1° trabalho de mineração

Marcelo da Silva Landim

2025-11-04

1.0 Descrição do trabalho

1.1 problema proposto

O seguinte trabalho tem como objetivo preparar um dataset para uma futura análise de risco, focando em micro-segmentos de risco para pessoa jurídica (PJ) no estado de Alagoas (AL). O processo envolve a filtragem e preparação de dados para posterior análise de risco creditício.

1.2 Descrição do processo

1.2.1 instalação dos pacotes utilizados

- **DuckDB**: Utilizado para processamento eficiente de dados em memória
- **dplyr**: Empregado para manipulação e transformação de dados
- **tibble**: Usado para melhor visualização de dataframes.

1.2.2 Configuração do Ambiente

- Conexão com banco de dados DuckDB em memória
- Definição do caminho do arquivo CSV contendo os dados originais.

1.2.3 Filtro aplicado nos Dados

- Aplicação de filtros específicos:
  - UF = 'AL' (Estado de Alagoas)
  - cliente = 'PJ' (Pessoa Jurídica)
- Execução via consulta SQL para eficiência.

1.2.4 Análise utilizadas

- Verificação da estrutura dos dados filtrados
- Estatísticas descritivas das variáveis numéricas
- Confirmação dos filtros aplicados.

1.3 Resultados Esperados

- Dataset filtrado contendo apenas PJs de Alagoas
- Base preparada para análises de risco subsequentes
- Documentação do processo de preparação.

