

DEPARTAMENTO DE ESTATÍSTICA

24 de dezembro de 2022 (Feliz Natal!)

Lista 2: Manipulação em Bancos de dados e em Spark com R.

Computação em Estatística para dados e cálculos massivos Tópicos especiais em Estatística 1

Prof. Guilherme Rodrigues César Augusto Fernandes Galvão (aluno colaborador) Gabriel Jose dos Reis Carvalho (aluno colaborador)

- 1. As questões deverão ser respondidas em um único relatório *PDF* ou *html*, produzido usando as funcionalidades do *Rmarkdown* ou outra ferramenta equivalente.
- 2. O aluno poderá consultar materiais relevantes disponíveis na internet, tais como livros, blogs e artigos.
- 3. O trabalho é individual. Suspeitas de plágio e compartilhamento de soluções serão tratadas com rigor.
- 4. Os códigos R utilizados devem ser disponibilizados na integra, seja no corpo do texto ou como anexo.
- 5. O aluno deverá enviar o trabalho até a data especificada na plataforma Microsoft Teams.
- 6. O trabalho será avaliado considerando o nível de qualidade do relatório, o que inclui a precisão das respostas, a pertinência das soluções encontradas, a formatação adotada, dentre outros aspectos correlatos.
- 7. Escreva seu código com esmero, evitando operações redundantes, visando eficiência computacional, otimizando o uso de memória, comentando os resultados e usando as melhores práticas em programação.

Por vezes, mesmo fazendo seleção de colunas e filtragem de linhas, o tamanho final da tabela extrapola o espaço disponível na memória RAM. Nesses casos, precisamos realizar as operações de manipulação fora do R, em um banco de dados ou em um sistema de armazenamento distribuído. Outas vezes, os dados já estão armazenados em algum servidor/cluster e queremos carregar para o R parte dele, possivelmente após algumas manipulações.

Nessa lista repetiremos parte do que fizemos na Lista 1. Se desejar, use o gabarito da Lista 1 em substituição à sua própria solução dos respectivos itens.

Questão 1: Criando bancos de dados.

- a) Crie um banco de dados SQLite e adicione as tabelas consideradas no item 2a) da Lista 1.
- b) Refaça as operações descritas no item 2b) da Lista 1 executando códigos sql diretamente no banco de dados criado no item a). Ao final, importe a tabela resultante para R. Não é necessário descrever novamente o que são as regiões de saúde.

Atenção: Pesquise e elabore os comandos sql sem usar a ferramenta de tradução de dplyr para sql.

- c) Refaça os itens a) e b), agora com um banco de dados MongoDB.
- d) Refaça os itens c), agora usando o Apache Spark.
- e) Compare o tempo de processamento das 3 abordagens (SQLite, MongoDB e Spark), desde o envio do comando sql até o recebimento dos resultados no R. Comente os resultados incluindo na análise os resultados obtidos no item 2d) da Lista 1.

Cuidado: A performance pode ser completamente diferente em outros cenários (com outras operações, diferentes tamanhos de tabelas, entre outros aspectos).