Pontifícia Universidade Católica do Rio de Janeiro – PUC-RJ
Disciplina de Engenharia de Dados
Projeto MVP
Aluno: Marcos Gabriel Fernandes Andrade

Objetivo do MVP

Este trabalho tem como objetivo analisar e investigar, de forma crítica e multidimensional, os padrões e tendências dos casos de suicídio no Brasil entre os anos de 2014 e 2018, buscando compreender não apenas os números, mas também os fatores sociais e geográficos que podem estar associados a esse fenômeno. Pretendo identificar possíveis correlações entre variáveis como faixa etária, gênero, região do país, raça e nível de escolaridade, a fim de contribuir para uma discussão mais embasada sobre a temática.

Perguntas a serem realizadas:

- **Pergunta 01: Analisar o número de suicídios por ano.**

  Esta pergunta visa avaliar se os óbitos por suicídio estão crescendo a cada ano, conforme o período da análise.

- **Pergunta 02: Analisar o número de suicídios por estado.**

  Esta pergunta tem como objetivo identificar possíveis disparidades regionais nas taxas de suicídio, permitindo compreender se há estados com números significativamente maiores ou menores e quais fatores locais podem influenciar esses dados.

- **Pergunta 03: Analisar o número de suicídios por gênero.**

  Esta pergunta busca entender a distribuição dos casos de suicídio entre homens e mulheres, visando identificar se há uma prevalência significativa em um dos gêneros

- **Pergunta 04: Analisar o número de suicídios por faixa etária**

  Esta pergunta tem como finalidade mapear as faixas etárias mais afetadas pelo suicídio, permitindo identificar se há grupos de risco.

- **Pergunta 05: Identificar o número de suicídio por estado civil.**

  Esta pergunta visa investigar se o estado civil (solteiro, casado, divorciado, viúvo) está associado a uma maior ou menor incidência de suicídios, contribuindo para compreender o impacto das relações interpessoais nesse contexto.

- **Pergunta 06: Analisar o número de suicídios por gênero e raça**

  Esta pergunta busca cruzar dados de gênero e raça para identificar possíveis interseccionalidades que possam influenciar as taxas de suicídio, permitindo uma análise mais detalhada de grupos específicos que podem estar em maior vulnerabilidade.

- **Pergunta 07: Identificar se o grau de escolaridade interfere na taxa de suicídio.**

  Esta pergunta busca avaliar se há uma correlação entre o nível de escolaridade e as taxas de suicídio, permitindo entender se fatores educacionais podem influenciar o comportamento suicida.

Modelagem

Modelo da tabela

Para o desenvolvimento deste MVP, os dados foram armazenados em uma única tabela, seguindo o modelo *flat file* (arquivo plano). A estrutura é composta por:

- 16 colunas: Representando as categorias das informações.
- 58.634 registros (linhas): Onde cada entrada corresponde a um dado individual.

Nesse formato, cada linha do arquivo equivale a um registro completo, enquanto as colunas organizam os atributos conforme suas classificações.

Linhagem dos dados

Os dados utilizados para a análise deste MVP foram obtidos por meio da plataforma **Kaggle**, um repositório público de conjuntos de dados. A fonte original desses registros é o **DATASUS** (*Departamento de Informática do Sistema Único de Saúde*), garantindo a confiabilidade e relevância das informações para o escopo do projeto.

A base de dados pode ser acessada pelo link:
https://www.kaggle.com/datasets/psicodata/dados-de-suicdio-no-brasil-2014-a-2018

Catálogo de dados

Este catálogo descreve a estrutura, os atributos e as descrições dos dados utilizados neste projeto, garantindo transparência e facilitando a interpretação das análises realizadas.

