Qualidade e transformao dos dados do SISPNCD

Gabriel Alves Castro

August 27, 2019

Quantidade de preenchimento dos dados de localidade

Quantidade de colunas:

```
## [1] 512846
```

Quantidade de dados com valor no preenchido por coluna:

```
##        coluna valoresVazios
## 1   SN_DENGUE        512846
## 2  SN_ESQUIST        512846
## 3    SN_LEISH        512846
## 4   SN_FMACUL        512846
## 5    SN_PESTE        512846
## 6   SN_CHAGAS        512846
## 7       SN_FA        512846
## 8  DIST_CENTR             0
## 9      ID_LOC             0
## 10     NM_LOC            52
## 11     ID_CAT             0
## 12    ID_STAT             0
## 13 QT_PRE_LOC             0
## 14 QT_HAB_LOC             0
## 15 DT_CAD_LOC         50576
## 16     ID_DMI        512846
## 17     ID_DSM        512846
## 18    ID_CROQ        490001
## 19     NU_CEP             0
## 20    NU_LONG             0
## 21     NU_LAT             0
## 22 NU_ALT_LOC             0
## 23    DT_ATUA        316511
## 24      QT_PE             0
## 25     QT_ARM             0
## 26  CS_URBRUR             0
## 27      NU_TB             0
## 28   NU_QUART             0
## 29     NU_RES             0
## 30  NU_COMERC             0
## 31   NU_OUTRO             0
## 32   SN_ELETR        512846
## 33    SN_AGUA        512846
## 34   SN_ESGOT        512846
## 35   SN_LAVAN        512846
## 36  SN_CSPRIV        512846
## 37    SN_LIXO        512846
```

```
## 38   SN_TELEF        512846
## 39  SN_TRANSP        512846
## 40   SN_PAVIM        512846
## 41  SN_ESCOLA        512846
## 42  SN_PSAUDE        512846
## 43  SN_ACESSO        512846
## 44 SN_PACSPSF        512846
## 45  QT_CACHOR             0
## 46    SN_GATO        512846
## 47  SN_ROEDOR        512846
## 48 SN_MALARIA        512846
```

Quantidade de dados com valor igual a zero:

```
##        coluna valoresVazios
## 1   SN_DENGUE             0
## 2  SN_ESQUIST             0
## 3    SN_LEISH             0
## 4   SN_FMACUL             0
## 5    SN_PESTE             0
## 6   SN_CHAGAS             0
## 7       SN_FA             0
## 8  DIST_CENTR        491757
## 9      ID_LOC             0
## 10     NM_LOC             8
## 11     ID_CAT             0
## 12    ID_STAT             0
## 13 QT_PRE_LOC        413302
## 14 QT_HAB_LOC        142964
## 15 DT_CAD_LOC             0
## 16     ID_DMI             0
## 17     ID_DSM             0
## 18    ID_CROQ          2347
## 19     NU_CEP        461808
## 20    NU_LONG        510600
## 21     NU_LAT        510360
## 22 NU_ALT_LOC        511389
## 23    DT_ATUA             0
## 24      QT_PE        492035
## 25     QT_ARM        509976
## 26  CS_URBRUR             0
## 27      NU_TB        458610
## 28   NU_QUART        449545
## 29     NU_RES        377477
## 30  NU_COMERC        455881
## 31   NU_OUTRO        448752
## 32   SN_ELETR             0
## 33    SN_AGUA             0
## 34   SN_ESGOT             0
## 35   SN_LAVAN             0
## 36  SN_CSPRIV             0
## 37    SN_LIXO             0
## 38   SN_TELEF             0
## 39  SN_TRANSP             0
```

```
## 40   SN_PAVIM             0
## 41  SN_ESCOLA             0
## 42  SN_PSAUDE             0
## 43  SN_ACESSO             0
## 44 SN_PACSPSF             0
## 45  QT_CACHOR        512846
## 46    SN_GATO             0
## 47  SN_ROEDOR             0
## 48 SN_MALARIA             0
```

Colunas com preenchimento acima de 30%:

```
## [1] "ID_LOC"     "NM_LOC"     "ID_CAT"     "ID_STAT"    "DT_CAD_LOC"
## [6] "DT_ATUA"    "CS_URBRUR"
```

Identificando sujeiras nos dados:

Anos para os quais foram indicados como a data do registro DT_ATUA:

```
##  [1] "2002" "2005" "2001" "2008" NA     "2006" "2004" "2007" "2009" "2010"
## [11] "2014" "2018" "2003" "2000" "2013" "2015" "2016" "2017" "2012" "2011"
## [21] "1906" "2066" "1908" "1907" "502-" "201-" "1910" "708-" "404-" "2210"
## [31] "709-" "2204" "206-" "1944" "207-" "2207" "1966" "1987" "2021" "806-"
## [41] "2099" "1997" "1999" "1996" "2073" "320-" "1994" "1995" "2205" "703-"
## [51] "1973" "1005" "2065" "2206" "1983" "202-" "200-" "2044" "1998" "807-"
## [61] "2087" "1965" "204-" "1984" "2201" "1990" "907-" "2088" "606-" "1974"
## [71] "2208" "707-" "208-" "1991" "607-" "909-"
```

O exemplo dos dados acima, demonstra uma situao que ocorre na maioria massiva dos bancos de dados existentes: Os dados no so perfeitos ou apresentam exatamente o que prometem. No possvel utilizar os dados, sem o devido tratamento. Ou simplesmente, pode no ser possvel utilizar os dados para o objetivo que prometem.

O problema da sujeira tambm est no fato de que, um nico dado sujo, que esteja com valores muito quem do esperado, em uma agregao, iro enviezar todos os dados na visualizao, ou em uma pesquisa. Uma maneira de encontrar estas incongruncias, por meio de variveis identificadoras incoerntes, ou atributos tambm incoerntes. Os quais podem indicar toda uma linha incoerente. Por exemplo: Uma coluna de datas, que possua uma data errada (uma data no futuro) pode contar todas as outras linhas erradas tambm, o que poder invalidar uma visualizao criada com os dados sem nenhum tratamento de dados.

O segundo fator refere-se ao fato de que, para gerar as visualizaes, os dados devem ser transformados para um determinado formato, devem ser unidos segundo suas chaves identificadores, e devem ser agregados segundo os objetivos. Diversas visualizaes, simplesmente no podem ser construdas sem o processo de transformao de dados.

Outro grande problema referente ao nvel de preenchimento dos dados, ou mesmo dos dados disponveis, a cobertura que a amostra trem sobre a populao. A depender da amostra, no possvel utilizar os dados para responder perguntas generalizadas sobre as amostras (perguntas que busquem caracterizar toda a populao(universo) pesquisada(o)).