (de float para int)

```python
# Substituindo o tipo da coluna de float para int
df_desmatamento2['contagem_prop_rurais'] = df_desmatamento2['contagem_prop_rurais'].astype(int)
```

Convertendo colunas para o tipo float

```python
# Convertendo as colunas para o tipo float
cols_to_convert = ['areaha','bacia_hidrografica_area', 'municipio_area', 'estados_area', 'area_reserva_legal', 'area_bioma']

for col in cols_to_convert:
    df_desmatamento2[col] = pd.to_numeric(df_desmatamento2[col], errors='coerce')

# Verificando os tipos de dado após a conversão
print(df_desmatamento2.dtypes)
```

Verificando mudanças feitas

```python
df_desmatamento2.dtypes
```

```python
df_desmatamento2
```

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

```python
# Agrupar por ano e calcular a média da área desmatada, formatando para 2 casas decimais
media_desmatada_por_ano = df_desmatamento2.groupby('ano')['areaha'].mean().round(2)

```

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

```python
# Criar um novo DataFrame com a média de área desmatada por ano
df_media_desmatada = pd.DataFrame({'Ano': media_desmatada_por_ano.index, 'Media_Area_Desmatada': media_desmatada_por_ano.values})

# Exibir o novo DataFrame
df_media_desmatada
```

Análise Dataset Desmatamento (Gráficos)
Visualizações em Gráficos de barras

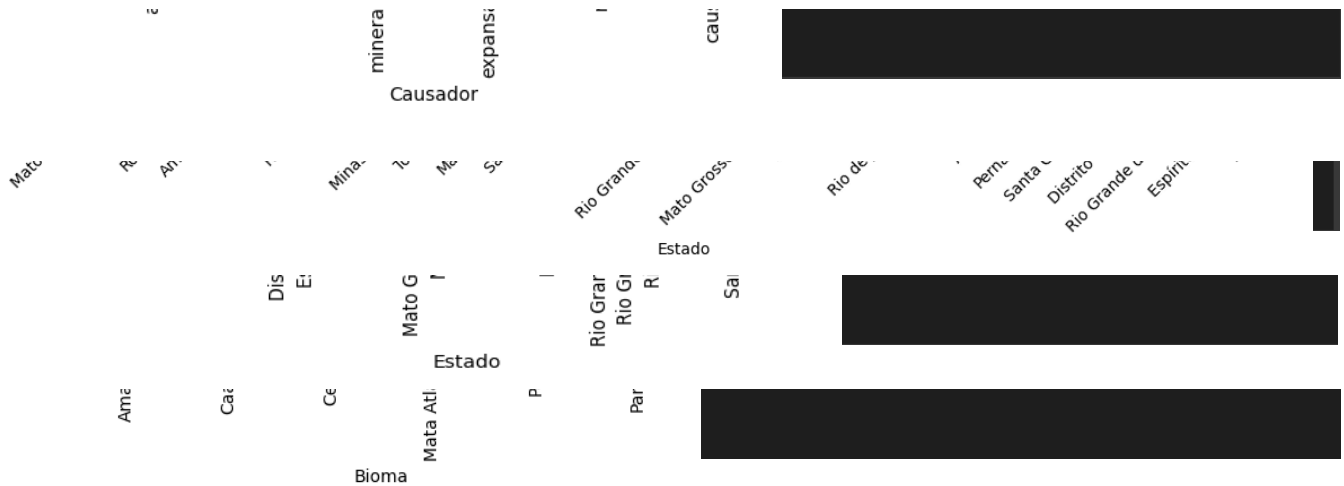
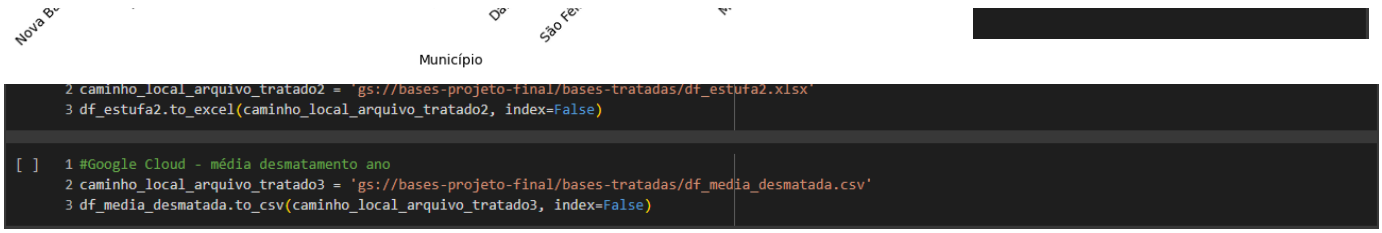


