



# **Apresentação da Analise Microdados Senso Educação Básica 2022**

## **Avaliação de Competências Técnicas de Analíticas para Desenvolvedor de Banco de Dados em Python (DBM)**

Analise introdutória ao censo 2022 com foco em prospecção de escolas privadas de ensino particular, com apresentação em pequenos insights sobre os microdados censo escolar 2022.

**Anderson Lopes**



## Cenário Inicial

Extração da base disponibilizada em :

<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-escolar>

Foram baixados os arquivos do censo de 2016 a 2022 em formato “.csv “ para o google drive no caminho:

[https://drive.google.com/drive/folders/1UrDDx0DKRgX4SR5nrssTdVLnjAOwcVFT?usp=drive\\_link](https://drive.google.com/drive/folders/1UrDDx0DKRgX4SR5nrssTdVLnjAOwcVFT?usp=drive_link)

As análises iniciais foram realizadas no ambiente do Google Colab em:

[https://colab.research.google.com/github/AndersonLopesbh/I.S/blob/main/Analises%20IS/Script\\_python/Analises\\_IS.ipynb](https://colab.research.google.com/github/AndersonLopesbh/I.S/blob/main/Analises%20IS/Script_python/Analises_IS.ipynb)

O script foi compartilhado no GitHub em:

[https://colab.research.google.com/github/AndersonLopesbh/I.S/blob/main/Analises%20IS/Script\\_python/Analises\\_IS.ipynb](https://colab.research.google.com/github/AndersonLopesbh/I.S/blob/main/Analises%20IS/Script_python/Analises_IS.ipynb)

Os insumos como apresentações e exemplo de base CSV se encontram em:

<https://github.com/AndersonLopesbh/I.S/tree/main>

## Informações iniciais

Embora realizado o download de 7 bases no formato .CSV e criado script para a leitura das mesmas, foram utilizados somente os dados de 2022, uma vez que o ambiente do Google Colab só fornece 12GB de RAM, o que em primeiro momento não atendeu toda a massa de dados e análises, assim as linhas não utilizadas foram comentadas:

```
[ ] # criando data set/frame com os dados / OBS: o foco da análise esta no arquivo 2022, devido ao limite de memoria do google colabs, mantereí os arquivos comentados

#csv_2016 = '/content/Drive_analise_IS/MyDrive/ANALISE_IS/microdados_2016.csv'
#csv_2017 = '/content/Drive_analise_IS/MyDrive/ANALISE_IS/microdados_2017.csv'
#csv_2018 = '/content/Drive_analise_IS/MyDrive/ANALISE_IS/microdados_2018.csv'
#csv_2019 = '/content/Drive_analise_IS/MyDrive/ANALISE_IS/microdados_2019.csv'
#csv_2020 = '/content/Drive_analise_IS/MyDrive/ANALISE_IS/microdados_2020.csv'
#csv_2021 = '/content/Drive_analise_IS/MyDrive/ANALISE_IS/microdados_2021.csv'
csv_2022 = '/content/Drive_analise_IS/MyDrive/ANALISE_IS/microdados_2022.csv'

[ ] # Carregando os dados CVS para Data Frames, já trabalhando a leitura destes dados que se encontram separados por
#';' que não se encontram no formato utf8 e lendo-os como Latin-1.

#df_2016 = pd.read_csv(csv_2016,header=0,sep=';',encoding='Latin-1')
#df_2017 = pd.read_csv(csv_2017,header=0,sep=';',encoding='Latin-1')
#df_2018 = pd.read_csv(csv_2018,header=0,sep=';',encoding='Latin-1')
#df_2019 = pd.read_csv(csv_2019,header=0,sep=';',encoding='Latin-1')
#df_2020 = pd.read_csv(csv_2020,header=0,sep=';',encoding='Latin-1')
#df_2021 = pd.read_csv(csv_2021,header=0,sep=';',encoding='Latin-1')
df_2022 = pd.read_csv(csv_2022,header=0,sep=';',encoding='Latin-1');
```



