Introdução à Análise de Dados com R

(Programa de Nivelamento para Lego 1)

Carlos Antônio Costa Ribeiro Rogério Jerônimo Barbosa

Março de 2020

1 Ementa e Justificativa

A aplicação de métodos e técnicas de análise quantitativa depende não apenas do conhecimento sobre Matemática e Estatística, como também do domínio de um ferramental computacional. De outra forma, o escopo de nossas análises estaria majoritariamente restrito a pequenos bancos de dados (manuseáveis apenas com papel e caneta) e a modelos e indicadores relativamente simples de calcular... Assim, softwares de análise são instrumentos obrigatórios nas pesquisas quantitativas já há muitas décadas – uma vez que nossos os bancos de dados tipicamente contém dezenas de variáveis e milhares (ou milhões) de observações.

Neste curso, ofereceremos uma introdução ao software R, que é gratuito, de código aberto e que nos últimos anos se tornou a principal plataforma para Estatística Aplicada, aplicação de modelos de Inteligência Artificial, coleta automatizada de dados (web scraping), elaboração de gráficos profissionais (no crescente campo da Visualização de Dados) – enfim, para a realização das atividades e tarefas que se convencionou denominar Data Science. Nossa opção pelo R justifica-se, assim, não apenas por sua licença de livre acesso (uma razão política, que visa a democratização do acesso), como também por estar alinhado aos desenvolvimentos mais recentes do campo da Metodologia das Ciências Sociais e às formas de análise de dados também em outros campos do conhecimento.

O R, porém, não é apenas um software, mas também uma linguagem de programação. Isso significa que não se trata apenas de aprender e executar "comandos", como usualmente se costuma pensar acerca de outros softwares estatísticos. Trata-se da aquisição de uma gramática e forma de estruturação de pensamento (e expressão). Seu aprendizado, como o de qualquer outra linguagem, depende de prática e exposição.

A consequência disso é que, inicialmente, avançaremos devagar, para fixarmos os conteúdos básicos, que servirão de pilares para os mais avançados. Ainda assim cobriremos muitos assuntos interessantes, passando desde a análise descritiva de dados, manuseio de diversos bancos de dados simultaneamente (e.g. uma série histórica de pesquisas do IBGE, bancos de resultados eleitorais etc.), passando por elementos de web scraping, até a criação de suas próprias funções/comandos. Enfatizaremos ainda boas práticas para organização dos códigos (i.e. sintaxes/scripts), abordaremos métodos e protocolos ótimos para organização da estrutura de pastas para análise de dados, bem como estratégias para maximizar a replicabilidade das análises.

O conteúdo deste curso é pré-requisito para a disciplina Lego 1.

2 Dinâmica das Aulas

O curso é composto de 4 (quatro) aulas expositivas, ministradas sempre em um laboratório de informática. O conteúdo expositivo é eminentemente prático.

3 Leituras e Materiais Suplementares

Não haverá leituras obrigatórias neste curso. Recomendamos, contudo, o excelente livro de Wickham and Grolemund (2016) como texto de referência. Outros materiais suplementares serão ainda indicados, para cada um dos temas abordados, incluindo vídeo-aulas, exercícios on-line e referências bibliográficas alternativas.

4 Avaliação

Quatro listas de exercício, que contarão pontos para a disciplina Lego 1.

5 Horários de Atendimento

Nos dias do curso, uma hora antes do início das aulas.

6 Conteúdo das aulas

- **Aula 1**
  - Visão geral
    - Download e instalação
    - R e RStudio
    - Obtendo ajuda
  - Algoritmos e Estrutura de dados
    - Objetos e Vetores simples
    - Matrizes, Listas e data.frames
- **Aula 2**
  - Lendo e gravando bancos de dados
    - Dados em CSV, Excel, Stata e SPSS
  - Estatísticas descritivas com R
    - Medidas de Tendência Central
    - Medidas de Dispersão

    - Medidas de Posição
    - Tabulações básicas: frequências e cruzamentos
  - O pacote dplyr

- **Aula 3**
  - Elaborando gráficos com o R
  - O pacote ggplot2: uma gramática para gráficos
  - Realizando tarefas repetitivas automaticamente: loops da família apply

- **Aula 4**
  - Realizando tarefas repetitivas automaticamente: loops do tipo for
  - Princípios de web scraping: coletando dados da internet automaticamente
  - Programação: construindo suas próprias funções no R

Referências

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.