

PASSO A PASSO E AUTOAVALIAÇÃO

Aluna: Jullya Letícia Marques da Silva

Eu pensei em utilizar minha base em graduação em economia para pensar num problema que pudesse ser trabalhado, então decidi pesquisar sobre a taxa de evasão dos alunos do ensino superior de instituições públicas, que é um tema bastante discutido e algo que os alunos sentem e veem acontecer com bastante frequência, principalmente nos cursos de exatas. Além disso, tive curiosidade de saber se havia diferença entre a taxa encontrada para os alunos que terminaram o ensino médio em escola pública e alunos que terminaram o ensino médio em escola privada. Assim, o primeiro passo foi coletar os microdados do Censo Superior que é disponibilizado no site do Inep (<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados>) entre os anos de 2009 e 2021.



Assim, criei um bucket no Google Cloud Storage para armazenar os dados coletados

← Detalhes do bucket ATUALIZAR

Intervalos > evasao_ensino_superior > Inep > censo_superior

FAZER UPLOAD DE ARQUIVOS CARREGAR PASTA CRIAR PASTA TRANSFERIR DADOS GERENCIAR RETENÇÕES FAZER O DOWNLOAD EXCLUIR

Filtrar apenas pelo prefixo do nome Filtro Filtrar objetos e pastas Mostrar dados excluídos

<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	Acesso público
<input type="checkbox"/>	2009/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2010/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2011/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2012/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2013/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2014/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2015/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2016/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2017/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2018/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2019/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2020/	—	Pasta	—	—	—	—
<input type="checkbox"/>	2021/	—	Pasta	—	—	—	—

Porém, eu não tenho (ainda) conhecimento o suficiente para responder minhas questões utilizando ETL da própria plataforma, então coloquei esses dados no python (localmente) para tratar os dados e formular um banco pronto para análise. Assim, importei, li os dados, os transformei e calculei a taxa de evasão, como mostrado no print abaixo. Observação: a proxy utilizada foi explicada no documento do relatório disponível no repositório do GitHub, assim como o script completo utilizado para o tratamento dos dados.

```
SELEÇÃO COLUNAS DE INTERESSE

colunas_ies = ['NU_ANO_CENSO', 'CO_IES', 'NO_IES', 'SG_IES', 'NO_REGIAO_IES', 'SG_UF_IES', 'TP_CATEGORIA_ADMINISTRATIVA']
colunas_cursos = ['NU_ANO_CENSO', 'CO_IES', 'SG_UF', 'QT_ING', 'QT_CONC', 'TP_ORGANIZACAO_ACADEMICA', 'TP_CATEGORIA_ADMINISTRATIVA',
                  'QT_ING_PROCESCPUBLICA', 'QT_CONC_PROCESCPUBLICA', 'QT_ING_PROCESCPRIVADA', 'QT_CONC_PROCESCPRIVADA',
                  'CO_CINE_AREA_GERAL', 'NO_CINE_AREA_GERAL']

LEITURA DE DADOS

# Transformações iniciais nos dados de cursos
pipeline = lambda x: (
    read_csv(x, encoding='ISO-8859-1', sep=';', usecols=colunas_cursos, low_memory=False)
    .rename(columns = {'NU_ANO_CENSO': 'ANO'})
    .query("TP_CATEGORIA_ADMINISTRATIVA <= 2 & CO_CINE_AREA_GERAL == 5 & ANO >= 2011")
    .groupby(['ANO', 'CO_IES', 'CO_CINE_AREA_GERAL', 'NO_CINE_AREA_GERAL'])
    .agg(QT_ING = ('QT_ING', 'sum'), QT_CONC = ('QT_CONC', 'sum'),
         QT_INGpb = ('QT_ING_PROCESCPUBLICA', 'sum'), QT_CONCpb = ('QT_CONC_PROCESCPUBLICA', 'sum'),
         QT_INGpv = ('QT_ING_PROCESCPRIVADA', 'sum'), QT_CONCpv = ('QT_CONC_PROCESCPRIVADA', 'sum'))
    .reset_index()
)
df_cursos = concat((pipeline(f) for f in cursos))
```

Restringi as instituições em apenas públicas federais e estaduais como mostra em “TP_CATEGORIA_ADMINISTRATIVA <= 2”, restringi a área relacionada aos cursos a apenas Ciências naturais, matemática e estatística como mostra

“CO_CINE_AREA_GERAL == 5” e por fim, para uma melhor análise dos dados obtidos, decidi coletar apenas dados a partir de 2011, que foi o ano posterior à criação do Sisu e o ano em que ele ganhou mais impulso “ANO >=2011”. Além disso, concatenei os dois tipos de tabela “cursos” e “ies” para me dar informações mais completas sobre as instituições, os cursos, utilizando o código das instituições como chave.

