**README da análise e limpeza inicial dos dados.**

**Texto em vermelho: apenas para orientação inicial. Será eliminado posteriormente.**

**Em elaboração: padronizar formatação do texto junto com os demais documentos semelhantes do projeto; revisão e aprovação pelos demais membros da equipe**

- *Download* do arquivo MICRODADOS\_ENEM\_2021.csv na página<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>;

- Carregamento dos dados deste arquivo em um objeto do tipo *Data Frame* da biblioteca *Pandas* da linguagem *Python*;

- Análise preliminar deste conjunto de dados, verificando adequação ao dicionário do ENEM 2021, tipos dos dados, estatísticas de colunas com valores vazios e seleção das colunas e variáveis que farão parte do contexto da análise posterior;

- Trabalhos de limpeza e tratamento dos dados, consistindo em:

- eliminação de linhas com indicadores socioeconômicos vazios;

- análise da condição de treineiro no Enem 2021 e verificação de sua aplicabilidade nas análises;

- criação de uma coluna adicional para as médias dos candidatos;

- mudança de tipo *float* da variável “Q005” para *int*;

- preenchimento de valores da variável “TP\_STATUS\_REDACAO” vazias com o valor 4;

- mudança de tipo *float* da variável “TP\_STATUS\_REDACAO” para *int*;

- preenchimento com 0 (zero) das células de notas vazias;

- cálculo e preenchimento das médias de cada inscrito;

- Criação de um *Data Frame* apenas com as variáveis de interesse para o projeto;

**A partir deste ponto, o *Data Frame* criado é carregado na *Bigquery* da Google Cloud.**