# Implementando um Leitor e Processador CSV

Um arquivo CSV (*Comma-Separated Values*) é um formato de arquivo amplamente utilizado para armazenar e trocar dados tabulares simples. O formato CSV é composto por linhas de texto que representam registros e colunas separadas por delimitadores, geralmente uma vírgula (,). Cada linha no arquivo CSV representa um registro ou uma entrada de dados e é dividida em campos ou valores separados pelo delimitador. Por exemplo, um arquivo CSV que armazena informações sobre funcionários de uma empresa pode ter uma linha para cada funcionário, com os campos separados por vírgulas (Fig. 1). Note que em alguns casos, os dados podem estar faltando, como a idade no terceiro registro.

```
Nome,Sexo,Idade,Profissao,Salario,EC
Gabriela,M,22.0,Gerente de Projetos,3284.34,Divorciado
Renata,F,60.0,Gerente de Projetos,3310.65,Solteiro
Aline,F,,Analista de Sistemas,5912.03,Divorciado
Mariana,M,21.0,Arquiteto,12030.86,Casado
…
```

**Exemplo de arquivo CSV**

O objetivo deste trabalho é construir um leitor de arquivos CSV. O programa deve detectar automaticamente a quantidade de linhas e colunas do arquivo e alocar memória para armazenar os dados. Além disso, o programa deve receber como argumento um arquivo CSV, por exemplo:

```
./csvreader <funcionarios.csv>
```

O programa deve então apresentar um menu para o usuário com as seguintes opções:

```
1) Sumário do Arquivo
2) Mostrar
3) Filtros
4) Descrição dos Dados
5) Ordenação
6) Seleção
7) Dados Faltantes
8) Salvar Dados
9) Fim
```

Descrição das funcionalidades que devem ser implementadas:

**Opção 1 (Sumário do Arquivo):**

Imprimir as variáveis encontradas e a quantidade de linhas do arquivo no seguinte formato. Cada uma das variáveis deve ser classificada como [S]tring ou [N]umérica. Veja o exemplo abaixo:

```
Nome [S]
Sexo [S]
Idade [N]
Profissao [S]
Salario [N]
EC [S]
```

```
6 variaveis encontradas

Pressione ENTER para continuar
```

**Opção 2 (Mostrar):**

Imprimir os *5* primeiros e 5 últimos registros do arquivo formatados, como no exemplo abaixo. Note a formação dos dados, inclusive com linha com pontos dividindo as primeiras das últimas linhas. A última linha deve informar o tamanho do arquivo em memória (linhas e colunas)

```
           Nome Sexo  Idade             Profissao   Salario          EC
0         Lucas    M   29.0              Dentista  19415.01  Divorciado
1         Paula    M   64.0               Gerente   7230.26    Solteiro
2       Luciana    M   47.0         Diretor (CTO)  12991.40       Viúvo
3       Eduardo    M   50.0               Gerente   7140.03    Solteiro
4        Thiago    M   46.0  Analista de Sistemas       NaN       Viúvo
...         ...  ...    ...                   ...       ...         ...
99995      José    F   65.0             Consultor   8946.47    Solteiro
99996    Fatima    F   62.0                Médico  11143.34    Solteiro
99997   Beatriz    F   49.0                Piloto  14601.53    Solteiro
99998  Fernando    F   40.0          Farmacêutico   4842.05  Divorciado
99999  Patricia    F   28.0          Farmacêutico   3753.70      Casado

[100000 rows x 6 columns]

Pressione ENTER para continuar
```

**Opção 3 (Filtros):**

Nesta opção, o usuário deve informar  o nome da variável, filtro e valor. Os filtros disponíveis são, igual (=), maior (>), maior igual (>=), menor (<), menor igual (<=) e diferente (!=). O programa deve imprimir os dados filtrados e/ou gravar todos os dados no arquivo informado pelo usuário. O usuário deve ter a opção de descartar os dados originais e usar os dados filtrados para as próximas operações. Veja o exemplo abaixo:

```
Entre com a variavel: Idade
Escolha um filtro ( == > >= < <= != ): >=
Digite um valor: 60
0           Nome Sexo  Idade      Profissao   Salario          EC
1          Paula    M   64.0        Gerente   7230.26    Solteiro
18        Amanda    F   64.0         Médico   9371.45  Divorciado
28         Diego    M   60.0     Engenheiro  13965.87  Divorciado
29        Sandra    M   62.0     Engenheiro   4281.44    Solteiro
31        Felipe    M   63.0        Gerente  19446.92    Solteiro
...          ...  ...    ...            ...       ...         ...
99972  Guilherme    F   60.0  Diretor (CTO)  11346.02    Solteiro
99976     Marcia    F   63.0  Gerente de TI  22385.03  Divorciado
99991   Anderson    M   60.0   Veterinário  10976.88      Casado
99995       José    F   65.0      Consultor   8946.47    Solteiro
99996     Fatima    F   62.0         Médico  11143.34    Solteiro

[12320 rows x 6 columns]
```

```
Deseja gravar um arquivo com os dados filtrados? [S|N]: S
Entre com o nome do arquivo: filtrados.csv
Arquivo gravado com sucesso
Deseja descartar os dados originais? [S|N]: N
Pressione ENTER para continuar
```

