



Data Science aplicada: do dado bruto até o insight analítico

Professor

Leonardo Monasterio

Carga Horária

26 horas

Datas dos Encontros

Remotos:

16 de Novembro – 19h00 às 21h30

28 de Novembro – 19h00 às 21h30

Presenciais:

22 de Novembro – 17h00 às 20h30

23 de Novembro – 08h30 às 16h30

06 de Dezembro – 17h00 às 20h30

07 de Dezembro – 08h30 às 16h30

Ementa do Curso

Introdução à Ciência de Dados com foco em aplicações práticas. Utilização do R moderno (tidyverse) para análise de dados, abrangendo desde a limpeza até a visualização avançada com ggplot2 e construção de modelos econométricos. Implementação de técnicas de análise exploratória, automação de fluxos de trabalho e práticas de pesquisa replicável. Uso de ferramentas como RStudio, GitHub, Posit Cloud e Large Language Models (LLMs) para otimizar o processo de análise de dados..

Objetivos do Curso

Ao final do curso, os alunos deverão estar aptos a:

1. Implementar métodos de limpeza e preparação de dados para desafios reais.
2. Transformar dados brutos em informações úteis usando o tidyverse.
3. Aplicar técnicas de análise exploratória para identificar padrões em conjuntos de dados complexos.
4. Criar visualizações eficazes com ggplot2 para comunicar resultados complexos.
5. Construir e interpretar modelos econométricos aplicáveis a problemas reais.
6. Implementar processos de trabalho automatizados para aumentar a eficiência das análises.
7. Utilizar o GitHub para controle de versão e colaboração em projetos de análise de dados.
8. Integrar Large Language Models (LLMs) no desenvolvimento de código R.
9. Desenvolver projetos de pesquisa replicáveis, seguindo práticas modernas de ciência de dados.
10. Apresentar resultados de análises de forma clara e compreensível a diversos públicos.

Metodologia de Ensino

A metodologia consiste em aulas expositivas e práticas, com foco na aplicação imediata dos conceitos aprendidos em problemas reais. As aulas incluirão:

- Demonstrações práticas de análise de dados utilizando R e tidyverse.
- Exercícios hands-on em sala de aula, simulando cenários reais de análise de dados.
- Discussões sobre casos de aplicação de ciência de dados em diversos setores.

- Workshops de utilização de GitHub, Posit Cloud e LLMs para otimizar o fluxo de trabalho.
- Sessões de revisão de código para promover boas práticas de programação e colaboração.
- Projetos práticos que integram múltiplas técnicas e ferramentas aprendidas no curso

Forma de Avaliação

A nota mínima para aprovação é 6,0 (seis) e a avaliação será composta por diversas atividades realizadas em sala de aula e em casa.

Conteúdo Programático

1 - Mise en Place: Conheça a sua cozinha e suas ferramentas

- R e RStudio
- tidyverse
- GitHub e Posit Cloud
- LLM, o seu Sous-chef hábil porém atrapalhado

2- Colhendo, Limpando e Cortando os Dados

- Limpando e descascando dados com dplyr
- Removendo as partes estragadas: tratamento de NAs e outliers
- Venenos, contaminações e outras armadilhas nos dados

3- Degustação de Dados: a análise exploratória

4- A Arte da Apresentação com ggplot2

5- O prato principal: a modelagem econométrica

6- Automatizando a sua cozinha e a importância das receitas documentadas

7 - O Banquete: fluxo para a elaboração de um projeto completo em ciência de dados

Plano das aulas

Aula 1 (Remota) - 16 de Novembro (19:00-21:30)

"Mise en Place: Conheça a sua cozinha e suas ferramentas"

- 19:00-19:30: Apresentação do curso e do professor
- 19:30-20:00: Apresentação dos alunos e suas expectativas
- 20:00-20:15: Intervalo para café
- 20:15-20:45: Introdução ao Posit Cloud e tidyverse
- 20:45-21:30: GitHub e LLM (seu Sous-chef hábil porém atrapalhado)

Aula 2 (Presencial) - 22 de Novembro (17:00-20:30)

"Colhendo, Limpando e Cortando os Dados"

- 17:00-18:15: Fundamentos do dplyr
- 18:15-18:30: Intervalo para café
- 18:30-19:45: Transformação de dados com tidyverse
- 19:45-20:30: Tratamento de NAs e outliers

Aula 3 (Presencial) - 23 de Novembro (08:30-16:30)

"Degustação de Dados: a análise exploratória"

- 08:30-10:00: Limpeza e preparação de dados
- 10:00-10:15: Intervalo para café
- 10:15-12:00: Análise exploratória inicial
- 12:00-13:00: Almoço
- 13:00-14:30: Práticas hands-on com dados reais
- 14:30-14:45: Intervalo para café
- 14:45-16:30: Continuação das práticas e discussão

Aula 4 (Remota) - 28 de Novembro (19:00-21:30)

"A Arte da Apresentação com ggplot2"

- 19:00-19:45: Fundamentos do ggplot2
- 19:45-20:00: Intervalo para café
- 20:00-21:00: Princípios de visualização de dados
- 21:00-21:30: Boas práticas em comunicação visual

Aula 5 (Presencial) - 06 de Dezembro (17:00-20:30)

"O prato principal: a modelagem econométrica"

- 17:00-18:15: Introdução à modelagem econométrica
- 18:15-18:30: Intervalo para café
- 18:30-19:45: Integração entre modelagem e tidyverse
- 19:45-20:30: Interpretação e diagnóstico de modelos

Aula 6 (Presencial) - 07 de Dezembro (08:30-16:30)

"O Banquete: fluxo para a elaboração de um projeto completo"

- 08:30-10:00: Automatização de fluxos de trabalho
- 10:00-10:15: Intervalo para café
- 10:15-12:00: Documentação e reprodutibilidade
- 12:00-13:00: Almoço
- 13:00-14:30: Desenvolvimento dos projetos finais
- 14:30-14:45: Intervalo para café
- 14:45-16:30: Apresentação dos projetos e encerramento

Míni Currículo do Docente

Leonardo Monasterio

Possui graduação em Ciências Econômicas na UFRJ (1992), mestrado em Economia pela UFRGS (1995) e doutorado em Desenvolvimento Econômico pela UFPR (2002), com estágio sanduíche na University of Cambridge. Foi professor visitante na University of London (2006-2007) e pesquisador visitante no Departamento de Economia da University of California -Los Angeles (2015-2016). É pesquisador de carreira do Ipea, foi coordenador de avaliação de políticas públicas da instituição e coordenador-geral de Ciência de Dados da ENAP.



Bibliografia (lista incompleta)

Bryan, J. (2024). Happy Git and GitHub for the useR. <https://happygitwithr.com/>

Gentzkow, Matthew, and Jesse M. Shapiro. 2014. "Code and Data for the Social Sciences: A Practitioner's Guide." *University of Chicago Mimeo*. Last Updated January.

Kuhn, M., & Silge, J. (2024). Tidy Modeling with R. <https://www.tmwr.org/>

Taylor, Linnet, Ralph Schroeder, and Eric Meyer. 2014. "Emerging Practices and Perspectives on Big Data Analysis in Economics: Bigger and Better or More of the Same?" *Big Data & Society* 1 (2):2053951714536877. <https://doi.org/10.1177/2053951714536877>.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Graphics Press.

Wickham, H., & Grolemund, G. (2024). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.