SmartEDA: Um pacote R para automatizar a Análise Exploratória de Dados

Sayan Putatunda; Dayananda Ubrangala; Kiran Rama; Ravi Kondapalli

04/01/2023

Introdução: Análise Exploratória de Dados

Hoje em dia, vemos aplicações de Data Science em quase todos os lugares.
Alguns dosaspectos mais bem destacados da ciência de dados são as várias técnicas estatísticas e de aprendizado de máquina aplicadas na resolução de problemas.
No entanto, qualquer atividade de ciência de dados começa com uma **Análise Exploratória de Dados** (EDA - *Exploratory Data Analysis*).
O termo "Análise Exploratória de Dados" foi criado pelo matemático e estatístico americano, John Tukey (SANDE, 2001). A EDA pode ser definida como *a arte e a ciência de realizar uma investigação inicial sobre os dados* por meio de técnicas estatísticas e de visualizações que possam trazer à tona os aspectos importantes nos dados e que podem ser utilizados posteriormente no processo de análise (Tukey, 1977).
Na literatura estatística existem muitos estudos sobre EDA. Alguns dos primeiros trabalhos realizados em Análise Exploratória de Dados (AED), incluindo susa definição e o estabelecimento das técnicas básicas em AED foram mostrados em Tukey (1977).
No entanto, muitos pesquisadores formularam diferentes definições de AED ao longo dos anos.
Chon Ho (2010) introduziu a AED no contexto de mineração de dados e da reamostragem com foco no reconhecimento de padrões, detecção de agrupamentos (cluster) e na seleção de variáveis.
Ao longo dos anos, a AED tem sido utilizada em diversos tipos de aplicações e em diferentes domínios, tais como: pesquisa em geociências (Ma et al., 2017), avaliações baseadas em jogos (DiCerbo et al., 2015), grupos de estudos clínicos (Konopka et al., 2018), dentre outros.
A AED pode ser categorizada em **técnicas estatísticas descritivas** e **técnicas gráficas**. A primeira categoria abrange várias técnicas estatísticas univariadas e multivariadas, enquanto a segunda categoria compreende várias técnicas de visualização. Ambas as técnicas são usadas para explorar e entender os padrões dos dados, compreender as relações existentes entre as variáveis e, o mais importante, gerar **insights** orientados por dados que podem ser usados pelas partes interessadas (stakeholders) nos negócios. No entanto, a AED requer muito esforço manual e também uma quantidade substancial de esforço de codificação em um ambiente de programação como o R (R Core Team, 2017). Existe uma grande necessidade de automação do processo EDA, e isso nos motivou a desenvolver o pacote **SmartEDA** e elaborar este artigo.

Principal Funcionalidade

O pacote SmartEDA seleciona automaticamente as variáveis e executa as funções estatisticas descritivas afins.
Além disso, também analisa o valor da informação, o peso da evidência, tabelas personalizadas, estatísticas resumidas e executa técnicas gráficas tanto para dados numéricos quanto categóricos.
Algumas das vantagens mais importantes do pacote SmartEDA são que ele pode ajudar na aplicação de ponta a ponta do processo EDA sem ter que lembrar os diferentes nomes de pacotes em R, escrever longos scripts em R, e nenhum esforço manual é necessário para preparar o relatório EDA e, finalmente, automaticamente categorizar as variáveis no tipo de dados correto (ou seja, caractere, numérico, Fator e mais) com base nos

dados de entrada.
Assim, os principais benefícios do SmartEDA estão na economia de tempo de desenvolvimento, menor percentual de erros e reprodutibilidade. Além disso, o pacote SmartEDA possui opções personalizadas para os dados. Pacote de tabelas como:
(1) Gera estatísticas de resumo apropriadas dependendo do tipo de dados;
(2) Remodelagem de dados usando data.table.dcast();
(3) Filtra linhas/casos em que as condições são verdadeiras. Opções para aplicar filtros em nível de variável ou conjunto de dados completo como subconjunto de base; e
(4) Opções para calcular medidas de tendência central (como Média, Mediana, Moda, etc.), medidas de variância/dispersão (como Desvio Padrão, Variância, etc.), Número de observações, Proporções, Quantis, IQR, Porcentagem de Ações (PS) para dados numéricos.
A Figura 1 resume as várias funcionalidades do pacote SmartEDA.