As colunas foram organizadas conforme os termos a seguir:

- **Nome variável** (nome da coluna no banco de dados)
- **Tipo de variável** (categórica ou numérica)
- **Descrição** (descrição completa do nome)
- **Valores permitidos** (valores presentes no banco de dados)

Nome variável	Tipo de variável	Descrição	Valores permitidos
ESTADO	categórica	Estado de residência	AC, AL, AM, AP, BA, CE, DF, ES, GO, MA, MG, MS, MT, PA, PB, PE, PI, PR, RJ, RN, RO, RR, RS, SC, SE, SP, TO
ANO	numérica	Ano do óbito	2014, 2015, 2016, 2017, 2018
CIRCOBITO	categórica	Circunstância do óbito	Suicidio
DTOBITO	numérica	Data do óbito	2014, 2015, 2016, 2017, 2018
DTNASC	numérica	Data de nascimento	1930 a 2010
SEXO	categórica	Gênero	Feminino, masculino, NA (não apresentado)
RACACOR	categórica	Raça	Amarela, branca, indígena, parda, preta, NA
ESTCIV	categórica	Estado civil	Solteiro, casado, viúvo, separado judicialmente, união consensual, NA.
ESC	numérica	Escolaridade	Nenhuma, 1 a 3 anos, 4 a 7 anos, 8 a 11 ano, 9 a 11 anos, 12 anos e mais, NA
OCUP	categórica	Ocupação	Segue-se a tabela CBO2002
CODMUNRES	categórica	Município de residência do falecido	Segue-se a classificação do IBGE
LOCOCOR	categórica	Local de ocorrência do óbito	Hospital, outro estabelecimento de saúde, domicílio, via pública, outros, NA
ASSITMED	categórica	Assistência médica	Sim, não, NA
CAUSABAS	numérica	Causa básica do óbito - Código CID 10	X60 a X84
IDADE	numérica	Idade do falecido	08 a 104
MES	numérica	Mês de ocorrência do óbito	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,12

Extração, Carga e Transformação dos Dados

Para a ingestão e preparação dos dados no ambiente **Databricks**, foram executadas as seguintes etapas:

1. Carga do Arquivo CSV

- **Configuração do DBFS (Databricks File System)**: Sistema de arquivos distribuído utilizado para interação com o armazenamento em nuvem.
- **Estruturação do Diretório**: Criação de uma pasta específica no catálogo para armazenamento do arquivo.

- **Upload do CSV**: Carregamento do arquivo para o ambiente Databricks.
- **Leitura dos Dados**: Utilização do comando df = spark.read.csv para importação do dataset no notebook.

2. Transformação dos Dados

Filtragem de Registros

- Aplicação de um comando **DELETE em SQL** para remoção de registros onde a coluna **"circunstância do óbito"** não apresentava o valor **"Suicídio"**, resultando na exclusão de **3.150 linhas**.
- Inclusão de condição adicional para eliminação de **valores nulos**, garantindo a qualidade dos dados.

Tratamento de Duplicidade

Identificou-se que a base de dados continha colunas duplicadas:

- **CAUSABAS** (original)
- **CAUSABAS_0** (duplicada, com os mesmos valores)

Como ambas as colunas armazenavam os mesmos dados, optou-se pela remoção de CAUSABAS_0 para evitar inconsistências.

Passos Executados:

1. **Tentativa inicial**: Uso do comando DROP COLUMN (SQL) para excluir CAUSABAS_0. **Não suportado** na versão do Databricks utilizada.
2. **Solução implementada**:

- Criação de uma nova tabela temporária via SQL, selecionando todas as colunas exceto CAUSABAS_0.
- Exclusão da tabela original e tentativa de renomeação da tabela temporária com RENAME (SQL). **Bloqueada** devido ao armazenamento em Amazon S3.
- Resolução final: Renomeação utilizando o comando RENAME do Spark.

Análise

A. Qualidade dos dados

Para garantir a confiabilidade e integridade dos dados utilizados neste estudo, foram realizadas verificações rigorosas em múltiplas etapas, assegurando que a base estivesse consistente, completa e dentro do escopo da pesquisa.

1. Validação do Período da Pesquisa (2014–2018)

- Foi executado um script em SQL para confirmar se todos os registros estavam dentro do intervalo temporal definido (2014 a 2018).
- Resultado: Todas as linhas da base apresentavam datas válidas, sem desvios ou inconsistências no período analisado.