**Opção 4 (Descrição dos Dados):**

A descrição de dados visa apresentar estatísticas da variável selecionada. Para uma dada variável numérica, as seguintes estatísticas devem ser apresentadas:

1. Total de Dados
2. Média,
3. Mediana,
4. Moda (informando quantas vezes o valor aparece),
5. Desvio Padrão,
6. Mínimo,
7. Máximo,
8. Número de valores únicos.

Considere a variável idade, por exemplo:

```
Entre com a variavel: Idade
Contador: 97980
Media: 41.5
Desvio: 13.9
Mediana: 42.0
Moda: 59.0 2151 vezes
Min.: 18.0
Max.: 65.0
Valores unicos: [18.0, 19.0, 20.0, 21.0, 22.0, 23.0, 24.0, 25.0, 26.0, 27.0,
     28.0, 29.0, 30.0, 31.0, 32.0, 33.0, 34.0, 35.0, 36.0, 37.0, 38.0, 39.0,
     40.0, 41.0, 42.0, 43.0, 44.0, 45.0, 46.0, 47.0, 48.0, 49.0, 50.0, 51.0,
     52.0, 53.0, 54.0, 55.0, 56.0, 57.0, 58.0, 59.0, 60.0, 61.0, 62.0, 63.0,
     64.0, 65.0]

Pressione ENTER para continuar
```

Para variáveis não numéricas, você deve apresentar, Total de Dados, Moda e Valores únicos. Por exemplo, considere a variável Sexo:

```
Contador: 97980
Moda: M 49121 vezes
Valores unicos: ['F', 'M']
Pressione ENTER para continuar
```

**Opção 5 (Ordenação):**

Ordenar os dados usando a variável indicada. O programa deve imprimir o cabeçalho dos dados ordenados. O usuário deve ter a possibilidade de gravar o arquivo ordenado. Veja o exemplo abaixo:

```
Entre com a variavel: Profissao
Selecione uma opcao [A]scendente ou [D]escrescente: A
0           Nome Sexo  Idade    Profissao   Salario        EC
22489    Antonio    F   31.0     Advogado  12106.14  Solteiro
61291     Eliane    F   59.0     Advogado  12153.36    Casado
78378  Ana Paula    F   22.0     Advogado  14935.03     Viúvo
34348    Ricardo    M   38.0     Advogado  20948.13    Casado
19898    Beatriz    M   58.0     Advogado   9062.51    Casado
...          ...  ...    ...          ...       ...       ...
92757      Paula    F   58.0  Veterinário  15860.65  Solteiro
15037    Ricardo    F   37.0  Veterinário  10201.33     Viúvo
15039     Marcos    F   20.0  Veterinário  21931.56  Solteiro
61346    Fabiana    M   45.0  Veterinário   3317.57  Solteiro
91099    Gustavo    M   52.0  Veterinário  21982.61    Casado

[100000 rows x 6 columns]

Deseja gravar um arquivo com os dados ordenados? [S|N] S
Entre com o nome do arquivo: profissao.csv
Arquivo gravado com sucesso
Deseja descartar os dados originais? [S|N]: N
Pressione ENTER para continuar
```

**Opção 6 (Seleção):**

Seleciona somente as variáveis informadas pelo usuário. O programa deve imprimir o cabeçalho com as variáveis selecionadas. O usuário deve ter a possibilidade de gravar o resultado no arquivo informado. Veja o exemplo abaixo:

```
Entre com a variaveis que deseja selecionar (separadas por espaço): Profissao
Salario
                        Profissao   Salario
0                       Dentista  19415.01
1                        Gerente   7230.26
2                  Diretor (CTO)  12991.40
3                        Gerente   7140.03
5        Engenheiro de Petróleo    5142.42
...                          ...       ...
99995                 Consultor   8946.47
99996                    Médico  11143.34
99997                    Piloto  14601.53
99998              Farmacêutico   4842.05
99999              Farmacêutico   3753.70
[97980 rows x 2 columns]

Deseja gravar um arquivo com as variáveis selecionadas? [S|N] S
Entre com o nome do arquivo: prof_sal.csv
Arquivo gravado com sucesso
Pressione ENTER para continuar
```