[./figuras/Figura 1 Funcionalidades SmarEDA.jpg]
Figura 1: As diversas funcionalidades do SmartEDA.

Ilustração

Aplicamos o SmartEDA para gerar insights sobre as vendas de cadeirinhas infantis em diferentes localidades. Usaremos os dados "Carseats" disponíveis no pacote ISLR (James, Witten, Hastie, & Tibshirani, 2017) que contém 11 variáveis, tais como: vendas unitárias em cada local (Sales); preçocobrado pelos concorrentes (CompPrice); nível de renda da comunidade (Income); tamanho da população na região (population); orçamento de publicidade (Advertising), preço que a empresa cobra pelas cadeirinhas em cada local (Price); qualidade da localização das prateleiras (ShelveLoc); média idade da população local (Age), nível educacional em cada local (Education), indicador de localização urbana/rural Utilizaremos o SmartEDA para entender as dimensões do conjunto de dados, nomes de variáveis, resumo geral ausente e tipos de dados de cada variável.

Inicialmente, iremos carregar as bibliotecas SmartEDA e ISLR.

```
#if(!SmartEDA %in% installed.packages()) install.packages("SmartEDA")
library(SmartEDA)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
#if(!ISLR %in% installed.packages()) install.packages("ISLR")
library("ISLR")
```

Agora usaremos o SmartEDA para compreender: as dimensões do conjunto de dados, os nomes e tipos de dados de cada variável e observar um resumo geral de dados ausentes (missing data) .

```
Carseats <- ISLR::Carseats
ExpData(data=Carseats,type=1)
```

```
##                                      Descriptions     Value
## 1                              Sample size (nrow)       400
## 2                         No. of variables (ncol)        11
## 3                 No. of numeric/interger variables         8
## 4                           No. of factor variables         3
## 5                             No. of text variables         0
```

```
## 6                              No. of logical variables         0
## 7                           No. of identifier variables         0
## 8                                No. of date variables          0
## 9              No. of zero variance variables (uniform)         0
## 10                %. of variables having complete cases 100% (11)
## 11   %. of variables having >0% and <50% missing cases    0% (0)
## 12 %. of variables having >=50% and <90% missing cases    0% (0)
## 13          %. of variables having >=90% missing cases    0% (0)
```

```
# saída
```

Agora, vejamos o resumo das variáveis numéricas/inteiras, como Advertising, Age,CompPrice, Income, Population, Price e Sales.

```
ExpNumStat(Carseats,by="A",gp=NULL,Qnt=NULL,MesofShape=2, Outlier=FALSE,round=2,Nlim=10)
```

```
##         Vname Group  TN nNeg nZero nPos NegInf PosInf NA_Value Per_of_Missing
## 4 Advertising   All 400    0   144  256      0      0        0              0
## 7         Age   All 400    0     0  400      0      0        0              0
## 2   CompPrice   All 400    0     0  400      0      0        0              0
## 3      Income   All 400    0     0  400      0      0        0              0
## 5  Population   All 400    0     0  400      0      0        0              0
## 6       Price   All 400    0     0  400      0      0        0              0
## 1       Sales   All 400    0     1  399      0      0        0              0
##         sum min    max   mean median     SD   CV    IQR Skewness Kurtosis
## 4   2654.00   0  29.00   6.64   5.00   6.65 1.00  12.00     0.64    -0.55
## 7  21329.00  25  80.00  53.32  54.50  16.20 0.30  26.25    -0.08    -1.14
## 2  49990.00  77 175.00 124.97 125.00  15.33 0.12  20.00    -0.04     0.03
## 3  27463.00  21 120.00  68.66  69.00  27.99 0.41  48.25     0.05    -1.09
## 5 105936.00  10 509.00 264.84 272.00 147.38 0.56 259.50    -0.05    -1.20
## 6  46318.00  24 191.00 115.80 117.00  23.68 0.20  31.00    -0.12     0.43
## 1   2998.53   0  16.27   7.50   7.49   2.82 0.38   3.93     0.18    -0.10
```

```
#Saída- Resumo das variáveis numéricas dos dados Carseats
```

