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**

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