**Opção 7 (Dados Faltantes):**

Como ilustrado na Fig 1, algumas variáveis podem ter dados faltantes ou valores numéricos inválidos (NaN - *Not a Number*) Nesses casos, o programa deve fornecer ao usuário formas de completar os dados para evitar problemas em cálculos e análises futuras ou remover as linhas que tenham dados inválidos. Sendo assim, o programa deve mostrar as seguintes opções para lidar com dados faltantes.

```
1) Listar registros com NaN
2) Substituir pela media
3) Substituir pelo proximo valor valido
4) Remover registros com NaN
5) Voltar ao menu principal
```

1) Listar todas as linhas que tenham algum valor NaN
   Por exemplo:

```
           Nome  Sexo  Idade              Profissao   Salario          EC
4         Thiago    M   46.0   Analista de Sistemas       NaN       Viúvo
208    Ana Paula    M    NaN      Gerente de Vendas  18102.43  Divorciado
280         Jose    M   47.0              Arquiteto       NaN  Divorciado
284        Luana    F   33.0  Engenheiro de Petróleo       NaN  Divorciado
326    Francisca    F    NaN             Engenheiro   3813.43    Solteiro
...          ...  ...    ...                    ...       ...         ...
99653     Felipe    M    NaN                Piloto   8369.20      Casado
99737      Paulo    F   24.0               Dentista       NaN       Viúvo
99756     Thiago    M    NaN          Gerente de TI  14451.69      Casado
99777   Francisco    F   42.0                Piloto       NaN      Casado
99895    Gabriela    F   65.0                Piloto       NaN      Casado

[2020 rows x 6 columns]

Deseja gravar um arquivo com os dados ordenados? [S|N] S
Entre com o nome do arquivo: nan.csv
Arquivo gravado com sucesso
Deseja descartar os dados originais? [S|N]: N
Pressione ENTER para continuar

Pressione Enter para continuar
```

2) Valores inválidos devem ser substituído pela média dos valores
3) Valores inválidos devem ser substituídos pelo próximo valor válido. Se o valor invalido for o último valor da coluna, ele não deverá ser substituído.
4) Nesta opção, todos os registros que contenham valores faltantes ou inválidos nas variáveis indicadas pelo usuário, devem ser excluídos.

Ao fim de qualquer uma das opções, os dados originais devem ser substituídos pelos dados originais.

**Opção 8 (Salvar Dados):**

Grava em disco os dados que estão em memória, por exemplo:

```
Entre com o nome do arquivo: novo.csv
```

```
Arquivo gravado com sucesso.
Pressione Enter para continuar
```

**Opção 9 (Fim):**

O programa deve ser encerrado.

**Como estruturar o programa:**

Além do programa csvreader.c que deve conter a função main, você deve construir as seguintes bibliotecas:

**1) io (.c/.h)**

Funções de leitura e gravação de arquivos. As funções que imprimem na tela também devem estar dentro desta biblioteca.

**2) Requisitos de Implementação:**

Na opção 3, você deve implementar uma função genérica chamada filtro, que recebe um ponteiro para função indicando o filtro que deve ser usado (além de outros argumentos necessários para a sua utilização).

Na opção 3, 5 e 7, os dados filtrados substituem a referência em memória principal apenas se o usuário decidir por descartar os dados originais. Na opção 6, os dados gerados NÃO substituem os dados originais em memória principal.

Para a aplicação de filtros, considere que um NaN é menor do que qualquer valor definido.

**Observações**:

Todas as entradas devem ser validadas. Se o usuário digitar uma opção inválida, o programa deve solicitar que o usuário informe uma opção válida.

O texto em azul corresponde a saída do programa, ou seja, o que será mostrado para o usuário na tela.

O texto em vermelho corresponde a entrada do usuário.

Você deve criar um **makefile** para o seu programa contendo, pelo menos:

- make all
- make clean
- make purge

##########
**IMPORTANTE**: A ENTREGA DO TRABALHO DEVE CONTEMPLAR OS REQUISITOS ELENCADOS NO PLANO DA DISCIPLINA. O NÃO CUMPRIMENTO DOS REQUISITOS DE ENTREGA RESULTARÁ EM UM DESCONTO DE 5 DÉCIMOS NA NOTA FINAL.
##########