Gráfico de barras verticais para visualizar os municípios mais devastados



Mongo DB

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

```
!pip install pandas pymongo
!pip install pymongo --upgrade
from pymongo import MongoClient
```

```
# Conector MongoDB
uri = 'mongodb+srv://cluster0.s2nanef.mongodb.net/?authSource=%24external&authMechanism=MONGODB-X509&retryWrites=true&w=majority' # Faça a cópia do Seu CÓDIGO
client = MongoClient(uri, tls=True, tlsCertificateKeyFile='/content/X509-cert-6982750129818228918.pem') # Colque SUA CHAVE
```

```
# Escolhendo a 2ª base de dados e coleção
db2 = client['pandasmongo']
collection_df_desmatamento2 = db2['collection_df_desmanatamento2']
```

```
# Contagem dos documentos (2ª base)
doc_count = collection_df_desmatamento2.count_documents({})
print(doc_count)
```

```
# Conversão para colocar no MongoDB (2ª base)
df_desmatamento2_dict = df_desmatamento2.to_dict("records")
collection_df_desmatamento2.insert_many(df_desmatamento2_dict)
```

```
# Checagem de valores no MongoDB
collection_df_desmatamento2.count_documents({})
```

Visualização das coleções no MongoDB

Análises em Pyspark

```
from pyspark.sql.types import StructType, StructField, DoubleType, IntegerType, StringType
spark = SparkSession.builder.master("local[*]").getOrCreate()
from pyspark.sql.functions import regexp_replace
spark.conf.set("spark.sql.repl.eagerEval.enabled", True) # Para deixar a visualição das tabelas mais amigável
spark
```

```
#lendo arquivo csv
df = spark.read.csv("/content/desmatamento_py", sep=',',
                    inferSchema=True, header=True, encoding='utf-8')
```

```
#mostra tipo de dado das colunas
df.printSchema()
```

```
df.orderBy("causador",ascending=True) # ordena df por coluna específica
```


```
df.orderBy("causador",ascending=True) # ordena df por coluna específica
```


Visualizando o tamanho do DataFrame

```
print(f'({df.count()}, {len(df.columns)})') # Visualizando a dimensão do DataFrame
```

Lendo a primeira linha do DataFrame

```
df.collect()[0] # Lendo a 1ª linha do DataFrame
```

Consultas na Big Query:

Top 10 munícipios mais devastados

Top 10 municipios

EXECUTAR SALVAR CONSULTA COMPARTIL

```
1  SELECT municipio, SUM(areaha) AS area_desmatada
2  FROM `copper-stacker-389812.Projeto_final.Projeto_final_dematamento`
3  GROUP BY municipio
4  ORDER BY area_desmatada DESC
5  LIMIT 10
```

Resultados da consulta SALV

INFORMAÇÕES DO JOB | RESULTADOS | JSON | DETALHES DA EXECUÇÃO

Linha	municipio	area_desmatada
1	Nova Bandeirantes	3569.0
2	Porto Velho	3224.0
3	Currais	2200.0
4	Balsas	1980.0
5	Lábrea	1887.0
6	Parnarama	1840.0
7	Damianópolis	1786.0
8	São Félix do Xingu	1732.0
9	Altamira	1702.0
10	Medicilândia	1319.0

Contagem de causadores

Contagem de ca...