Sobre os arquivos temos as seguintes mensurações iniciais:

- Censo 2016 - 370 colunas e **237506 linhas**
- Censo 2017 - 370 colunas e **236481 linhas**
- Censo 2018 - 370 colunas e **236460 linhas**
- Censo 2019 - 370 colunas e 228521 linhas
- Censo 2020 - 370 colunas e 224229 linhas
- Censo 2021 - 370 colunas e 221140 linhas
- Censo 2022 - 385 colunas e 224649 linhas

Pode-se perceber uma diferença no numero de registros a partir de 2019, além do acréscimo de 15 colunas no censo de 2022 onde temos pela leitura do Pandas :

```
<class 'pandas.core.frame.DataFrame'> RangeIndex: 224649 entries, 0 to 224648 Columns: 385 entries, NU_ANO_CENSO to QT_TUR_ESP_CE dtypes: float64(358), int64(11), object(16) memory usage: 659.9+ MB
```

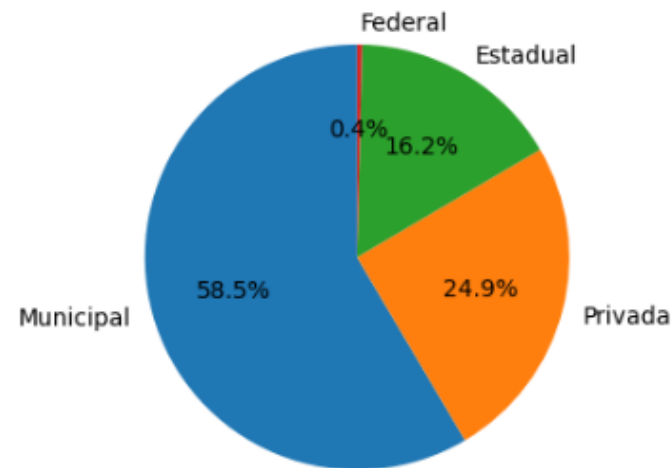
Para a análise, as 15 colunas acrescidas, não serão utilizadas para manter a modelagem, junto aos censos anteriores, caso os mesmos sejam acrescidos as análises futuramente.



### Informações iniciais:

Realizando as primeiras limpezas, com o foco no público alvo, observando as entidades “Ativas” segundo o censo, possuímos um total de 24,9% de entidades privadas, a serem prospectadas\*

Distribuição de Tipos de Entidades Ativas no Censo 2022

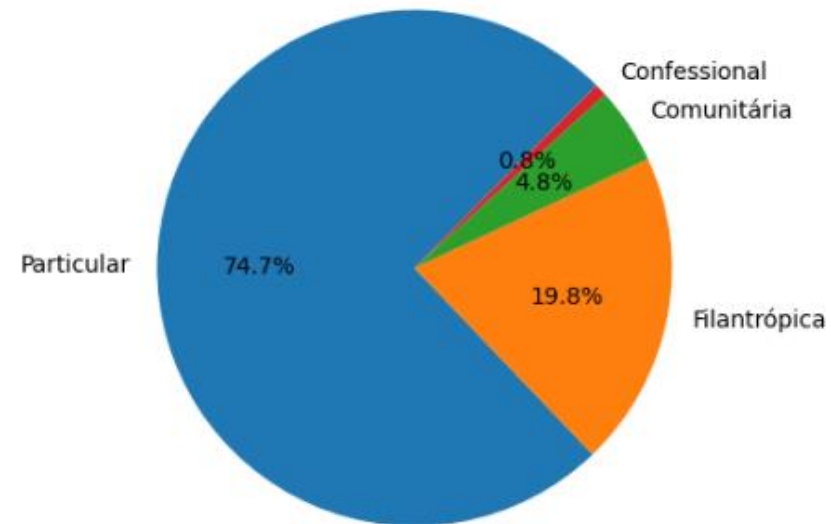


\* Todas as análises partem do pressuposto anama vez que não possuo acesso a base de cliente atuais, estou considerando que estamos prospectando desde o ponto 0 (sem clientes)

