Aula03\_Trilha03-Análise\_Exploratoria\_Dados

André Gustavo Silva Lovo

03/11/2020

# Introdução a Análise Exploratória de Dados (AED) ou do inglês Exploratory Data Analysis (EDA):

O filosofia de análise de dados que emprega uma variedade de técnicas para:

1. Maximizar insights sobre o conjunto de dados;
2. Descobrir estruturas subjacentes;
3. Extrair variávies importantes;
4. Detectar Anomalias e outliers;
5. Testar hipóteses subjacentes ou premissas iniciais;
6. Desenvolver modelos parsimoniosos; e
7. Determinar um conjunto ótimo de fatores/variáveis para a modelagem.

Checklist da Análise Exploratória de Dados (Passo a passo):

### 1. Formule sua(s) questão(ões)

* Uma questão precisa ou hipótese pode servir como uma ferramenta de redução de dimensão (podemos diminuir a quantidade de trabalho e ganhar tempo quando formulamos questões objetivas e com foco na resolução do problema);

### 2. Leia seus dados

* Os dados de que dispomos para analisar podem vir bem bagunçados. Podem precisar de limpeza. Algumas vezes, alguem já fez o trabalho de limpeza e você não precisará sujar as mãos.

Os dados podem vir de diversas fontes diferentes, em diferentes formatos, aprenda como ler/importá-los.

### 3. Verifique o conteúdo

* Abrir os arquivos usando o head(), tail(), verifique células NA e valores 0 (zero) ou nulos antes do início das análises.
* Verifique as quantidades de linhas e colunas, bem como os tipos de dados de cada coluna, bem como suas caracteristicas.

### 4. Rode str()

1. Veja o topo e o final de seus dados
2. Verifique as quantidades (os “n”s)
3. Valide com pelo uma fonte de dados externa
4. Tente a solução fácil primeiro
5. Desafie sua solução
6. Continue

## Fazendo a programação em R:

# 1- Entrar com as bibliotecas:  
# Biblioteca para ler o arquivo  
if(!require (readr))  
 install.packages("readr")

## Loading required package: readr

library(readr)   
  
# Biblioteca para transformar o arquivo em um dataframe (tabela de trabalho)  
# É um pacote do R que fornece uma versão melhorada de data.frames  
if(!require (data.table))  
 install.packages("data.table")

## Loading required package: data.table

library(data.table)  
  
# Verificar a diferença entre leitura do BD em tabelas x dplyr  
# https://stats4good.github.io/blog/2016/09/09/2016-09-09-dt-dplyr/  
  
# Biblioteca para manipulação de dados  
if(!require (dplyr))  
 install.packages("dplyr")

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(dplyr)  
  
# Biblioteca para trabalharmos com o Markdown  
if(!require (knitr))  
 install.packages("knitr")

## Loading required package: knitr

library(knitr)  
  
if(!require (tidyr))  
 install.packages("tidyr")

## Loading required package: tidyr

library(tidyr)

# 2- Lendo o arquivo csv  
arquivo <- 'ozonesmall.csv'  
dados<- read.csv(arquivo, sep=',')#, dec='.'  
dados <- as.data.table(dados)  
  
#Lendo as 5 primeiras linhas  
head(dados,5)