Para evitar erro, restringi os limites numéricos para inferior= 0, superior= 100 e calculei a taxa por um group_by que me desse o máximo de precisão, como mostrado no print abaixo. Além disso, no script também possui algumas análises feitas no próprio python, só a nível de informação.

INDICADORES DE EVASÃO


```
# Preparação dos dados
# Funções
xlag = lambda x: x.shift(5)
def dropout(x, y): return round(clip((1-(x/y))*100,0,100),2)
df = (
    df_cursos.sort_values(['CO_IES', 'ANO'])
    .assign(QT_ING5 = lambda x: x.groupby(['CO_IES', 'CO_CINE_AREA_GERAL', 'NO_CINE_AREA_GERAL']).QT_ING.transform(xlag),
    EVASAO = lambda x: dropout(x.QT_CONC, x.QT_ING5),
    QT_ING5pv = lambda x: x.groupby(['CO_IES', 'CO_CINE_AREA_GERAL', 'NO_CINE_AREA_GERAL']).QT_INGpv.transform(xlag),
    EVASAOpv = lambda x: dropout(x.QT_CONCpv, x.QT_ING5pv),
    QT_ING5pb = lambda x: x.groupby(['CO_IES', 'CO_CINE_AREA_GERAL', 'NO_CINE_AREA_GERAL']).QT_INGpb.transform(xlag),
    EVASAOpb = lambda x: dropout(x.QT_CONCpb, x.QT_ING5pb))
    .dropna(subset=['QT_ING5'])
    .merge(df_ies, on='CO_IES', how='inner')
)
df.head(7)
```

Python

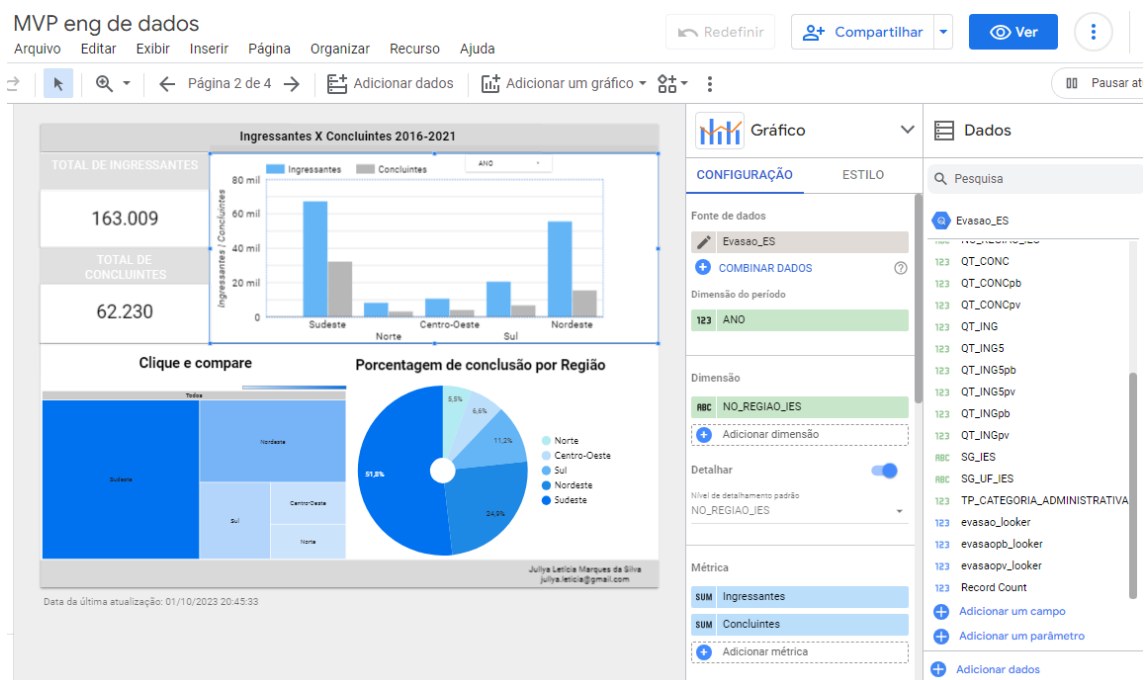
Depois de salvo o csv pronto para a análise, carreguei ele no BigQuery do google como banco de dados final para a análise e lá fiz algumas descrições sobre minhas variáveis, como nos prints abaixo (não estão todas):

The screenshot shows the Google BigQuery interface. On the left is the Explorer pane with a search bar and a list of datasets. The 'Evasao_ES' dataset is selected. The main pane shows the 'ESQUEMA' (Schema) tab for the 'Evasao_ES' table. It includes a filter bar and a table with the following columns: Nome do campo, Tipo, Modo, Chave, Compilação, Valor padrão, and Tags de políticas. The table lists two fields: 'ANO' and 'CO_IES', both of type INTEGER and nullable.

Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas
ANO	INTEGER	NULLABLE				
CO_IES	INTEGER	NULLABLE				

<input type="checkbox"/>	Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas 	Descrição
<input type="checkbox"/>	ANO	INTEGER	NULLABLE					Ano de referência do Censo da Educação Superior
<input type="checkbox"/>	CO_IES	INTEGER	NULLABLE					Código único de identificação da IES
<input type="checkbox"/>	CO_CINE_AREA_GERAL	INTEGER	NULLABLE					Código de identificação da área geral, conforme adaptação da Classificação Internacional Normalizada da Educação Cine/Unesco
<input type="checkbox"/>	NO_CINE_AREA_GERAL	STRING	NULLABLE					Nome da área geral, conforme adaptação da Classificação Internacional Normalizada da Educação Cine/Unesco
<input type="checkbox"/>	QT_ING	FLOAT	NULLABLE					Quantidade de ingressantes
<input type="checkbox"/>	QT_CONC	FLOAT	NULLABLE					Quantidade de concluintes
<input type="checkbox"/>	QT_INGpb	FLOAT	NULLABLE					Quantidade de ingressantes que terminaram o ensino médio em escolas públicas
<input type="checkbox"/>	QT_ING5pb	FLOAT	NULLABLE					Quantidade de ingressantes (5 anos anterior ao ano do censo) que terminaram o ensino médio em escolas públicas
<input type="checkbox"/>	EVASAOpb	FLOAT	NULLABLE					Taxa de evasão calculada terminaram o ensino médio em escolas públicas
<input type="checkbox"/>	ANO_MIN	INTEGER	NULLABLE					Ano que a instituição ingressou no censo superior do Inep
<input type="checkbox"/>	NO_REGIAO_IES	STRING	NULLABLE					Nome da região geográfica da sede administrativa ou reitoria da IES
<input type="checkbox"/>	SG_UF_IES	STRING	NULLABLE					Sigla da Unidade da Federação da sede administrativa ou reitoria da IES
<input type="checkbox"/>	TP_CATEGORIA_ADMINISTRATIVA	INTEGER	NULLABLE					Tipo de Categoria Administrativa da IES
<input type="checkbox"/>	NO_IES	STRING	NULLABLE					Nome da IES
<input type="checkbox"/>	SG_IES	STRING	NULLABLE					Sigla da IES

Decidi fazer minha análise pelo Looker Studio, outra plataforma da Google, importando os dados diretamente do BigQuery, onde estruturei alguns gráficos interativos para responder as perguntas iniciais. Link do dashboard: (<https://lookerstudio.google.com/reporting/a2312f2a-0534-442e-82ca-fc5da152997d>).



Dentro do Looker, eu adicionei mais 3 campos: “evasao_looker”, “evasaopb_looker” e “evasaopv_looker”, que são, respectivamente, a taxa de evasão, taxa de evasão de egressos de escola pública e taxa de evasão de egressos de escola privada, como mostrado no print abaixo:

The screenshot shows the Looker field editor interface. At the top, there are two labels: 'Nome do campo' (Field Name) and 'Código do campo' (Field Code). Below 'Nome do campo' is the text 'evasao_looker'. Below 'Código do campo' is the text 'calc_ysrmhg7cad'. In the center, there is a section labeled 'Fórmula' (Formula) with a question mark icon. To the right of this section is a button labeled 'FORMATAR FÓRMULA'. Below the 'Fórmula' section, there is a text area containing the following formula: `IF(ROUND((1-SUM(QT_CONC)/SUM(QT_ING5))*100,4)<0,0,ROUND((1-SUM(QT_CONC)/SUM(QT_ING5))*100,4))`. The formula is displayed on a grid background.

Por fim, como explicado no relatório, consegui responder e analisar as características desse problema, sugerindo uma solução nas considerações finais. Além disso, possivelmente, ainda há alguns dados faltantes (advindos do próprio Inep) fazendo com que a taxa de certos estados fugisse do resultado médio encontrado, como mostrado no dashboard criado pelo looker. Também há dados faltantes nas siglas das instituições, mas essa falta não foi impactante na análise, então não foi mexida.

Sobre minhas dificuldades, eu confesso que fiquei mais confusa sobre a proposta do “MVP”, pois nunca tinha feito nada desse gênero antes, então fiquei com medo de fugir do objetivo proposto e fazer algo totalmente diferente, desde a produção do documento “relatório” e quanto de informação eu precisaria escrever sobre meus dados nesse passo a passo e autoavaliação. Além disso, fiquei receosa por fazer certa parte no python, mas ainda assim, espero ter atingido o objetivo proposto do MVP.