### Informações iniciais:

Ao ampliarmos nossa visão nos dados (Drill Down) veremos que no conjunto de entidades privadas, temos uma fatia de 74,7% de entidades de ensino particular, publico alvo da nossa prospecção e para onde vamos apontar nosso DataFrame (df\_pvp)

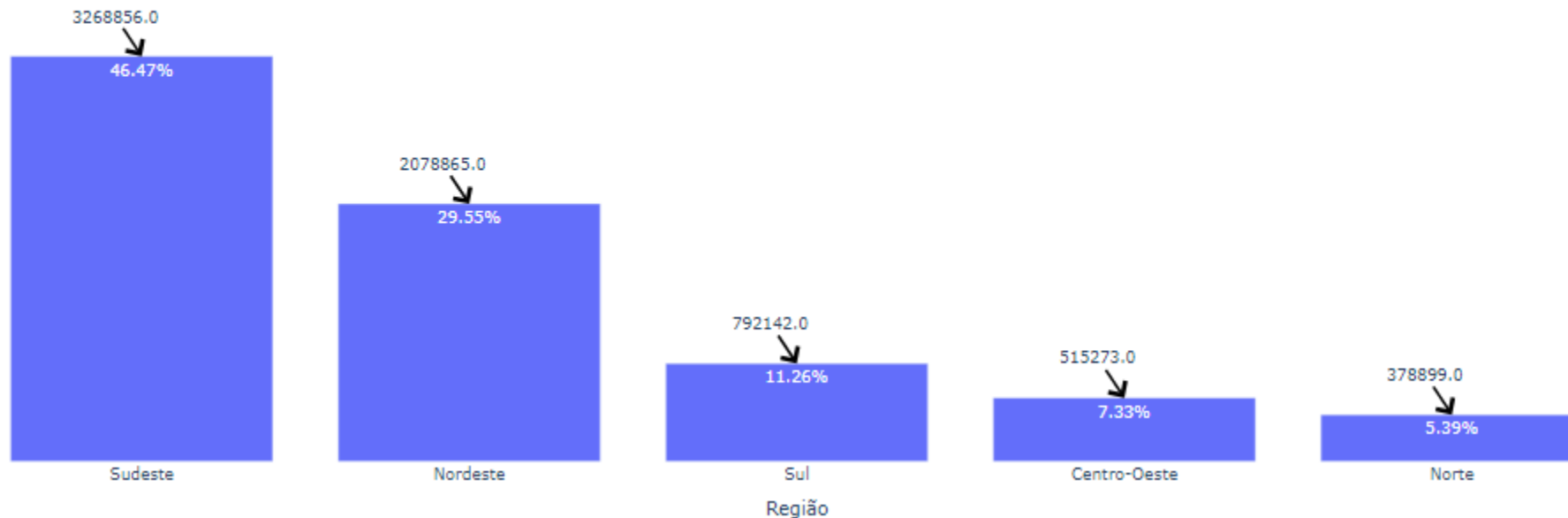
Distribuição categorias de Entidades Privadas Ativas no Censo 2022



## Matriculas por regiões

Conforme abaixo, podemos verificar que temos uma concentração de matriculas nas regiões Sudeste e Nordeste e um índice de matrículas muito baixo na região norte do país, o que pode indicar que nossa estratégia de prospecção , pode encontrar maiores sucessos na região Sudeste e Nordeste, para uma melhor estratégia. caberia um estudo das melhores entidades ramificadas nestas regiões.

