**R para Ciência de Dados**

Atividade Prática Supervisionada

Professor: Paulo C. Marques F.

Primeiro semestre de 2021

**Configuração dos grupos**

Esta atividade prática supervisionada deverá ser feita em grupos de até 4 integrantes. Se desejado, a atividade poderá ser feita individualmente. Não será permitida a formação de grupos com integrantes de outra turma. Um dos integrantes do grupo deve enviar um e-mail para o professor ([PauloCMF1@insper.edu.br](mailto:PauloCMF1@insper.edu.br)) até o dia 7 de Maio informando os nomes dos integrantes.

**Apresentação e entrega**

Os grupos apresentarão o relatório final da atividade durante a aula do dia 31 de Maio. Todos os integrantes do grupo deverão ser capazes de discutir o conteúdo do relatório. O relatório da atividade deverá ser redigido em R Markdown. A partir do R Markdown, gere um arquivo HTML e faça a entrega, até 23h59 do dia 31 de Maio, via link do Blackboard, juntando o arquivo HTML, o arquivo .Rmd original e o arquivo de dados. Coloque todos os arquivos em um único .zip para fazer a entrega. Apenas um dos integrantes do grupo precisa fazer a entrega no Blackboard. Contacte o professor assistente ([MagnoTFS@insper.edu.br](mailto:MagnoTFS@insper.edu.br)) caso haja dúvidas quanto à formatação do relatório.

**Conteúdo**

O grupo deve escolher algum conjunto de dados para análise utilizando as técnicas desenvolvidas no curso. Os links abaixo contém algumas possibilidades.

1. Base dos dados (<https://basedosdados.org/>)
2. Portal Brasileiro de Dados Abertos (<https://dados.gov.br/>)
3. UCI Datasets (<https://archive.ics.uci.edu/ml/datasets.php>)
4. Google Dataset Search (<https://datasetsearch.research.google.com/>)
5. Ipea Data (<http://www.ipeadata.gov.br/Default.aspx>)

Cada grupo deverá utilizar um conjunto de dados diferente dos utilizados pelos demais grupos da turma. O grupo deverá informar ao professor por e-mail qual conjunto de dados foi escolhido, além de fornecer o link para acesso aos dados originais. **O primeiro grupo a solicitar um determinado conjunto de dados terá prioridade na escolha.** Nos avisos do Blackboard ficarão listados os conjuntos de dados já escolhidos pelos grupos.

**Muito importante: não será permitido utilizar o mesmo conjunto de dados utilizado por qualquer um dos integrantes do grupo na disciplina de Python no semestre passado.**

Para o conjunto de dados escolhido, faça o maior número possível de visualizações. O site *From Data to Viz* (<https://www.data-to-viz.com/>) é uma boa fonte de ideias para a construção de gráficos. Sempre utilize a gramática dos gráficos e a biblioteca ggplot. As transformações de dados devem seguir o estilo dos “cinco verbos fundamentais” desenvolvido nas aulas. A ênfase deve ser na utilização de dados em formato *tidy* de maneira a facilitar as transformações e visualizações dos dados. No início do relatório faça uma descrição impecável que justifique o interesse nos dados analisados. Para atingir a nota máxima, o grupo deve incluir algum modelo de regressão ou classificação, seja com o propósito de inferência, ou de predição.