A Data Science Workshop: R Fundamentals and Biostatistics

Hosted by the Bakeng se Africa.

26-30 August 2019

Mon-Thurs: 09:00 – 17:00

Fri: 09:00 – 16:00

Location: Laboratory of Forensic Anthropology, Centre for Functional Ecology, Department of Life Sciences, University of Coimbra

Course Description

The goal of this course is provide you with the most important tools so you can do data science. The course provides an introduction to the R language and the environment, both of which are crucial for success using the program. Once the foundation is set, the course will transition into statistical analyses that are commonly employed in biomedical research ranging from hypothesis testing and outlier detection to supervised and unsupervised regression and classification techniques. The workshop will not focus on any one analysis or specific field; instead, the aim of the workshop is to teach the basic knowledge necessary to use R independently, thus helping participants initiate their own research projects and produce publications.

The course is structured around hands-on practicals and interactive sessions that will ensure participants are comfortable with the R environment, running statistical analyses, and interpreting statistical outputs by the time the course is completed.

No advanced statistics knowledge or prior R/programming experience is required.

The following topics will be addressed:

- R and the RStudio environment
- Data Manipulation: Import, Cleaning, Transforming
- R programming: The tidyverse
- Data Visualizations: GGplot2
- Descriptive Statistics, Outlier Detection, and Hypothesis testing
- Introduction to Supervised and Unsupervised Machine Learning Techniques
    - Categorical Analyses, Linear and Logistic Regression, Discriminant Function Analysis, Multiple Correspondence Analysis, Principal Component Analysis, Cluster Analysis, and Random Forest

Who should enrol?

Anyone that is interested in becoming a better data scientist: postgraduate students, researchers, and faculty. Course is limited to 25 participants.

No fees are required.

Technical aspects:

The course will be held in a computer laboratory and therefore participants are not required to bring their own laptop. However, if it is possible to bring a personal laptop, it is recommended so that you can establish a workspace that you can use once the course is finished. If you

bring a personal laptop it needs to runs Windows (2000, XP/2003/Vista/7/8/2012 Server/8.1/10) or Mac and have at least 32 MB of RAM and enough disk space for recovered files, image files, etc.

Certificates

The University of Coimbra will award delegates who successfully complete the required assignment and attend all sessions in the workshop a certificate of successful completion and/or attendance.

Program:

Monday August 26	Topic
8:00 – 8:30	Registration
8:30 – 9:00	Introduction to the Course
9:00 – 10:30	Introduction: Introducing R, Markdown, and Importing Data
10:30 – 10:45	Break
10:45 – 13:00	Introduction: Introducing R, Markdown, and Importing Data → Data Munge
13:00– 13:50	Lunch
13:50 – 15:20	Data Munge
15:20 – 15:30	Break
15:30 – 16:30	Data Munge
Tuesday August 27	
9:00 – 10:30	Visualizations with ggplot2
10:30 – 10:45	*Break*
10:45 – 13:00	Visualizations with ggplot2
13:00 – 13:50	***Lunch***
13:50 – 15:20	The Tidyverse
15:20 – 15:30	*Break*
15:30 – 17:00	*In-Class Practical #1*
Wednesday August 28	
9:00 – 11:00	Exploring Assumptions & Outliers
11:00 – 11:15	*Break*
11:15 – 12:00	Descriptive Statistics
12:00 – 13:00	Comparing Means
13:00 – 13:50	***Lunch***
13:50 – 14:20	Comparing Several Means
14:20 – 15:20	Categorical Data and Multiple Correspondance Analysis
15:20 – 15:30	*Break*
15:30 – 17:00	Correlations and *In-class Practical #2*
Thursday August 29	
9:00 – 9:30	Machine Learning Introduction
9:30 – 10:30	Linear and Logistic Regression
10:30 – 10:45	Break

10:45 – 13:00	Linear and Logistic Regression
13:00 – 13:50	***Lunch***
13:50 – 14:30	Linear and Logistic Regression
14:30 – 14:45	*Break*
14:45 – 17:00	Discriminant Function Analysis (DFA)
Friday August 30	
9:00 – 10:30	DFA
10:30 – 10:45	*Tea*
10:45 – 13:00	Data Reduction Techniques: PCA
13:00 – 13:50	***Lunch***
13:50 – 15:30	Cluster Analyses, Decision Trees, Shake-Rattle-Roll
15:30 – 16:00	Presentation of Certificates, Feedback, and Final Comments

Instructors

Kyra E. Stull, PhD, D-ABFA is an Assistant Professor in the Department of Anthropology at the University of Nevada, Reno. Dr. Stull earned her B.A. in Anthropology from the University of Tennessee, Knoxville (2006), a MS in Biological and Forensic Anthropology at Mercyhurst University (2008), and her PhD in Anatomy, with a concentration in Biological Anthropology, from the University of Pretoria (2014). Dr. Stull has research interests in forensic anthropology, human growth and development, modern human variation, and quantitative methods. Specifically, her research seeks to explore the predictive ability of age and sex indicators and to develop techniques to accurately identify unknown decedents. A large portion of her current research is dedicated to increasing the utility and accessibility of modern techniques and methods through intuitive graphical user interfaces that operate through R. She has provided professional workshops and has taught graduate and undergraduate courses on applied biostatistics using R since 2014.

David Navega, MS is an Anthropology doctoral candidate in the Department of Life Science at the University of Coimbra. David holds a bachelor's degree in Anthropology (2010), a master's degree in Legal Medicine and Forensic Sciences (2012), and a postgraduate diploma in Forensic Anthropology (2012). His research interests focus biological profile assessment, with emphasis on age and ancestry estimation, using statistical and machine learning tools. He developed several software tools for anthropological applications using the R programming language and statistical environment.