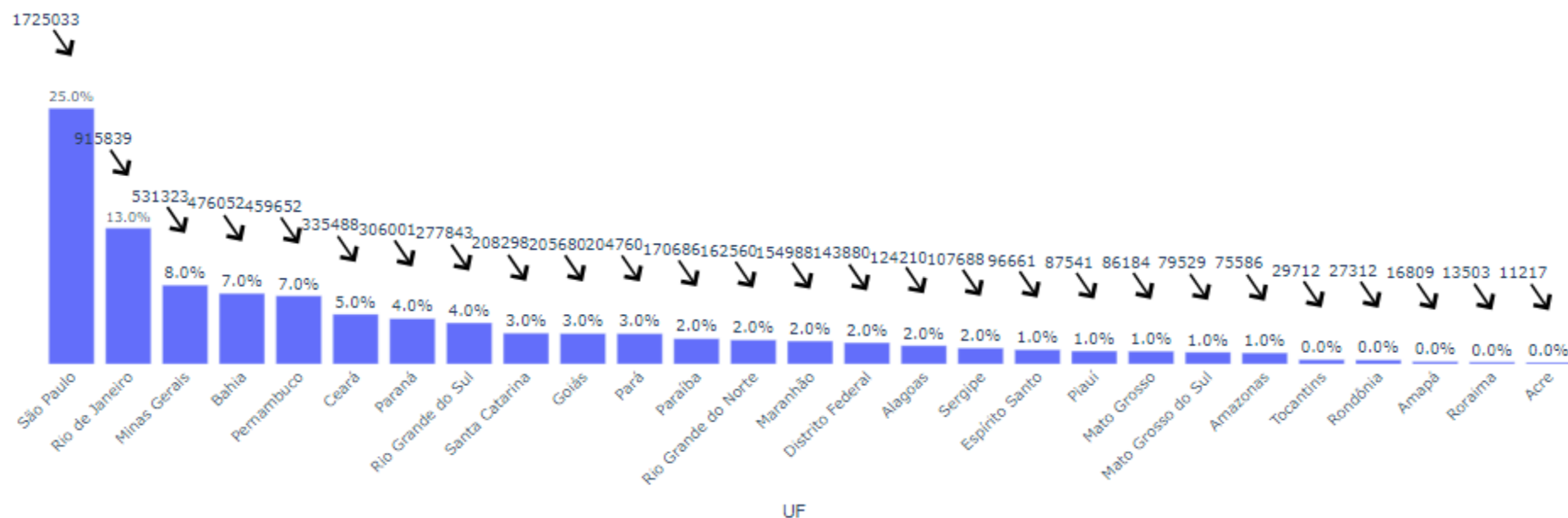
Matriculas por Região



## Matriculas por estados

Temos no ranking 6 estados principais, sendo um deles da origem da empresa, para os estados mais ao norte, temos que ter em mente o tipo de material mais adequado (tendo em vista a logística e infraestrutura) já que podemos ter um déficit na infraestrutura de comunicações, não sendo aplicável os meios de mídia digital.

