June 2018

Data Science in Global Health
Programming Library/Resource List

Overview

Getting started with data science can be a daunting task. Identifying and understanding the appropriate methods and techniques presents one challenge, while producing the code necessary to implement those methods and techniques presents another challenge.

MEASURE Evaluation's Programming Library/Resource list seeks to address those twin challenges by providing links to existing resources that present an overview of data science techniques and example code for applying those techniques. The links in this resource list are intended for global health professionals with basic to moderate programming skills and who are interested in strengthening their knowledge and experience in data science methods and techniques.

Links with R code		
Title	Description	Link
UCLA Resources	This site provides multiple links to R resources including links for downloading software.	https://stats.idre.ucla.edu/r/
	The site includes links to free online modules introducing fundamentals of R including importing data, variable construction, and data visual representation.	https://stats.idre.ucla.edu/r/modules/
	Additional links provide code for data visual representation and various statistical methods including descriptive statistics, logistic regression, and multilevel models.	https://stats.idre.ucla.edu/r/codefragments/introduction/ https://stats.idre.ucla.edu/other/dae/
UNC Resources	Open course materials presented as separate lectures for various statistical methods, including multiple negative binomial regression. The site also includes lectures introducing Bayesian methods for regression analysis, and graphical representations in R. Each lecture page provides sample R code for each method presented.	https://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm
Rspatial	This site provides code and descriptions for spatial data analysis in R, as well as links to download various spatial analysis packages. The site also includes an introduction to manipulating spatial data in R.	http://www.rspatial.org/
Code Index	Collection of code snippets for specific tasks	https://source.opennews.org/code/
Starting data analysis/ wrangling with R: Things I wish I had been told	Descriptions of data wrangling techniques and associated R code	http://reganmian.net/blog/2014/10/14/starting-data-analysiswrangling-with-r-things-i-wish-id-been-told/
R Reference Card for Data Mining	List of R packages and functions that can assist with data mining	http://www.rdatamining.com/docs/r-reference-card-for-data-mining
CausalImpact: Estimating causal effects in time series	Open-source R package developed by Google that makes causal analyses simple and fast. based on Bayesian structural time-series models. We use these models to construct a synthetic control—what would have happened to our outcome metric in the absence of the intervention. This approach makes it possible to estimate the causal effect that can be attributed to the intervention, as well as its evolution over time.	https://opensource.googleblog.com/2014/09/causalimpact-new-open-source-package.html
RAPPOR	R and Python open source project from Google to facilitate analysis of data while preserving privacy of individuals.	https://github.com/google/rappor https://ai.googleblog.com/2014/10/learning-statistics-with-privacy-aided.html

Mapping with ggplot: Create a nice choropleth map in R	Basic tutorial on how to produce choropleth maps in R	http://rforpublichealth.blogspot.com/2015/10/mapping-with-ggplot-create-nice.html
Introduction to R Markdown (R Notebooks)	R Markdown is the R version of Jupyter notebooks. Code can be written and executed in a document that makes it possible to present findings and the output of analysis but also include the code used to conduct analysis in a way that makes reproducible results possible.	https://rmarkdown.rstudio.com/r_notebooks.html
Flexdashboard: Easy interactive dashboards for R	Interactive dashboards produced using R Markdown. Can be linked with Shiny to make visualizations dynamic.	https://rmarkdown.rstudio.com/flexdashboard/
Implementation of a reproducible data analysis workflow	Overview of best practice for creating a workflow in R that can be replicated. Replicable workflows make it possible to execute analysis of data by launching one R script and can facilitate collaboration.	http://blog.jom.link/implementation_basic_reproductible_workflow.html
Access WHO Global Health Observatory Data from R	Github repository of code to access data from the WHO Global Health Observatory within R	https://github.com/pierucci/rgho https://cran.r-project.org/web/packages/WHO/vignettes/who_vignette.html
Obtaining DHS Data via API	Code snippets for access to DHS data via an API	https://api.dhsprogram.com/#/samples-python.cfm http://api.dhsprogram.com/#/samples-r.cfm
	Apps to visualize or map data	http://api.dhsprogram.com/#/sample-apps.cfm
	DHS.rates R package	https://cran.r-project.org/web/packages/DHS.rates/vignettes/DHS.rates.html

Links with Python code

Title	Description	Link
Python for data science	This short primer on Python is designed to provide a rapid "on-ramp" to enable computer programmers who are already familiar with concepts and constructs in other programming languages learn enough about Python to facilitate the effective to use of open-source and proprietary Python-based machine learning and data science tools.	http://nbviewer.jupyter.org/github/gumption/Python_for_Data_Science/blob/master/Python_for_Data_Science_all.ipynb
Python basics for data science	Tutorial on using Python and Jupyter notebooks. Provides an overview of key Python libraries and techniques and an introduction to Jupyter notebooks.	https://data36.com/python-for-data-science-python-basics-1/
Bayesian First Aid: Pearson Correlation Test	Understanding the degree of correlation between two variables is a key step in effective use of data for decisions. Bayesian analysis offers the opportunity to assess correlation in data that doesn't have a normal distribution.	http://www.sumsar.net/blog/2014/03/bayesian-first-aid-pearson-correlation-test/

General Data Science Links

Title	Description	Link
Data Science Toolbox	The Data Science Toolbox is a virtual environment based on Ubuntu Linux that is specifically suited for doing data science. Its purpose is to get you started in a matter of minutes. You can run the Data Science Toolbox either locally (using VirtualBox and Vagrant) or in the cloud.	http://datasciencetoolbox.org/
Cross Validated	Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization.	https://stats.stackexchange.com/

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of the MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. FS-18-280

USAID FROM THE AMERICAN PEOPLE PEPFAR MEASURE Evaluation