Relatório de Qualidade dos Dados

Sarah Souza Pontes

Importância da qualidade dos dados:

- ☐ Tomada de Decisão Precisa: Decisões baseadas em dados precisos tendem a ser mais acertadas e orientadas para os resultados.
- ☐ Eficiência Operacional: Dados de alta qualidade garantem que os processos de negócios sejam executados de forma eficiente e sem interrupções.
- ☐ Conformidade Regulatória: Em muitos setores, há requisitos legais para garantir a precisão e integridade dos dados.
- ☐ Credibilidade: Dados imprecisos podem prejudicar a reputação da organização e a confiança dos clientes, parceiros e partes interessadas.

Aumentar a qualidade dos dados:

Padrões de Qualidade de Dados: Estabelecer critérios claros para o que constitui dados de alta qualidade em sua organização. Isso pode incluir precisão, integridade, consistência e atualidade.

Coleta de Dados: Garanta que os dados sejam coletados de fontes confiáveis e que os processos de entrada de dados sejam precisos e eficientes (tanto a mensuração, como o treinamento da equipe por exemplo causam viés).

Limpeza e Normalização de Dados: Identifique e corrija inconsistências, erros e duplicatas nos dados. Normalize os dados para garantir consistência e integração entre diferentes conjuntos de dados, balancear os dados para análise como foi realizado é uma boa prática .

Implemente Controles de Qualidade de Dados: Desenvolva procedimentos para verificar regularmente a qualidade dos dados, incluindo verificações automáticas e manuais.

Utilize Ferramentas de Qualidade de Dados: Adotar software de limpeza de dados, ferramentas de deduplicação e soluções de gerenciamento de metadados.

Cultura de Qualidade de Dados: Formatação das variáveis e Unificar as colunas e linhas sobre como representar os registros.

Monitoramento Contínuo: Processo contínuo de monitoramento e melhoria da qualidade.

Há dados faltantes, mas pela descrição do banco é provável que sejam valores que não deveriam ser preenchidos devido a característica do dado. Mas, uma boa conduta seria assumir um valor de NAN diferente de vazio. A 'Data_Internação' e ' Data_Óbito' na linha 6845 apresentam inconsistência, pois a internação está ocorrendo após o óbito :2013-05-25 2013-05-24 respectivamente.

Há inconsistência nos dados na coluna 'POPULAÇÃO ESTIMADA 2012' exemplo abaixo nas linhas de 1 a 4 e de 38564 a 38568 com "-" entre números.

```
0       2317-03-06
1       2317-03-06
2       1910-05-09
3       1944-09-08
4       1944-09-08
         ...
38564   3455-05-26
38565   9322-05-17
38566   9322-05-17
38567   9322-05-17
38568   9322-05-17
```

Há registros duplicados como é caso abaixo, poderia ter duas pessoas com o mesmo nome, mas não com os mesmos dados:

```
38548   CELINA MACEDO SIQUEIRA MIRANDA AMORIM      NaN     291080
38549   CELINA MACEDO SIQUEIRA MIRANDA AMORIM      NaN     291080
```

Há valores extremos de idade, o que levanta suspeita dos valores abaixo de 18 anos, porém os valores estão corretos quando compara a data de notificação, nascimento e calcula a idade.

```
[30. 28. 24.  8. 51. 42. 70. 12. 18. 20. 53. 27. 17. 19. 45. 60. 65. 43.
 5. 21. 32.  6. 15. 13. 23. 11. 34.  4. 31. 29. 47. 35. 39. 26. 37.  7.
40. 22. 57.  3. 61. 25. 46. 62.  9. 33.  1. 14. 44. 63. 16. 49. 91. 36.
41. 10.  2. 76. 55. 38. 56. 52. 69. 50. 54. 59. 58. 73. 75. 48.  0. 72.
87. 64. 68. 77. 71. 67. 78. 66. 82. 74. 83. 79. 80. 84. 81. 99. 86. nan
97. 90. 85. 88. 94. 96. 89. 92. 95. 93.]
```

Na variável ID_OCUPA_N os valores são representados por números, mas há uma representação: 'XXX'.

A qualidade da base é duvidosa, precisaria de passar por um processamento com validação dos dados ou imputação dos mesmos.