Matriculas por Estado

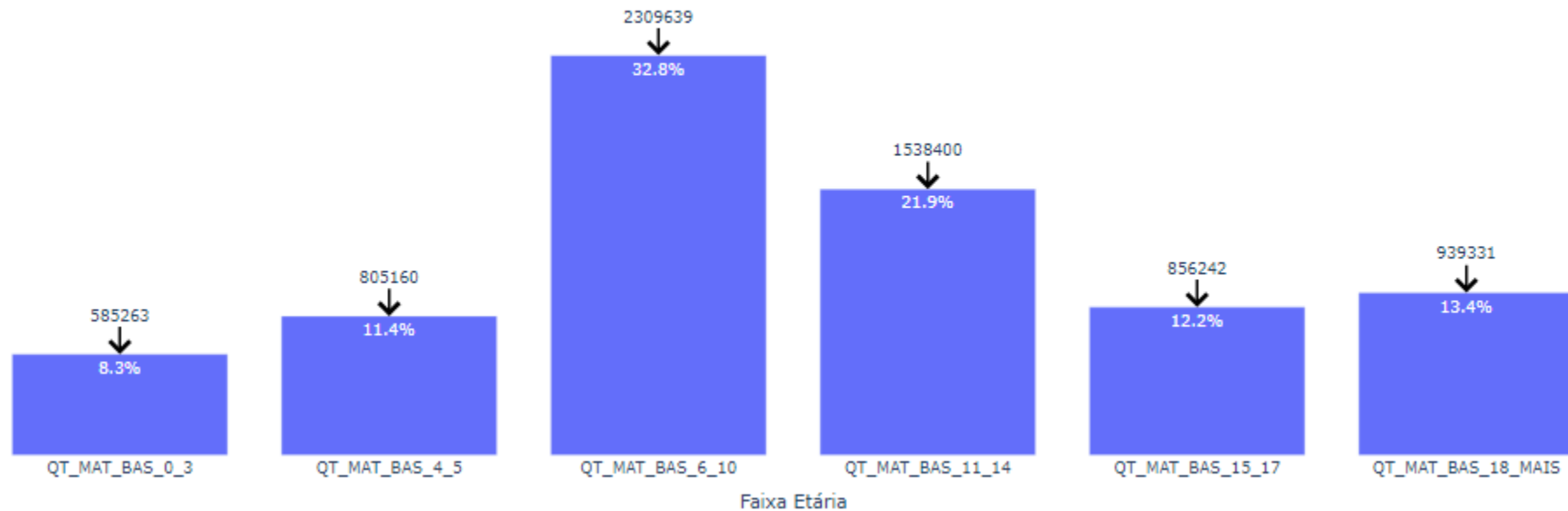




## Matriculas por Faixas Etárias

Observando ao nível Brasil, temos um grande público concentrado na faixa etária de 6 a 10 anos, um modelo mais exato de prospecção, seria a estatística com % de aprovação e reprovação, para criarmos uma previsão de distribuição de materiais na logística e precificação, caso a empresa tenha em contrato o fornecimento de materiais em caso de reprovações.

