Short Checklist for Learning Health Data Science

A short guide to getting started learning data science to use in health and related fields

by Ginger Dixon, MS, One Life Epi Solutions www.onelifeepisolutions.com

Learn the Basics

- ☐ Basic statistics: means and medians, normality, hypothesis testing, t-tests, z-tests, one-tail versus two-tail testing, line of best fit, residuals, positive and negative correlations, left and right skews, samples of study population
- ☐ Basic principles of study design: correlation versus causation, representative samples for a population, measures of association for case control, cohort, and clinical trial studies

Decide on a Programming Platform

- ☐ SAS: paid platform, good for statistical analysis, data cleaning, modeling, used frequently in government, large organizations, university/academic work
- ☐ Stata: paid platform, good for statistical analysis, data cleaning, and modeling
- ☐ Python: free platform, good for a wide variety of programming, data cleaning, and modeling
- ☐ R: free platform, good for a wide variety of programming, data cleaning, and modeling; used frequently in biological sciences
- ☐ MATLAB: paid platform, good for programming, data cleaning, and modeling; used frequently in engineering
- ☐ Excel: free and paid versions, good for small datasets, calculations, visualizations; used to import and export files to and from other programs

Additional Tools

- ☐ Tableau: free and paid versions, used to build interactive visualizations and dashboards
- ☐ Power BI: free and paid versions, used to build interactive visualizations and dashboards
- ☐ SQL: variety of versions and integrations with other platforms, language used to work with data in databases
- ☐ Jupyter Notebook: free web-based platform, used to write text and code for Python

Cleaning Data

- ☐ Import and export data into a platform
- ☐ Merge and/or split data as needed: evaluate type of merge needed
- ☐ Group, filter, calculate, and/or drop variables/columns as needed
- ☐ Evaluate extraneous datapoints and decide how to deal with them
- ☐ Evaluate missing datapoints and patterns and decide how to deal with them

Processes of Exploring Data

- ☐ Check distribution and patterns of single data variables/columns with histograms, scatter plots, line plots, box plots, bar graphs, density plots, heat maps, geographic heat maps, etc.
- ☐ Check relationship between variables/columns with scatterplots as appropriate
- ☐ Explore potential trends in time series data: plot by day, week, month, year, etc. and evaluate trends and lags
- ☐ Decide if and how you want to explore data further
- ☐ If modeling data, check assumptions for model type
- ☐ Calculate odds ratio or relative risk measures of association as appropriate

Basic Modeling

- ☐ Linear regression: used for data that is continuous, can be used to identify dependent variables that are relevant to the outcome of interest and develop a function model
- ☐ Logistic regression: used for data that is binary or categorical, can be used to identify dependent variables that are relevant to the outcome of interest and develop a function model
- ☐ Survival/Time-to-event analysis: can be used to identify time to event of interest, useful in population and pharmacologic studies

Advanced Modeling

- ☐ ARIMA/SARIMAX
- ☐ Neural Networks
- ☐ Vector Autoregression (VAR)
- ☐ Markov chain modeling methods
- ☐ Bayesian regression methods
- ☐ and many, many others!

Troubleshooting and Useful Resources

- ☐ www.onelifeepisolutions.com
- ☐ https://towardsdatascience.com/
- ☐ www.dataindependent.com
- ☐ www.geeksforgeeks.org
- ☐ www.datascienceplus.com
- ☐ www.dataquest.io
- ☐ www.correlation-one.com

For more information:

Take the course

https://www.onelifeepisolutions.com/hdspresale

Sign up for our newsletter

https://mailchi.mp/9f8fd2116983/subscribe