Vamos agora verificar o resumo das variáveis categóricas, ou seja, ShelveLoc, Urban e US.

```
ExpCTable(Carseats)
```

```
##     Variable  Valid Frequency Percent CumPercent
## 1  ShelveLoc    Bad        96   24.00      24.00
## 2  ShelveLoc   Good        85   21.25      45.25
## 3  ShelveLoc Medium       219   54.75     100.00
## 4  ShelveLoc  TOTAL       400      NA         NA
## 5      Urban     No       118   29.50      29.50
## 6      Urban    Yes       282   70.50     100.00
## 7      Urban  TOTAL       400      NA         NA
## 8         US     No       142   35.50      35.50
## 9         US    Yes       258   64.50     100.00
## 10        US  TOTAL       400      NA         NA
## 11 Education     10        48   12.00      12.00
## 12 Education     11        48   12.00      24.00
```

```
## 13 Education     12        49   12.25      36.25
## 14 Education     13        43   10.75      47.00
## 15 Education     14        40   10.00      57.00
## 16 Education     15        36    9.00      66.00
## 17 Education     16        47   11.75      77.75
## 18 Education     17        49   12.25      90.00
## 19 Education     18        40   10.00     100.00
## 20 Education  TOTAL       400      NA         NA
```

```
#Output- Resumo das variáveis categóricas dos dados Carseats
```

Podemos visualizar as diferentes representações gráficas usando o pacote SmartEDA quando aplicado no conjunto de dados "Carseats".
A Figura 2 mostra as diferentes visualizações gráficas, ou seja, gráfico de dispersão, gráfico de densidade, gráfico de barras, gráfico de caixa, gráfico de normalidade e gráfico de coordenadas.

Figura 2: Representações gráficas dos dados Carseats usando SmartEDA

Comparação com outros pacotes R

A Figura 3 compara o pacote SmartEDA (Ubrangala, Rama, Kondapalli, & Putatunda, 2018) com outros pacotes semelhantes disponíveis no CRAN para análise exploratória de dados viz. dlookr (Ryu, 2018), DataExplorer (Cui, 2018), Hmisc (Harrell et al., 2018), exploreR (Coates, 2016), Tutor (Nair, 2018) e summarytools (Comtois, 2018).
A métrica para avaliação é a disponibilidade de vários recursos desejados para realizar uma análise exploratória de dados, tais como:
(a) Descrever informações básicas para dados de entrada;
(b) Função a fornecer;
(c) estatísticas resumidas para todas as variáveis numéricas;
(d) Função para fornecer gráficos para todas as variáveis numéricas;
(e) Função para fornecer estatísticas de resumo para todos os caracteres ou variáveis categóricas;
(f) Função para fornecer gráficos para todos os caracteres ou variáveis categóricas;
(g) Estatísticas de resumo personalizadas - extensão para dados. Pacote data.table;
(h) Gráficos de normalidade/coordenadas;
(i) Binarização/binning de recursos;
(j) Padronizar/faltar imputação/diagnosticar outliers; e
(k) Relatório HTML usando rmarkdown/Shiny.

[./figuras/Figura 3 - Tabela comparativa de pacotes EDA.jpg]
Figura 3: Comparação do SmartEDA com os pacotes R disponíveis.

Conclusão

A contribuição deste trabalho está no desenvolvimento de um novo pacote em R ou seja, SmartEDA para Análise Exploratória de Dados automatizada.
O pacote SmartEDA ajuda na implementação da Análise Exploratória de Dados completa apenas executando a função em vez de escrever um código R demorado. Os usuários do SmartEDA podem automatizar todo o processo de EDA em qualquer conjunto de dados com funções fáceis de implementar e exportar relatórios de EDA que seguem as melhores práticas da indústria e da academia.
O SmartEDA pode fornecer estatísticas resumidas junto com gráficos para variáveis numéricas e categóricas. Ele também fornece uma extensão para o pacote data.table que nenhum dos outros pacotes disponíveis no CRAN oferece.
No geral, os principais benefícios do SmartEDA estão na economia de tempo de desenvolvimento, menor