2.0 Código R Utilizado

```
# Projeto- Preparação de Datasets para Análise de Risco
# Opção escolhida: 1 - Micro-Segmentos de Risco (Pessoa jurídica)
# UF escolhida: Alagoas (AL)

# Carregar as bibliotecas necessárias
library(tinytex)
library(duckdb)
library(dplyr)
library(tibble)

# Configurar conexão com DuckDB (banco de dados em memória)
con <- dbConnect(duckdb::duckdb(), dbdir = ":memory:")

#  o caminho para o arquivo CSV baixado
caminho_csv <- "C:/Users/Katia/OneDrive/Documentos/marcelo/mineração de dados/minera-o/planil

# Filtros aplicados: UF = 'AL' AND cliente = 'PJ'
consulta_sql <- "
SELECT
    *
FROM
    read_csv_auto(?)
WHERE
    UF = 'AL'
    AND cliente = 'PJ'
```

```
"

# Executar a consulta e carregar os resultados em um dataframe
df_filtrado <- dbGetQuery(con, consulta_sql, list(caminho_csv))

# Fechar a conexão com o banco de dados
dbDisconnect(con)

# Análise exploratória do dataframe resultante
# Visualização rápida da estrutura dos dados
cat("=== ESTRUTURA DO DATAFRAME (glimpse) ===\n")
```

```
=== ESTRUTURA DO DATAFRAME (glimpse) ===
```

```
glimpse(df_filtrado)
```

```
Rows: 12,520
Columns: 23
$ data_base                       <date> 2025-08-31, 2025-08-31, 2025-08-31, 20~
$ uf                              <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL~
$ tcb                             <chr> "Bancário", "Bancário", "Bancário", "Ba~
$ sr                              <chr> "S1", "S1", "S1", "S1", "S1", "S1", "S1~
$ cliente                         <chr> "PJ", "PJ", "PJ", "PJ", "PJ", "PJ", "PJ~
$ ocupacao                        <chr> "-", "-", "-", "-", "-", "-", "-", "-",~
$ cnae_secao                      <chr> "PJ - Administração pública, defesa e s~
$ cnae_subclasse                  <chr> "-", "-", "-", "-", "-", "-", "PJ - Adm~
$ porte                           <chr> "PJ - Indisponível                      ~
$ modalidade                      <chr> "PJ - Capital de giro", "PJ - Financiam~
$ origem                          <chr> "Sem destinação específica", "Com desti~
$ indexador                       <chr> "Flutuantes", "Flutuantes", "Pós-fixado~
$ numero_de_operacoes             <chr> "<= 15", "<= 15", "<= 15", "<= 15", "<=~
$ a_vencer_ate_90_dias            <chr> "27299837,92", "2546283,51", "332153,67~
$ a_vencer_de_91_ate_360_dias     <chr> "67380965,36", "7994412,51", "971099,09~
$ a_vencer_de_361_ate_1080_dias   <chr> "174605193,19", "24637011,88", "2412955~
$ a_vencer_de_1081_ate_1800_dias  <chr> "172585915,84", "24637011,84", "2176193~
$ a_vencer_de_1801_ate_5400_dias  <chr> "353847507,83", "46669004,91", "712940,~
$ a_vencer_acima_de_5400_dias     <chr> "0,00", "0,00", "0,00", "0,00", "0,00",~
$ vencido_acima_de_15_dias        <chr> "0,00", "0,00", "0,00", "0,00", "0,00",~
$ carteira_ativa                  <chr> "795719420,14", "106483724,65", "660534~
$ carteira_inadimplida_arrastada  <chr> "0,00", "0,00", "0,00", "0,00", "0,00",~
$ ativo_problematico              <chr> "0,00", "0,00", "0,00", "0,00", "0,00",~
```

```r
# Estatísticas descritivas das variáveis numéricas
cat("\n=== ESTATÍSTICAS DESCRITIVAS (summary) ===\n")
```

```
=== ESTATÍSTICAS DESCRITIVAS (summary) ===
```

```r
summary(df_filtrado)
```

```
  data_base               uf                 tcb                 sr
 Min.   :2025-08-31   Length:12520       Length:12520       Length:12520
 1st Qu.:2025-08-31   Class :character   Class :character   Class :character
 Median :2025-08-31   Mode  :character   Mode  :character   Mode  :character
 Mean   :2025-08-31
 3rd Qu.:2025-08-31
 Max.   :2025-08-31
  cliente            ocupacao          cnae_secao        cnae_subclasse
 Length:12520       Length:12520       Length:12520       Length:12520
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


   porte             modalidade           origem           indexador
 Length:12520       Length:12520       Length:12520       Length:12520
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


 numero_de_operacoes a_vencer_ate_90_dias a_vencer_de_91_ate_360_dias
 Length:12520        Length:12520         Length:12520
 Class :character    Class :character     Class :character
 Mode  :character    Mode  :character     Mode  :character


 a_vencer_de_361_ate_1080_dias a_vencer_de_1081_ate_1800_dias
 Length:12520                  Length:12520
 Class :character              Class :character
 Mode  :character              Mode  :character
```

```
 a_vencer_de_1801_ate_5400_dias a_vencer_acima_de_5400_dias
 Length:12520                   Length:12520
 Class :character               Class :character
 Mode  :character               Mode  :character
```

```
 vencido_acima_de_15_dias carteira_ativa     carteira_inadimplida_arrastada
 Length:12520             Length:12520       Length:12520
 Class :character         Class :character   Class :character
 Mode  :character         Mode  :character   Mode  :character
```

```
 ativo_problematico
 Length:12520
 Class :character
 Mode  :character
```

```
# Verificação adicional dos filtros aplicados
cat("\n=== VERIFICAÇÃO DOS FILTROS ===\n")
```

```
=== VERIFICAÇÃO DOS FILTROS ===
```

```
cat("Valores únicos na coluna UF:", unique(df_filtrado$UF), "\n")
```

```
Valores únicos na coluna UF:
```

```
cat("Valores únicos na coluna cliente:", unique(df_filtrado$cliente), "\n")
```

```
Valores únicos na coluna cliente: PJ
```

```r
cat("Dimensões do dataframe (linhas x colunas):", dim(df_filtrado), "\n")
```

```
Dimensões do dataframe (linhas x colunas): 12520 23
```