## State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude Datum  
## 1: 1 113 2 44201 1 32.46797 85.08384 WGS84  
## 2: 56 39 8 44201 1 43.67083 110.59947 NAD83  
## 3: 48 231 1006 44201 1 33.15309 96.11557 NAD83  
## 4: 47 187 106 44201 1 35.95153 87.13701 NAD83  
## 5: 40 27 49 44201 1 35.32010 97.48410 WGS84  
## Parameter.Name Date.Local Time.Local Date.GMT Time.GMT Sample.Measurement  
## 1: Ozone 2014-10-22 02:00 2014-10-22 08:00 0.018  
## 2: Ozone 2014-08-19 19:00 2014-08-20 02:00 0.039  
## 3: Ozone 2014-01-15 09:00 2014-01-15 15:00 0.027  
## 4: Ozone 2014-10-02 18:00 2014-10-02 23:00 0.045  
## 5: Ozone 2014-03-25 04:00 2014-03-25 10:00 0.007  
## Units.of.Measure MDL Uncertainty Qualifier Method.Type Method.Code  
## 1: Parts per million 0.005 NA FEM 47  
## 2: Parts per million 0.005 NA FEM 87  
## 3: Parts per million 0.005 NA FEM 87  
## 4: Parts per million 0.005 NA FEM 47  
## 5: Parts per million 0.005 NA FEM 47  
## Method.Name State.Name County.Name  
## 1: INSTRUMENTAL - ULTRA VIOLET Alabama Russell  
## 2: INSTRUMENTAL - ULTRA VIOLET ABSORPTION Wyoming Teton  
## 3: INSTRUMENTAL - ULTRA VIOLET ABSORPTION Texas Hunt  
## 4: INSTRUMENTAL - ULTRA VIOLET Tennessee Williamson  
## 5: INSTRUMENTAL - ULTRA VIOLET Oklahoma Cleveland  
## Date.of.Last.Change  
## 1: 2015-02-02  
## 2: 2014-10-08  
## 3: 2014-04-18  
## 4: 2014-12-11  
## 5: 2014-05-06

#Lendo as 10 ultimas linhas  
tail(dados,10)

## State.Code County.Code Site.Num Parameter.Code POC Latitude Longitude Datum  
## 1: 56 37 100 44201 1 41.57500 109.20700 NAD83  
## 2: 4 13 7021 44201 1 33.50799 111.75533 NAD83  
## 3: 18 145 1 44201 1 39.61342 85.87065 WGS84  
## 4: 12 105 6006 44201 1 28.02889 81.97222 NAD83  
## 5: 27 53 962 44201 1 44.96524 93.25476 NAD83  
## 6: 18 97 78 44201 1 39.81110 86.11447 WGS84  
## 7: 5 143 6 44201 1 36.01170 94.16744 NAD83  
## 8: 39 165 7 44201 1 39.42689 84.20077 NAD83  
## 9: 6 111 7 44201 1 34.21015 118.87050 NAD83  
## 10: 8 45 21 44201 1 39.41250 107.23056 NAD83  
## Parameter.Name Date.Local Time.Local Date.GMT Time.GMT Sample.Measurement  
## 1: Ozone 2014-03-01 18:00 2014-03-02 01:00 0.024  
## 2: Ozone 2014-04-07 12:00 2014-04-07 19:00 0.052  
## 3: Ozone 2014-07-27 12:00 2014-07-27 17:00 0.032  
## 4: Ozone 2014-03-16 19:00 2014-03-17 00:00 0.035  
## 5: Ozone 2014-01-22 08:00 2014-01-22 14:00 0.015  
## 6: Ozone 2014-06-24 19:00 2014-06-25 00:00 0.014  
## 7: Ozone 2014-03-14 18:00 2014-03-15 00:00 0.047  
## 8: Ozone 2014-06-01 11:00 2014-06-01 16:00 0.057  
## 9: Ozone 2014-04-01 20:00 2014-04-02 04:00 0.042  
## 10: Ozone 2014-04-27 10:00 2014-04-27 17:00 0.042  
## Units.of.Measure MDL Uncertainty Qualifier Method.Type Method.Code  
## 1: Parts per million 0.0050 NA FEM 87  
## 2: Parts per million 0.0050 NA FEM 47  
## 3: Parts per million 0.0050 NA FEM 47  
## 4: Parts per million 0.0050 NA FEM 47  
## 5: Parts per million 0.0050 NA FEM 87  
## 6: Parts per million 0.0050 NA FEM 47  
## 7: Parts per million 0.0050 NA FEM 19  
## 8: Parts per million 0.0050 NA FEM 87  
## 9: Parts per million 0.0050 NA FEM 87  
## 10: Parts per million 0.0015 NA Non-FRM 901  
## Method.Name State.Name County.Name  
## 1: INSTRUMENTAL - ULTRA VIOLET ABSORPTION Wyoming Sweetwater  
## 2: INSTRUMENTAL - ULTRA VIOLET Arizona Maricopa  
## 3: INSTRUMENTAL - ULTRA VIOLET Indiana Shelby  
## 4: INSTRUMENTAL - ULTRA VIOLET Florida Polk  
## 5: INSTRUMENTAL - ULTRA VIOLET ABSORPTION Minnesota Hennepin  
## 6: INSTRUMENTAL - ULTRA VIOLET Indiana Marion  
## 7: INSTRUMENTAL - ULTRA VIOLET Arkansas Washington  
## 8: INSTRUMENTAL - ULTRA VIOLET ABSORPTION Ohio Warren  
## 9: INSTRUMENTAL - ULTRA VIOLET ABSORPTION California Ventura  
## 10: Instrumental - Ultra Violet 2B Model 202 Colorado Garfield  
## Date.of.Last.Change  
## 1: 2014-06-06  
## 2: 2014-09-02  
## 3: 2014-10-17  
## 4: 2014-06-18  
## 5: 2014-06-30  
## 6: 2014-09-03  
## 7: 2014-04-17  
## 8: 2014-07-28  
## 9: 2014-05-27  
## 10: 2015-07-21

