Análise exploratória de dados

[Letícia Raposo](https://leticiaraposo.netlify.app/author/leticia-raposo/)

29-08-2020

Fonte: <https://leticiaraposo.netlify.app/post/analise-exploratoria/>

Banco de madeira em um poste

Descrição gerada automaticamente com confiança médiaCréditos da Imagem: [Canva](https://www.canva.com/)

**Você sabia que o R possui pacotes que realizam uma análise exploratória dos dados de maneira automática? Ao rodar as funções especializadas para isso, ele gera relatórios em HTML ou PDF que auxiliam a visualizar os tipos de variáveis, a existência de dados ausentes, dentre outras características. Aproveite o script e faça a sua própria análise!**

📌 Pacotes

As ferramentas automatizadas de análise exploratória de dados visam tornar a fase de compreensão dos dados mais rápida e fácil. Ela pode ser dividida em tarefas como: descrição de um conjunto de dados, exploração de dados e verificação da qualidade dos dados. Os pacotes aqui apresentados fornecem funções para lidar com esses desafios. Alguns deles também se preocupam com a fase de preparação de dados, que precede e dá suporte à fase de construção de modelos.

Veja aqui 👇 os pacotes disponíveis que realizam análise exploratória de dados de forma automática:

👉️ [dataMaid](https://cran.r-project.org/web/packages/dataMaid/index.html" \t "_blank): possui duas funções centrais: a função de verificação, que realiza avaliações de consistência e validade dos dados, e de resumo, que sumariza cada coluna do banco de dados. O pacote também cria automaticamente um relatório em formato PDF, DOCX ou HTML. O relatório contém um resumo completo do conjunto de dados: variáveis e seus tipos, número de valores ausentes, e resumos univariados na forma de estatísticas descritivas, histogramas/gráficos de barra e uma indicação de possíveis problemas.

👉️ [xray](https://cran.r-project.org/web/packages/xray/index.html" \t "_blank): possui três funções para a análise de dados: (i) anomalies: analisa todas as colunas do banco de dados em busca de anomalias, sejam elas NAs, zeros, infinitos, etc, e avisa se detectar variáveis com pelo menos 80% das linhas com essas anomalias. Ele também avisa quando todas as linhas têm o mesmo valor; (ii) distributions: tenta analisar a distribuição das variáveis, para que você possa entender como cada variável está estatisticamente estruturada. Ele também retorna uma tabela de percentis de variáveis numéricas como resultado; (iii) timebased: também investiga as distribuições, mas mostra a mudança ao longo do tempo.

👉️ [visdat](https://cran.r-project.org/web/packages/visdat/index.html" \t "_blank): formado por seis funções que ajudam a visualizar: (i) vis\_dat: apresenta o tipo das variáveis e se há dados ausentes; (ii) vis\_guess: apresenta o tipo de cada valor em cada coluna; (iii) vis\_miss: apresenta clusters de valores ausentes; (iv) vis\_compare: avalia diferenças entre dois conjuntos de dados; (v) vis\_expect: onde determinadas condições são satisfeitas nos dados; (vi) vis\_cor: gera uma matriz de correlação entre as variáveis numéricas.

👉️ [dlookr](https://cran.r-project.org/web/packages/dlookr/index.html" \t "_blank): fornece ferramentas para 3 tipos de análise: (i) diagnóstico dos dados, incluindo correção, detecção de valores ausentes e *outliers*; (ii) análise exploratória de dados; (iii) e engenharia de recursos, como imputação, dicotomização e transformação de variáveis contínuas. Ele também gera automaticamente um relatório em PDF ou HTML para todas essas análises.

👉️ [DataExplorer](https://cran.r-project.org/web/packages/DataExplorer/index.html" \t "_blank): fornece funções para: (i) introduce e plot\_intro: resumo do conjunto de dados completo: dimensões, tipos de variáveis, valores ausentes etc; (ii) plot\_missing e profile\_missing: visualização dos valores ausentes; (iii) plot\_histogram e plot\_bar: gera gráficos representando as distribuições de variáveis; (iv) plot\_qq: gera gráficos Q-Q; (v) plot\_correlation: gera matrizes de correlação; (vi) plot\_prcomp: permite visualizar os resultados de uma PCA plotando a porcentagem de variância explicada e correlações de cada variável original para cada componente principal; (vii) plot\_scatterplot e plot\_boxplot: traça relações entre a variável resposta e preditores; (viii) set\_missing: substitui os valores ausentes por uma constante; (ix) group\_category: agrupa categorias esparsas; (x) dummify e drop\_columns: cria variáveis dummy e descarta variáveis existentes.

Para conhecer outro pacotes, acesse [este link](https://arxiv.org/pdf/1904.02101.pdf).

Para ver um exemplo com os pacotes dlookr e dataMaid, acesse: , [Script](https://drive.google.com/file/d/1_eiQ-RTHV-RifFIeSWP21Tg2hXF-h24z/view?usp=sharing), [Dados](https://drive.google.com/file/d/1ZzCh-SGYvjbujiM-3oeGB7oarAFiSRSh/view?usp=sharing)

**#***#%######################################################%##*

**#** *#*

**#***### ANÁLISE EXPLORATÓRIA DOS DADOS ####*

**#***### PACOTES PARA GERAR RELATÓRIOS AUTOMÁTICOS ####*

**#** *#*

**#***#%######################################################%##*

**#** Banco de dados usado como exemplo

data(iris)

**#***#%######################################################%##*

**#** *#*

**#***### PACOTE DATAMAID ####*

**#** *#*

**#***#%######################################################%##*

**#** install.packages("dataMaid")

library(dataMaid)

makeDataReport(iris)

**#***#%######################################################%##*

**#** *#*

**#***### PACOTE XRAY ####*

**#** *#*

**#***#%######################################################%##*

**#** install.packages("xray")

library(xray)

anomalies(iris)

distributions(iris)

**#***#%######################################################%##*

**#** *#*

**#***### PACOTE VISDAT ####*

**#** *#*

**#***#%######################################################%##*

**#** install.packages("visdat")

library(visdat)

vis\_dat(iris)

vis\_guess(iris)

vis\_miss(iris)

vis\_cor(iris[,1:4])

**#***#%######################################################%##*

**#** *#*

**#***### PACOTE DLOOKR ####*

**#** *#*

**#***#%######################################################%##*

**#** install.packages("dlookr")

library(dlookr)

eda\_report(iris, "Species", output\_format = "html")

**#***#%######################################################%##*

**#** *#*

**#***### PACOTE DATAEXPLORER ####*

**#** *#*

**#***#%######################################################%##*

**#** install.packages("DataExplorer")

library(DataExplorer)

introduce(iris)

plot\_intro(iris)

plot\_missing(iris)

profile\_missing(iris)

plot\_histogram(iris)

plot\_bar(iris)

plot\_qq(iris)

plot\_correlation(iris)

plot\_prcomp(iris)

plot\_scatterplot(iris, by = "Species")

plot\_boxplot(iris, by = "Species")

create\_report(iris)