porcentagem de erros e reprodutibilidade.
Em setembro de 2019, o pacote SmartEDA tinha mais de 6.000 downloads, o que indica sua aceitabilidade e maturidade nas estatísticas e na comunidade de aprendizado de máquina.
Podemos ver na Figura 3 que a versão atual do SmartEDA possui quase todas as características desejadas mencionadas acima, exceto os pontos (h) e (i), ou seja, gráficos de normalidade e binning de recursos, respectivamente. Esses dois recursos seriam incorporados na próxima versão e estamos trabalhando nisso.
No entanto, a funcionalidade exclusiva e mais forte fornecida pelo SmartEDA é o ponto (f), ou seja, uma extensão para o pacote data.table que nenhum dos outros pacotes oferece. Assim, o SmartEDA agrega valor devido à importância e popularidade do data.table entre os usuários do R para analisar grandes conjuntos de dados.
A Figura 3 mostra que o SmartEDA é melhor do que quase todos os outros pacotes disponíveis no CRAN.
O concorrente mais próximo do SmartEDA parece ser o pacote DataExplorer, mas este não possui os recursos de dataviz (b) e (f). Função para fornecer estatísticas de resumo para todas as variáveis numéricas e extensão para dados e para o acote data.table, respectivamente.
Além disso, outra característica distintiva que o SmartEDA possui, mas nenhum dos outros pacotes semelhantes possuem, é a capacidade de exportar todos os gráficos em um pdf.

Disponibilidade

O software é distribuído sob uma LICENÇA de arquivo MIT + (Repositório: CRAN) e está disponível em https://github.com/daya6489/SmartEDA.

Reconhecimentos

Queremos agradecer à VMware e à liderança do Enterprise and Data Analytics (EDA) por nos fornecer a infraestrutura e o suporte necessários para este trabalho. Somos gratos à comunidade R por sua aceitação e feedback para melhorar ainda mais nosso pacote.

Referências Bibliográficas

Chon Ho, Y. (2010). Exploratory data analysis in the context of data mining and resampling. International Journal of Psychological Research, 3(1), 9–22. doi:https://doi.org/10.21500/ 20112084.819
Coates, M. (2016). exploreR: Tools for Quickly Exploring Data. Retrieved from {https://CRAN.R-project.org/package=exploreR}
Comtois, D. (2018). summarytools: Tools to Quickly and Neatly Summarize Data. Retrieved from {https://CRAN.R-project.org/package=summarytools}
Cui, B. (2018). DataExplorer: Data Explorer. Retrieved from {https://CRAN.Rproject. org/package=DataExplorer}
DiCerbo et al. (2015). Serious Games Analytics. Advances in Game-Based Learning. In C. Loh, Y. Sheng, & D. Ifenthaler (Eds.),. Cham: Springer. doi:10.1007/978-3-319-05834-4 Harrell et al. (2018). Hmisc: Harrell Miscellaneous. Retrieved from {https://CRAN.Rproject. org/package=Hmisc}
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). ISLR: Data for an Introduction to Statistical Learning with Applications in R. doi:https://doi.org/10.1007/ 978-1-4614-7138-7_1
Konopka et al. (2018). Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. PLoS ONE, 13(8). doi:https://doi.org/10. 1371/journal.pone.0201950
Ma, X., Hummer, D., Golden, J. J., Fox, P. A., Hazen, R. M., Morrison, S. M., Downs, R. T., et al. (2017). Using Visual Exploratory Data Analysis to Facilitate Collaboration and Hypothesis Generation in Cross-Disciplinary Research. International Journal of Geo-Information, 6(368), 1–11. doi:https://doi.org/10.3390/ijgi6110368
Nair, A. (2018). RtutoR: Shiny Apps for Plotting and Exploratory Analysis. Retrieved from {https://CRAN.R-project.org/package=RtutoR}
R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ryu, C. (2018). dlookr: Tools for Data Diagnosis, Exploration, Transformation. Retrieved from {https://CRAN.R-project.org/package=dlookr}
Sande, Gordon (July 2001). "Obituary: John Wilder Tukey". Physics Today. 54 (7): 80–81. doi:10.1063/1.1397408.
Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.
Ubrangala, D., Rama, K., Kondapalli, R. P., & Putatunda, S. (2018). SmartEDA: Summarize and Explore the Data. Retrieved from {https://CRAN.R-project.org/package=SmartEDA}