#Lendo e exibindo o arquivo de forma resumida e detalhada (importante)  
glimpse(dados)

## Rows: 1,000,000  
## Columns: 24  
## $ State.Code <int> 1, 56, 48, 47, 40, 4, 45, 20, 6, 42, 35, 39, 22...  
## $ County.Code <int> 113, 39, 231, 187, 27, 13, 73, 133, 11, 99, 45,...  
## $ Site.Num <int> 2, 8, 1006, 106, 49, 2005, 1, 3, 1002, 301, 18,...  
## $ Parameter.Code <int> 44201, 44201, 44201, 44201, 44201, 44201, 44201...  
## $ POC <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,...  
## $ Latitude <dbl> 32.46797, 43.67083, 33.15309, 35.95153, 35.3201...  
## $ Longitude <dbl> 85.08384, 110.59947, 96.11557, 87.13701, 97.484...  
## $ Datum <chr> "WGS84", "NAD83", "NAD83", "NAD83", "WGS84", "N...  
## $ Parameter.Name <chr> "Ozone", "Ozone", "Ozone", "Ozone", "Ozone", "O...  
## $ Date.Local <chr> "2014-10-22", "2014-08-19", "2014-01-15", "2014...  
## $ Time.Local <chr> "02:00", "19:00", "09:00", "18:00", "04:00", "1...  
## $ Date.GMT <chr> "2014-10-22", "2014-08-20", "2014-01-15", "2014...  
## $ Time.GMT <chr> "08:00", "02:00", "15:00", "23:00", "10:00", "2...  
## $ Sample.Measurement <dbl> 0.018, 0.039, 0.027, 0.045, 0.007, 0.043, 0.024...  
## $ Units.of.Measure <chr> "Parts per million", "Parts per million", "Part...  
## $ MDL <dbl> 0.005, 0.005, 0.005, 0.005, 0.005, 0.005, 0.005...  
## $ Uncertainty <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ Qualifier <chr> "", "", "", "", "", "", "", "", "", "", "", "",...  
## $ Method.Type <chr> "FEM", "FEM", "FEM", "FEM", "FEM", "FEM", "FEM"...  
## $ Method.Code <int> 47, 87, 87, 47, 47, 87, 47, 87, 87, 87, 47, 47,...  
## $ Method.Name <chr> "INSTRUMENTAL - ULTRA VIOLET", "INSTRUMENTAL - ...  
## $ State.Name <chr> "Alabama", "Wyoming", "Texas", "Tennessee", "Ok...  
## $ County.Name <chr> "Russell", "Teton", "Hunt", "Williamson", "Clev...  
## $ Date.of.Last.Change <chr> "2015-02-02", "2014-10-08", "2014-04-18", "2014...

Podemos utilizar a variável Time.Local para ver quais medidas de tempo são registradas.