Opções de Conjuntos de Dados

Escolha um dos conjuntos de dados abaixo ou use um seu próprio (veja abaixo considerações para este caso).

Conjunto de Dados	Visão Geral	Questão Guia	Estimativa de Tempo
Red Wine Quality[1] Leia este arquivo texto que descreve as variáveis e como os dados foram coletados.	Este conjunto de dados contém 1.599 vinhos tintos com 11 variáveis de propriedades químicas do vinho. Ao menos 3 especialistas em vinhos avaliaram cada vinho, fornecendo uma nota entre 0 (muito ruim) e 10 (muito excelente).	Quais propriedades químicas influenciam a qualidade dos vinhos tintos?	10-20 horas
White Wine Quality[2] Leia este arquivo texto que descreve as variáveis e como os dados foram coletados	Este conjunto de dados contém 4.898 vinhos brancos com 11 variáveis de propriedades químicas do vinho. Ao menos 3 especialistas em vinhos avaliaram cada vinho, fornecendo uma nota entre 0 (muito ruim) e 10 (muito excelente).	Quais propriedades químicas influenciam a qualidade dos vinhos brancos?	10-20 horas
Financial Contributions to Presidential Campaigns by State	Selecione uma eleição usando os botões e clique em “Export Contributor Data” para obter os conjuntos de dados. Escolha UM estado e explore as contribuições feitas para um candidato em um ano de eleições.	Faça suas próprias perguntas sobre este conjunto de dados. Você pode adicionar variáveis a este conjunto de dados como sexo ou partido político do candidato.	15-30 horas
Loan Data from Prosper	Este conjunto de dados possui 113.937 empréstimos com 81 variáveis em cada um, incluindo o valor, taxa de juros, status do	Faça suas próprias perguntas sobre este conjunto de dados. Existem MUITAS variáveis	15-30 horas

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236. Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

Última atualização em 11/03/2014 Este dicionário de variáveis explica as variáveis do conjunto de dados.	pagamento, receita do mutuário, seu emprego atual, histórico do cartão de crédito e informações sobre seu último pagamento.	neste conjunto de dados e você não deverá explorar todas. Escolha entre 10 a 15 variáveis para sua análise.	
Encontre seu próprio conjunto de dados!	**Lembre-se que encontrar e limpar o conjunto de dados é uma tarefa que demanda tempo e esforço significativos!** Veja a lista abaixo caso você deseje utilizar seu próprio conjunto de dados.	Faça suas próprias perguntas sobre o conjunto de dados!	30+ horas

Caso você esteja usando um conjunto de dados próprio...

Seu conjunto de dados deve:

- ❏ possuir ao menos 1.000 observações
- ❏ conter ao menos uma variável categórica (você pode criar uma)
- ❏ conter ao menos 8 variáveis diferentes
- ❏ estar em um formato "limpo"[1] (você pode ter que realizar a limpeza e formatação dos seus dados como parte da exploração)
- ❏ estar em um formato usual, como .csv, .tsv, .txt, ou .xls

Aqui estão algumas fontes para encontrar conjuntos de dados:

- http://www.inside-r.org/howto/finding-data-internet (não utilize o conjunto de dados do Titanic)
- http://opendata.stackexchange.com/
- http://www.data.gov/

[1] Conjuntos de dados "limpos" (tidy) são aqueles que possuem uma estrutura particular. Leia mais sobre este tipo de conjunto de dados no artigo de Hadley Wickham's, http://vita.had.co.nz/papers/tidy-data.pdf