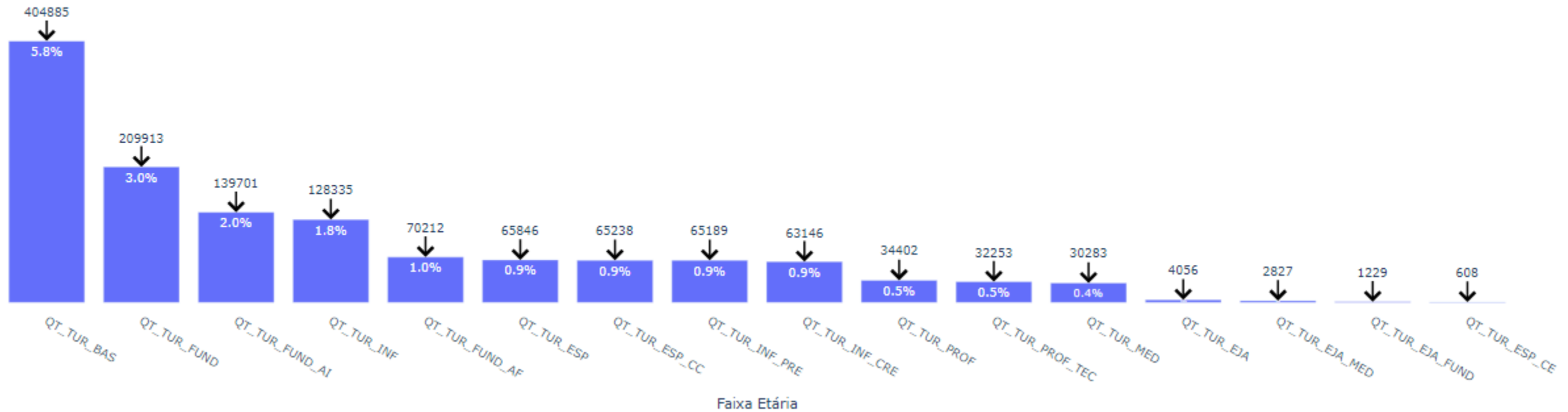
Matrículas por Faixa Etária - Brasil



## Matriculas por Turmas

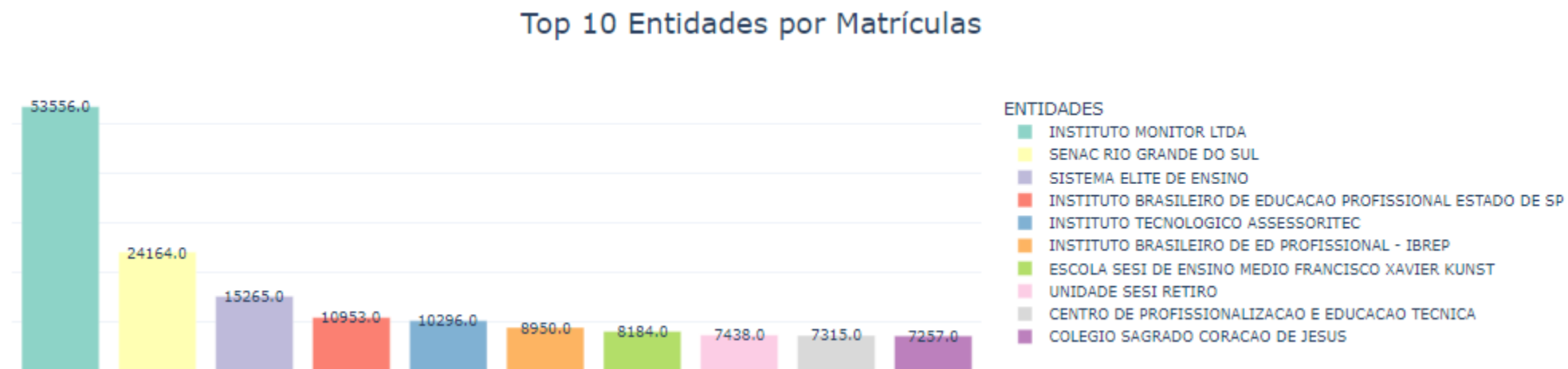
Quando observamos a formação a tipo de turmas, por quantidade, não temos uma definição clara no dicionário de dados, de qual faixa etária esta inclusa neste tipo de turma, mas poderemos perceber que temos um nicho muito especifico apontando para a educação infantil e primeiros anos: Turma de Educação Básica, Fundamental, Fundamental Anos Iniciais e educação infantil.

### Matrículas Turmas



## Top 10 entidades

Estão listadas abaixo as Top 10 entidades, com os maiores números de matrículas registradas em 2022, chamando a atenção que muitas delas são entidades profissionalizantes, o que levanta a questão se deveriam estar constando como entidade de ensino privado e particular no censo.





### **Conclusão:**

A base fornecida pelo censo, nos fornece informações com pouca granularidade, quanto ao que tange a faixa etária, que deveria ser mais aberta, o que impede dimensionar as migrações de alunos por idade e criar outras faixas etárias personalizadas e adequar nossos materiais de acordo com esta necessidade.

Do ponto de vista estratégico das análises, seria necessário termos em mãos, a lista de entidades parceiras, para refinarmos a base, através de cruzamentos de dados retirando os clientes fidelizados, podendo projetar nossos índices e metas de prospecção.

Outras análises de infraestrutura das entidades seriam interessantes, caso necessário, como por exemplo se estas entidades possuem espaço físico para música, quem sabe não poderíamos criar um material voltado ao áudio visual com foco em aprendizagem por música, principalmente nas fases iniciais da infância, onde há facilidade em aprender novas linguagens através do canto.

Em pontos mais distantes do Brasil, para fins de economia na logística, seria necessário avaliar a infraestrutura para implantação de aulas inteiramente virtuais e digitais, sendo desnecessário o envio de materiais físicos.

São muitas especulações e possibilidades, cabe ao negócio as perguntas corretas e a nós o conhecimento do negócio em sua totalidade, para criarmos laços cognitivos, para criação de insights realista e pontuais para criação de micro e macro estratégias de conversão de novos clientes.