Departamento de Ciência Política

Disciplina: ELT – extração, leitura e tratamento de dados com R – 2021

Professor: Hugo Medeiros

Objetivo Geral: Usar a linguagem de programação R para realizar processos de extração, leitura e tratamentos de dados de múltiplas fontes e formatos.

Objetivos Específicos

- Configurar o RStudio para utilização integrada e produtiva

- Ler dados de diferentes formatos e tamanhos com o R

- Transformar e higienizar dados com o R

**1. Tópicos e Calendário**

|  |  |  |
| --- | --- | --- |
| **Tópico** | **Data** | **C/H** |
| Introdução à disciplina – Bibliografia – Avaliação e Métodos.  Introdução ao R: o que é? Como baixar e configurar. Objetos. Pacotes.  Introdução ao RStudio: funcionalidades, customização, projetos e integração com github | 05/04 | 4 |
| Objetos e tipos de dados no R: fatores, vetores, matrizes, listas e dataframes  Programação em R: escrevendo funções, condicionais e laços | 12/04 | 8 |
| Exercícios | 19/04 | 12 |
| Extração, Tratamento e Leitura x Extração, Leitura e Tratamento  Data warehouse e Data Lake | 26/04 | 16 |
| Extração e leitura de dados no R: tabulares, csv, excel, bancos de dados, html etc.  Salvando e exportando objetos no R | 03/05 | 20 |
| Small, Large e Big Data  Lidando com Large Data no R | 10/05 | 24 |
| Exercícios | 17/05 | 28 |
| Data Wrangling (Transformação) e Data Cleaning (Higienização) de Dados  Introdução ao Tidyverse | 24/05 | 32 |
| Estruturas, manipulações e transformações de dados no R | 31/05 | 36 |
| Estruturas, manipulações e transformações de dados no R | 07/06 | 40 |
| Trabalhando com textos no R | 14/06 | 44 |
| Trabalhando com datas e séries temporais no R | 21/06 | 48 |
| Exercícios | 28/06 | 52 |
| Avaliação | 05/07 | 56 |
| Seminários | 12/07 | 60 |

\* Tópicos e datas previstas podem ser alterados no futuro.

**2. Bibliografia**

Sites Úteis

https://www.r-bloggers.com/

https://blog.rstudio.com/

https://www.linkedin.com/company/r-icymi/posts/?feedView=all

https://rweekly.org/2021-W03.html#get-

https://www.statmethods.net/

http://www.sthda.com/english/

https://style.tidyverse.org/

Cheat Sheets do RStudio

https://www.rstudio.com/resources/cheatsheets/

Referências

BARUFFA, Oscar. Big Book of R. 2021. Disponível em: https://www.bigbookofr.com/

BIGDATA DOC. Fuzzy String Matching – a survival skill to tackle unstructured information. R-bloggers. Disponível em: https://www.r-bloggers.com/2015/02/fuzzy-string-matching-a-survival-skill-to-tackle-unstructured-information/

DE JONGE, Edwin; VAN DER LOO, Mark. An introduction to data cleaning with R. Statistics Netherlands, 2013. Disponível em: https://cran.r-project.org/doc/contrib/de\_Jonge+van\_der\_Loo-Introduction\_to\_data\_cleaning\_with\_R.pdf.

GOEL, Ajay Kumar. ETL vs ELT: Must-Know Benefits and Differences. https://codestoresolutions.com/etl-vs-elt-benefits-differences/.

GROLEMUND, Garrett. Hands-On Programming with R. Editora O'Reilly Media, 2017. Disponível em: https://rstudio-education.github.io/hopr/

KABACOFF, Robert I. R in Action: Data analysis and graphics with R. Manning Publication Co, 2011. Disponível em: http://www.cs.uni.edu/~jacobson/4772/week11/R\_in\_Action.pdf

KANG, Hyun. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May; 64(5): 402–406. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/

KIMBALL, Ralph; ROSS, Margy. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Indianapolis: Wiley, 2013.

PHUOC LE, Khac. String Manipulation in R with stringr. 2018. Disponível em: https://rpubs.com/iPhuoc/stringr\_manipulation

MAYER, Fernando et al. Estatística Computacional com R. Disponível em: http://cursos.leg.ufpr.br/ecr/index.html

MOREIRA, Davi. Texto como Dado para Ciências Sociais: guia prático com aplicações. 2021. Disponível: https://bookdown.org/davi\_moreira/txt4cs/

MÜLLER, Heiko; FREYTAG, Johann-Christoph. Problems, methods, and challenges in comprehensive data cleansing. 2002. Disponível em: https://www.researchgate.net/publication/228929938\_Problems\_methods\_and\_challenges\_in\_comprehensive\_data\_cleansing.

PENG, Roger D; KROSS, Sean; ANDERSON, Brooke. Mastering Software Development in R. 2002. Disponível em: https://bookdown.org/rdpeng/RProgDA/.

PRADEEP, Sundar; MOV, Philip. Handling large data sets in R. Disponível em: https://rpubs.com/msundar/large\_data\_analysis.

RSTUDIO. Data Wrangling with dplyr and tidyr: Cheat Sheet. Disponível em: https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf.

R Software Handbook. Disponível em: https://bookdown.org/aschmi11/RESMHandbook/.

SANCHEZ, Gaston. Handling and Processing Strings in R. 2014. Disponível em: https://gotellilab.github.io/Bio381/Scripts/Feb07/HandlingAndProcessingStringsInR.pdf.

SILGE, Julia; ROBINSON, David. Text Mining with R: A Tidy Approach. Editora O'Reilly Media, 2017.

SMALLCOMBE, Mark. ETL vs ELT: 5 Critical Differences. 2020. Disponível em: https://www.xplenty.com/blog/etl-vs-elt.

TEETOR, Paul. R Cookbook 2e: Proven Recipes for Data Analysis, Statistics, and Graphics. O’Reilly, 2019. Disponível em: https://rc2e.com/.

TORRES-REYNA, Oscar. Fuzzy merge in R. Pricenton University: 2015. Disponível em: https://www.princeton.edu/~otorres/FuzzyMergeR101.pdf

VAN DEN BROECK, Jan. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS, October 2005 | Volume 2 | Issue 10 | e267. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198040/pdf/pmed.0020267.pdf .

WICKHAM, Hadley; GROLEMUND, Garrett. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Editora O'Reilly Media, 2016.

\* Tópicos e datas previstas podem ser alterados no futuro.

**3. Avaliação da disciplina**

* **Exercícios** individuais sobre blocos de conteúdo do curso.
* **Prova** individual, abordando todos os conteúdos do curso.
* **Seminários** em grupo, sobre aspectos teóricos e tópicos avançados das técnicas aprendidas no curso.

A **nota final** será a *média ponderada* das atividades assinaladas, de acordo com os pesos abaixo:

|  |  |
| --- | --- |
| **Atividades** | **Peso** |
| Exercícios | 40% |
| Prova | 40% |
| Seminários | 20% |