Data Set Options

Choose from one of the following data sets or find your own (see below if you're finding your own).

Data Set	Overview	Guiding Question	Time Estimate
Red Wine Quality[1] Read this text file which describes the variables and how the data was collected.	This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).	Which chemical properties influence the quality of red wines?	10-20 hours
White Wine Quality[2] Read this text file which describes the variables and how the data was collected.	This tidy data set contains 4,898 white wines with 11 variables on quantifying the chemical properties of each wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).	Which chemical properties influence the quality of white wines?	10-20 hours
Financial Contributions to Presidential Campaigns by State **(2/6/18 Note: This site has changed and data is currently inaccessible. We're working to update our access.)**	Select an election year from the radio buttons and press the "Export Contributor Data" button to get downloadable datasets. Choose ONE state and explore financial contributions made to Presidential candidates in a given election year.	Ask your own questions about this data set. You may want to add variables to this data set such as the gender or political party of the candidate.	15-30 hours
Loan Data from Prosper Last updated 03/11/2014 This variable dictionary explains the variables in the data set.	This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information.	Ask your own questions about this data set. There are MANY variables in this data set and you are not expected to explore all of them. You should explore between 10-15 variables in your analysis.	15-30 hours
Find your own data set!	**Remember that finding and cleaning your own data set**	Ask your own questions about your data set!	30+ hours

	could take significant time and effort! See the checklist below if you want to choose your own data set.		

If you’re finding your own data...

Your data set should include:

❑ at least 1,000 observations
❑ contain at least one categorical variable (you may create one)
❑ contain at least 8 different variables
❑ be in a tidy format[1] (you may need to clean and reshape the data as part of your exploration)
❑ the data set should be in a commonly used format such as .csv, .tsv, .txt, or .xls

Here are a few resources to find a data set:

- http://www.inside-r.org/howto/finding-data-internet (do not use the Titanic data set)
- http://opendata.stackexchange.com/
- http://www.data.gov/

[1] Tidy data sets are data sets that have a particular structure. Read more about tidy data in Hadley Wickham’s paper, http://vita.had.co.nz/papers/tidy-data.pdf

[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