2. Verificação de Valores Nulos em Campos principais

- Colunas verificadas: Estado e Ano (campos essenciais para a análise).
- Método: Consulta SQL para identificar registros nulos ou incompletos.
- Resultado: Nenhum valor nulo foi encontrado nessas colunas, garantindo que:

  Todos os registros possuíam localização geográfica definida (Estado).
  Todos os óbitos estavam corretamente datados (Ano).

3. Tratamento de Valores Nulos em Dados Categóricos

- Contexto: Algumas colunas, como escolaridade e estado civil, apresentavam valores ou não apresentados (NA).
- Decisão: Optou-se por não excluir esses registros devido a três fatores:

  1. Impacto na Análise Principal: A remoção afetaria a contagem de óbitos por suicídio, comprometendo a representatividade dos dados.
  2. Natureza dos Dados: Essas informações são categóricas e, muitas vezes, não coletadas pelo sistema de saúde devido às limitações operacionais (ex.: falta de registro no momento do atendimento).
  3. Preservação da Integridade: Como as colunas principais (Estado e Ano) estavam completas é possível determinar que o registro representa de fato um dado real.

B. Solução do problema

Os dados analisados neste projeto de MPV nos revelam um aumento preocupante de 16% nos casos de suicídio no Brasil entre os anos de 2014 e 2018, com padrões consistentes com pesquisas globais.

Conforme os resultados encontrados, foi possível identificar os seguintes cenários:

- Homens são 3 vezes mais vulneráveis que mulheres, refletindo questões culturais e de acesso à saúde mental.
- A faixa etária 30-39 anos concentra o maior número de casos, associada a pressões profissionais e pessoais.
- Solteiros representam 50% dos registros, possivelmente pela falta de redes de apoio estruturadas. Este resultado pode apresentar supernotificação, considerando que

muitos indivíduos mantêm relacionamentos estáveis sem efetuar a formalização do estado civil.

- Homens brancos e pardos lideram as estatísticas, enquanto escolaridade média (4–11 anos) mostra maior risco, indicando uma relação complexa entre frustração socioeconômica e saúde mental.
- Diferenças regionais persistem mesmo após ajuste populacional, sugerindo influência de fatores locais (como acesso a serviços e condições de vida).

Esses resultados reforçam a urgência de políticas públicas direcionadas, especialmente para grupos em maior risco (homens adultos, solteiros e população com escolaridade intermediária), com foco em prevenção, diagnóstico precoce e quebra de estigmas culturais. A base de dados utilizada mostrou-se confiável para orientar as decisões.

As respostas para as perguntas elaborados no objetivo do projeto foram respondidas separadamente nos comentários do notebook, após o resultado de cada consulta.

Autoavaliação

Quando comecei este projeto, confesso que me senti um pouco preocupado. Este é o meu primeiro projeto de MVP, então estava com receio de como seria o meu desenvolvimento no projeto. No entanto, com o apoio dedicado de vocês professores durante os plantões de dúvidas, fui gradualmente compreendendo a estruturação adequada do MVP e quais serias os passos necessários para a entrega do projeto. Encontrar a base de dados no Kaggle logo nos primeiros dias de pesquisa me deu mais tempo para me familiarizar com o Databricks, permitindo testar soluções e corrigir os erros que apareciam no decorrer das análises. Optei por realizar uma abordagem mais direta, com isso, foi possível responder todas as perguntas elaboradas no meu objetivo. Para deixar minhas respostas mais completas, pesquisei informações técnicas sobre cada resultado encontrado, buscando sempre explicar melhor os dados com fontes confiáveis. Ao entregar este trabalho, me sinto feliz e realizado, tenho plena consciência de que há espaço para aprimoramento, mas também reconheço o valioso esforço e aprendizado adquirido. Quero agradecer especialmente aos professores pela paciência e pelo tempo dedicado conosco. Não poderia deixar de registrar também um elogio para a toda a turma de colegas, que todos os dias se ajudavam no Discord e WhatsApp, compartilhando dúvidas, soluções e incentivos. Esse apoio fez toda a diferença.