

# Política e Ciência de Dados

## Informações de Contato

**Thiago Moreira** (DCP-UFMG)

✉ thiagomoreira@ufmg.br

📅 Segundas e Quartas, 14h às 17h

📍 A definir



## Visão Geral e Objetivos

A ciência de dados está mudando rapidamente a ciência política e a forma como as políticas públicas são implementadas. Neste curso, as alunas e os alunos aprenderão ferramentas fundamentais de programação, com aplicações relacionadas a questões políticas e sociais. Logo, serão ensinadas técnicas utilizadas para gerenciamento e análise de dados. Ao final do curso, as estudantes serão capazes de i) construir bancos de dados, ii) resumir e visualizar informações e iii) usar regressões para analisar resultados. Os conteúdos serão expostos no ambiente do *software* R.

## Pré-Requisitos

Não se espera o conhecimento prévio de programação ou de estatística básica. O curso será lecionado em etapas, que ajudarão a desenvolver habilidades de programação a partir do zero.

## Metodologia de Ensino

A disciplina será pautada por aulas expositivas e exercícios semanais, praticados em sala de aula. A leitura prévia dos textos indicados é opcional, mas altamente recomendada. Todos os encontros semanais contarão com notas de aula, colocadas na pasta

da disciplina, para facilitar o aprendizado de todas as pessoas presentes no curso.

## Avaliação

A avaliação da disciplina será baseada na aplicação das técnicas ensinadas em dois exercícios: i) no meio e ii) ao final do curso. A primeira avaliação será um desafio proposto por mim, ao passo que a segunda será pautada por propostas das próprias alunas, com base de dados e projetos escolhidos pelas discentes.

As datas de entrega dos trabalhos estão especificadas na seção **Cronograma das Aulas**. Recomendo que as alunas e os alunos façam os trabalhos em duplas. Para a escrita dos exercícios, sugiro que usem o LaTeX (para um aprendizado rápido, ver o [site](#)) ou o *Rmarkdown* e não o *Word*, já que o *software* da *Microsoft* não funciona bem com códigos, tabelas e gráficos.

As notas serão dispostas da seguinte maneira:

- Avaliação I (30%) da nota final.
- Avaliação II (70%) da nota final.

## Software

Neste curso usaremos o *software* R, a ferramenta mais popular de análise de dados para cientistas sociais atualmente. O *software* pode ser baixado gratuitamente [aqui](#). Recomendo também que as alunas e os alunos façam o *download* da interface gráfica Rstudio, que facilita a operacionalização do *software*. O download pode ser feito neste [site](#). Por fim, sugiro o uso do *Posit Cloud* para evitar problemas de compatibilidade com os pacotes. Para abrir uma conta gratuita no *Posit Cloud*, entrem neste [endereço](#).

## Expectativas

A expectativa é que as(os) alunas(os) consigam abrir e montar bancos de dados diversos, criar sumários estatísticos e boas visualizações e aprendem o básico para reportar modelos de regressão.

## Livros Recomendados

- Wickham,(2016). *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Houston: Springer.
- Wickham, H., & Grolemund, G. (2017). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc. (Há uma versão em português).
- Silge, J., Robinson, D., & Robinson, D. (2017). *Text mining with R: A tidy approach* (p. 194). Boston (MA): O'reilly.
- Wilke, C. O. (2019). *Fundamentals of Data Visualization: a primer on making informative and compelling figures*. O'Reilly Media.

## Cronograma das Aulas

\* *O cronograma pode ser modificado ao longo do curso.*

### Aula 1 (08/10) - Introdução ao R

- Apresentação do curso, introdução ao R e ao *Rmarkdown*. Uso de pacotes e aplicações básicas no R.

Leituras Recomendadas:

- Wickham, H., & Grolemund, G. (2017), caps. 2 e 3.

### Aula 2 (15/10) - Funções e Manipulação de Dados

- Organização de dados com *tidyverse* (I): selecionar, filtrar, renomear e modificar variáveis.

Leituras Recomendadas:

- Wickham, H., & Grolemund, G. (2017), caps. 2 e 3.

### **Aula 3 (22/10) - Sumários Estatísticos**

- Sumarizar dados: médias, medianas, desvio-padrão, frequências e tabelas.

Leituras Recomendadas:

- Wickham, H., & Grolemund, G. (2017), cap 3.
- Ver o básico do pacote [gt](#).
- Dica: outras possibilidades de pacote no site [rfortherestofus](#)

### **Aula 4 (29/10) - Combinar Bases de Dados**

- Junção de bases de dados diversas.

Leituras Recomendadas:

- Wickham & Grolemund (2017), cap. 10.

**Exercício I:** entrega até 05/11.

### **Aula 5 (06/11) - Visualização de Dados I**

- Visualização de dados com *ggplot* (I): histogramas, barras, linhas, caixa e dispersão.

Leituras Recomendadas:

- Wickham (2016), cap. 2.
- Wilke (2019), caps. 6, 7, 10 e 12.

Vídeo recomendado:

- Palestra de John Rauser: "[How humans see data?](#)"

### **Aula 7 (13/11) - Visualização de Dados II**

- Visualização de dados com *ggplot* (II): mapas e gráfico interativos.

Leituras Recomendadas:

- Wickham (2016), cap. 3.
- Wilke (2019), cap. 15.

## **Aula 8 (20/11) - Texto como Dado**

- Manipulação de palavras (*strings*) no R.

Leituras Recomendadas:

- Wickham (2016), cap. 11.
- Silge & Robinson (2017), cap. 1.
- Denny, M. & Spirling, A. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Anal.* 26, 168–189 (2018).

## **Aula 9 (27/11) - Modelos de Regressão no R**

- Testes estatísticos e modelos de regressão no R.

Leituras Recomendadas:

- Wickham & Grolemund (2017), cap. 18.
- Dica: pacotes [marginaleffects](#) e [sjPlot](#) podem te ajudar muito.

**Exercício II:** entrega até 15/12.