**Template para entrega do projeto da disciplina**
**Ciência de Dados e Inteligência Artificial**
**Fase 1**

| | |
|---|---|
| **Nome do estudante** | **KHENED BATISTA DOS SANTOS** |

## SÍNTESE DO CONJUNTO DE DADOS COMO APRESENTADO NA SUA ORIGEM

*Apresente uma breve descrição, acerca do conjunto de dados escolhido, indicando as informações contidas nos dados selecionados. É importante, ao apresentar sua síntese, que você inclua elementos que permitam compreender o significado e a aplicação dos dados escolhidos.*

O conjunto de dados reúne informações sobre emissões de Carteiras de Trabalho e Previdência Social (CTPS) no Brasil de 2020 a 2022, documento essencial para registrar vínculos empregatícios. Com mais de 485.000 registros, os dados incluem o tipo de protocolo (primeira ou segunda via), datas de emissão e protocolo, além de detalhes sobre o órgão emissor e o município de atendimento.

As informações pessoais dos solicitantes – como sexo, escolaridade, raça/cor e estado civil – permitem uma visão detalhada do perfil dos trabalhadores que buscaram o documento nesse período. Esse panorama possibilita identificar padrões demográficos e regionais, auxiliando na compreensão das características da força de trabalho formal no Brasil.

## DESCRIÇÃO DO SEU INTERESSE EM EXPLORAR ESSE CONJUNTO DE DADOS

*Em até 3 parágrafos, explique o interesse em explorá-lo.*

O interesse em explorar este conjunto de dados surge da relevância social e econômica da Carteira de Trabalho e Previdência Social (CTPS) no Brasil, um documento essencial para o registro formal de vínculos empregatícios. Analisar as emissões de CTPS oferece uma oportunidade única de entender o perfil dos trabalhadores que buscam o documento, o que inclui informações demográficas, educacionais e regionais. Esse tipo de análise é relevante para identificar quem está acessando o mercado de trabalho formal e como diferentes fatores, como escolaridade e região, influenciam esse acesso.

Além disso, o período coberto pelos dados (2020 a 2022) inclui a pandemia de COVID-19, um evento que impactou profundamente o mercado de trabalho. Compreender como o perfil de emissão de CTPS foi afetado durante esses anos pode ajudar a revelar mudanças na dinâmica de empregos formais e o impacto socioeconômico da pandemia no Brasil. Isso pode trazer insights valiosos para o planejamento de políticas públicas voltadas à empregabilidade e inclusão social.

Por fim, essa análise é de grande utilidade para órgãos governamentais e empresas interessadas em traçar estratégias de desenvolvimento regional. Entender as variações na emissão da CTPS por estado e município permite identificar áreas com maior demanda por formalização do trabalho, o que pode guiar ações para facilitar o acesso a empregos formais, com impacto direto na economia e no bem-estar da população.

## FINALIDADE DO CONJUNTO DE DADOS ESCOLHIDO

*Descreve para que serve esse conjunto de dados selecionado, ou seja, a razão pela qual eles foram coletados e estruturados.*

A finalidade deste conjunto de dados é registrar e acompanhar as emissões de Carteiras de Trabalho e Previdência Social (CTPS) no Brasil, um documento essencial para a formalização do trabalhador no mercado de trabalho. Esses dados foram coletados e estruturados para fornecer informações detalhadas sobre cada emissão de CTPS realizada entre 2020 e 2022, incluindo dados demográficos, educacionais e regionais dos solicitantes.

A coleta e estruturação desse conjunto de dados possibilitam que órgãos governamentais, pesquisadores e profissionais de diversas áreas analisem padrões e tendências nas emissões de CTPS, o que é útil para entender o perfil dos trabalhadores formais no país. Além disso, esses dados são importantes para a formulação de políticas públicas voltadas à inclusão no mercado de trabalho, ao desenvolvimento regional e ao apoio à empregabilidade, especialmente em contextos de mudanças econômicas e sociais, como a pandemia de COVID-19.

## QUANTIDADE DE LINHAS E COLUNAS DO CONJUNTO DE DADOS

*Descreva apenas a quantidade de linhas e de colunas selecionadas.*

O conjunto de dados possui 485.430 Linhas e 18 colunas.

## QUAL O FORMATO QUE O CONJUNTO DE DADOS É DISPONIBILIZADO

*CSV, JSON, XLSX, etc.*

O arwuivo está no formato CSV.

Escolha pelo menos 10 colunas totalmente preenchidas (as mais importantes) e, para cada coluna (inclusive para a coluna alvo):

- Qual o nome e o que representa?
- Qual o tipo de dados? Nominal/Ordinal/Numérico/Data e/ou hora?
- Quais são os valores considerados válidos?
- Quantos valores distintos aparecem na coluna?
- Qual o menor e o maior valor, e qual a moda?
- Os valores da coluna são numéricos? Qual a média e qual o desvio-padrão? Qual a mediana?

*Você pode apresentar essas informações da maneira como entender mais conveniente, como por exemplo em uma tabela, parágrafos, etc. É importante que sejam inseridas as imagens e as evidências do passo a passo realizado na atividade.*

1. **Tipo Protocolo**

- **Representação**: Tipo de emissão da CTPS (por exemplo, "1ª Via" ou "2ª Via").
- **Tipo de Dados**: Nominal.
- **Valores Válidos**: "1ª Via", "2ª Via" e possivelmente outros tipos específicos de emissão.
- **Valores Distintos**: 2.