EXECUTAR SALVAR CONSULTA

```
1  SELECT causador, SUM(areaha) AS area_desmatada
2  FROM `copper-stacker-389812.Projeto_final.Projeto_final_dematamento`
3  GROUP BY causador
4  ORDER BY area_desmatada DESC
```

Resultados da consulta

INFORMAÇÕES DO JOB | RESULTADOS | JSON | DETALHES DA E

Linha	causador	area_desmatada
1	agricultura	69576.0
2	outros	1213.0
3	mineracao_ilegal	336.0
4	expansao_urbana	65.0
5	mineracao	10.0
6	causa_natural	7.0

Total de área desmatada:

Total de área des... EXECUTAR SALVAR CONSULTA

```
SELECT
  ano,
  SUM(area_desmat_acum) AS total_area_desmatada
FROM
  `copper-stacker-389812.Projeto_final.projeto_final_estufa`
GROUP BY
  ano
ORDER BY
  ano
```

Resultados da consulta

INFORMAÇÕES DO JOB | RESULTADOS | JSON | DETALHE

Linha	ano	total_area_desmatad
1	1960	842754
2	1961	1685508
3	1962	2528262
4	1963	3371016
5	1964	4213770
6	1965	5056524
7	1966	5899278
8	1967	6742032
9	1968	7584786
10	1969	8427540

Área desmatada acumulada ao longo dos anos

Total de Área De... EXECUTAR SALVAR CONSULTA

```
SELECT ano, SUM(areaha) AS area_desmatada
FROM `copper-stacker-389812.Projeto_final.Projeto_final_dematamento`
GROUP BY ano
ORDER BY ano
```

Resultados da consulta

INFORMAÇÕES DO JOB | RESULTADOS | JSON | DETALHES DA E

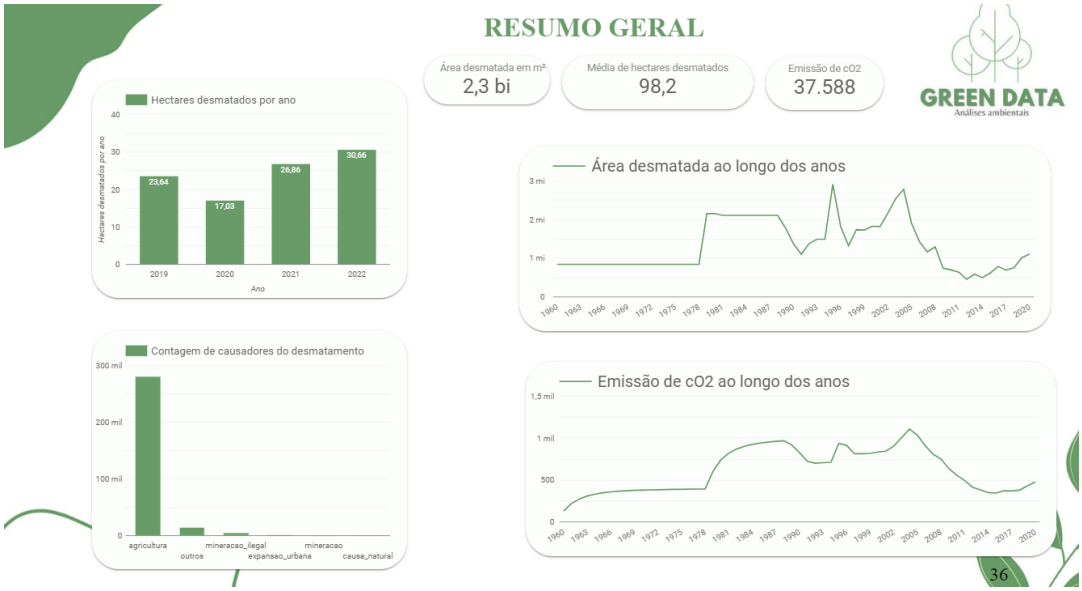
Linha	ano	area_desmatada
1	2019	13214.0
2	2020	15978.0
3	2021	19018.0
4	2022	22997.0

Dashboards:

Looker Studio:

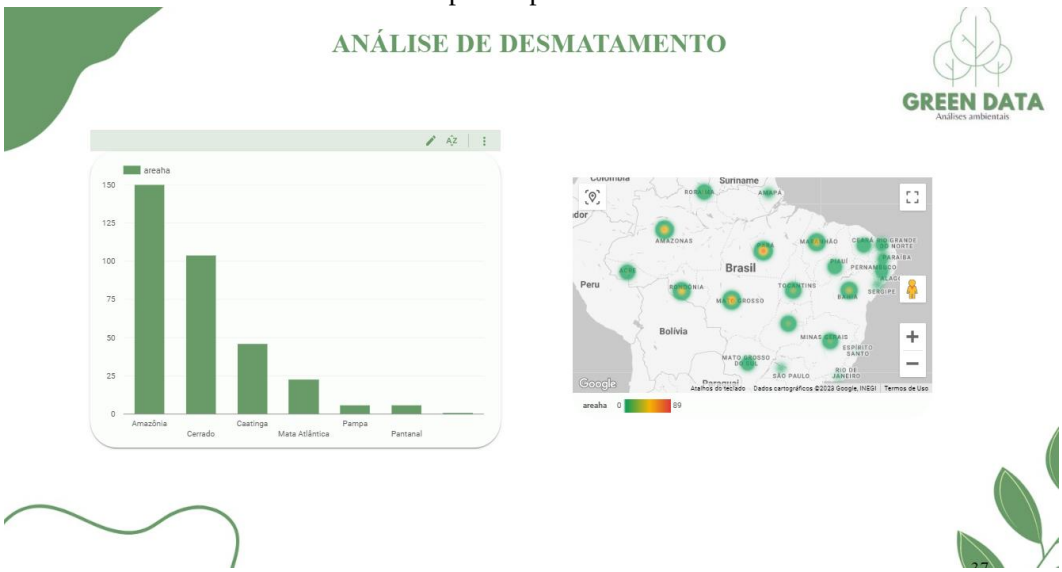
Dashboard 1:

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.



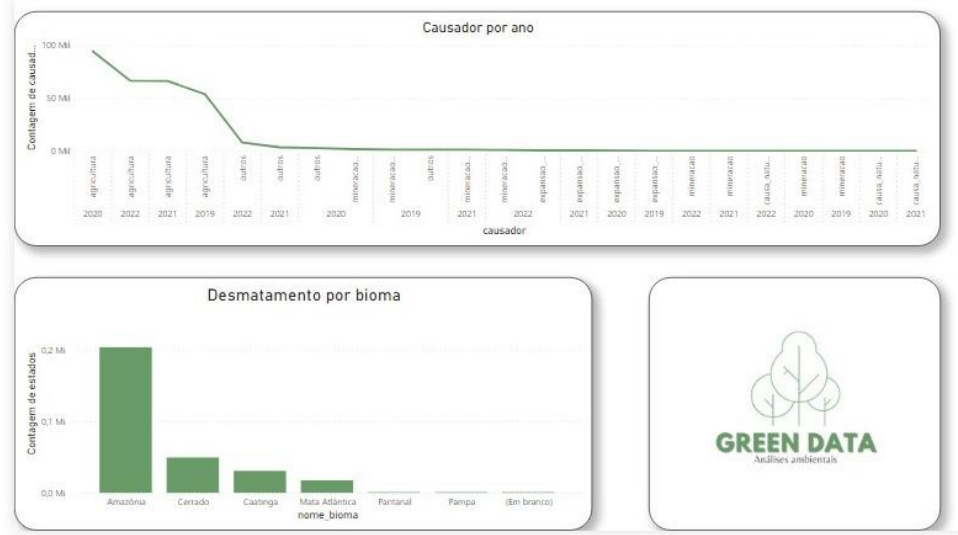
Dashboard 2:

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.

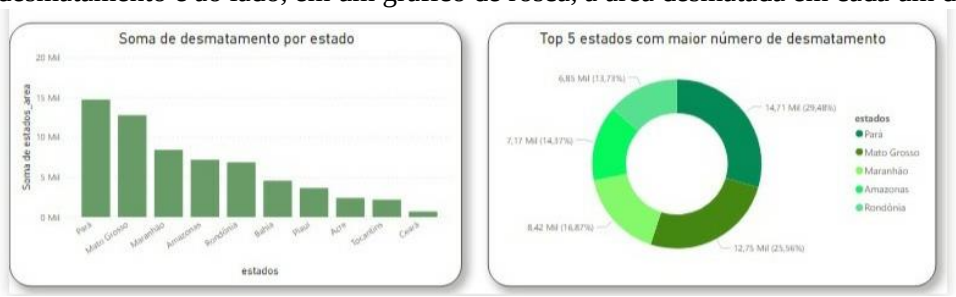


Power BI

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).