- **Moda**: "1ª Via".
- **Numérico**: Não aplicável.

2. **Data Protocolo**

- **Representação**: Data em que o protocolo de emissão foi realizado.
- **Tipo de Dados**: Data.
- **Valores Válidos**: Datas entre 2020 e 2022.
- **Valores Distintos**: Várias datas exclusivas.
- **Menor e Maior Valor**: Datas do início e final do período (exemplo: "2020-01-01" a "2022-12-31").
- **Numérico**: Não aplicável.

3. **Tipo CTPS**

- **Representação**: Classificação da CTPS, geralmente indica o tipo de trabalhador.
- **Tipo de Dados**: Nominal.
- **Valores Válidos**: Predominantemente "Brasileiro".
- **Valores Distintos**: 1 valor predominante.
- **Moda**: "Brasileiro".
- **Numérico**: Não aplicável.

4. **Nome Órgão**

- **Representação**: Nome do órgão emissor da CTPS.
- **Tipo de Dados**: Nominal.
- **Valores Válidos**: Nomes de diferentes órgãos emissores.
- **Valores Distintos**: Vários órgãos emissores.
- **Moda**: Exemplo, "SINEBAHIA".
- **Numérico**: Não aplicável.

5. **Nome Município Órgão**

- **Representação**: Nome da cidade onde está o órgão emissor.
- **Tipo de Dados**: Nominal.
- **Valores Válidos**: Cidades brasileiras.
- **Valores Distintos**: Muitas cidades.
- **Moda**: Cidade com maior número de emissões (exemplo: "Salvador").
- **Numérico**: Não aplicável.

6. **Sexo**

- **Representação**: Gênero do titular da CTPS.
- **Tipo de Dados**: Nominal.
- **Valores Válidos**: "Masculino" e "Feminino".
- **Valores Distintos**: 2.
- **Moda**: Exemplo, "Masculino".
- **Numérico**: Não aplicável.

7. **Nível Escolaridade**

- **Representação**: Grau de escolaridade do titular.

- **Tipo de Dados**: Ordinal.
- **Valores Válidos**: "Fundamental Incompleto", "Médio Completo", "Superior Completo", entre outros.
- **Valores Distintos**: Vários níveis de escolaridade.
- **Moda**: Nível mais frequente.
- **Numérico**: Não aplicável.

8. **Raça e Cor**

- **Representação**: Cor ou raça autodeclarada do titular.
- **Tipo de Dados**: Nominal.
- **Valores Válidos**: "Branco", "Pardo", "Negro", entre outros.
- **Valores Distintos**: Várias opções.
- **Moda**: Categoria mais frequente.
- **Numérico**: Não aplicável.

9. **Estado Civil**

- **Representação**: Estado civil do titular da CTPS.
- **Tipo de Dados**: Nominal.
- **Valores Válidos**: "Solteiro", "Casado", "Divorciado", etc.
- **Valores Distintos**: Diversos estados civis.
- **Moda**: Estado civil mais frequente.
- **Numérico**: Não aplicável.

10. **Data Nascimento (Coluna alvo)**

- **Representação**: Data de nascimento do titular, para calcular a idade.
- **Tipo de Dados**: Data.
- **Valores Válidos**: Datas válidas, correspondentes a idades de trabalhadores.
- **Valores Distintos**: Muitas datas exclusivas.
- **Menor e Maior Valor**: As menores e maiores idades no conjunto.
- **Numérico**: Não aplicável diretamente, mas a idade derivada pode ser numérica.

Qual a oportunidade para um projeto de ciência de dados foi identificada a partir dessa análise? Justifique sua resposta!

*Descreva as potencialidades que o conjunto de dados selecionado apresenta do ponto de vista da Ciência de Dados e Inteligência Artificial.*

A análise desse conjunto de dados apresenta uma excelente oportunidade para explorar o perfil dos trabalhadores formais no Brasil, especialmente durante o período impactado pela pandemia de COVID-19. Esse contexto oferece potencial para desenvolver modelos preditivos e identificar padrões no acesso ao mercado de trabalho formal com base em fatores demográficos, como escolaridade, sexo, região e estado civil. Por exemplo, é possível prever a demanda por emissões de CTPS em diferentes regiões ou investigar como características sociodemográficas influenciam o tipo de emissão (primeira ou segunda via).

Do ponto de vista da Ciência de Dados e da Inteligência Artificial, o conjunto é particularmente valioso para a criação de modelos de segmentação, classificação e análise de séries temporais. A segmentação pode ajudar a identificar grupos de trabalhadores com necessidades

específicas, como regiões com maiores índices de informalidade que buscam formalização. Já a análise de séries temporais pode capturar tendências de emissão de CTPS ao longo do tempo, revelando padrões sazonais ou o impacto de eventos econômicos. Essas informações são fundamentais para que governos e empresas formulem políticas de empregabilidade e ações de inclusão no mercado de trabalho, com base em insights gerados a partir de dados reais e robustos.

## LINK PARA ACESSO AOS DADOS SELECIONADOS

*Inclua aqui o link para o conjunto de dados escolhido. No caso de ter sido feito um recorte/amostragem do conjunto original, documentar quais foram os passos feitos para produzir o recorte.*

https://empregabrasil.mte.gov.br/wp-content/uploads/2024/06/Dados_CTPS_2020-a-2022.zip