8th GESIS Summer School in Survey Methodology
Cologne, August 2019

Syllabus for Course 05: "Introductory Course to R with Applications from Data Analysis"

Instructors:	Dr. Jan-Philipp Kolb	Alexander Murray-Watters
E-mail:	Jan-Philipp.Kolb@gesis.org	Alexander.murray-watters@gesis.org
Homepage:	https://github.com/Japhilko	http://amurrayw.com/

Date: 12.-16. August 2019
Time: 09:00-13:00, 14:00-16:00
Course starts Monday morning at 09:00

About the Instructors:

Dr. Jan-Philipp Kolb is a senior researcher at the Leibniz Institute for Social Sciences (GESIS), working as a survey statistician in the GESIS Panel team. He previously worked in the GESIS Survey Statistics team and as a research assistant in the Economic and Social Statistics Department at the University of Trier teaching sampling techniques and applied statistics using R.

Alexander Murray-Watters is a research associate/doctoral candidate at the Leibniz institute for Social Sciences (GESIS) working as a statistics/machine learning expert. He previously worked as a Research Associate/Systems Analyst at Carnegie Mellon University in Pennsylvania, USA, where he also received a MS in Logic, Computation, and Methodology.

Selected Publications:

- Burgard, Jan Pablo, Jan-Philipp Kolb, Hariolf Merkle, and Ralf Münnich. 2017. "Synthetic data for open and reproducible methodological research in social sciences and official statistics." AStA Wirtschafts- und Sozialstatistisches Archiv first online 1-12.
- Münnich, Ralf, Jan P. Burgard, Siegfried Gabler, Matthias Ganninger, and Jan-Philipp Kolb. 2016. " Small area estimation in the German Census 2011." Statistics in Transition new series and Survey Methodology 17 (1): 25-40.
- Kolb, Jan-Philipp. 2016. "Visualizing GeoData with R." Austrian Journal of Statistics 45 (1): 45-54. doi: dx.doi.org/10.17713/ajs.v45i1.88.
- Murray-Watters, A. & Glymour, C. (2015). What Is Going on Inside the Arrows? Discovering the Hidden Springs in Causal Models. Philosophy of Science, 82 (4): 556-586. [Erratum: 2016. 83 (1): 170]. DOI: 10.1086/682962 and DOI: 10.1086/684247. http://www.journals.uchicago.edu/doi/abs/10.1086/682962

Short Course Description:

The open source software package R is free of charge and offers standard data analysis procedures as well as a comprehensive repertoire of highly specialized processes and procedures, even for complex applications. Emphasis in this course will be on methods of graphically-based data analysis as R is particularly suitable for it.

Keywords:

R, data import and export, data cleaning, statistical graphics, data analysis, survey data, web scrapping

Course Prerequisites:

- prior experience with data analysis, basic statistics, and regression;
- as we will work with GESIS Panel data, it would be good to download the campus file of the GESIS Panel (https://www.gesis.org/en/gesis-panel/data/gesis-panel-campus-file/);
- basic familiarity with the use of a computer.

Target Group:

- this course is for people that work with survey data and want to use R as an additional tool;
- the participants should have already attended an introductory event in statistics. Experience in dealing with other statistics packages is helpful, but not a requirement.

Course and Learning Objectives:

Our workshop provides a hands-on introduction to R and lays the foundations for independently developing your skills in dealing with the programming language R. The participants can expect to receive an overview of the functional scope of R, master the import and export of data, and how to perform basic data analysis in R.

Organizational Structure of the Course:

The best way to learn R is to try things out and apply the presented concepts. Therefore, we will have a mixture of classroom instruction (about three hours per day), hands-on exercises/lab sessions (about two hours per day) and contact time for individual consultations on participants' projects.

Software and Hardware Requirements:

Course participants will need to bring a personal laptop in order to do the hands-on exercises. Ideally, the laptop already has R (https://cran.r-project.org/) and Rstudio installed (https://www.rstudio.com/). Both programs are free and open source.

It would be useful to have the following R-packages installed: lattice, ggplot2, tidyr, plyr, rvest, lme4, survey, readstata13, haven, xlsx, foreign, MASS, devtools, rio, visreg and reshape2.

Long Course Description:

Getting started
The first session will cover all preliminary topics. For example, Rstudio is a graphical user interface which makes beginning with R easier. Many of its available features and Add-Ins will be explained on the first day. In addition we will cover data import and export as well as the data processing.

Basic data analysis
In this part we will give an introduction to basic data analysis. We will use the campus file of the GESIS Panel for those purposes. We will cover basic statistical tests, various regression techniques (linear, logit, lasso and multilevel regression) and clustering methods. In addition, we will give a short introduction in how to use the survey package.

Graphics
Students will learn how to visualize data using the basic functions, the lattice package and the ggplot2 package.

Web scraping
Students will learn how to scrape data off of a website using several different methods and process the scrapped data into a usable format (e.g., using the tidyr package and regular expressions). There will also be a discussion of how APIs function and why they are used.

Functional programming paradigm
In this part we plan an introduction to basic programming. The basics of debugging and how the functional programming paradigm makes debugging easier will also be covered.

Specific topics
If time permits, the use of ggplot2 on geospatial data will also be covered.

Day-to-day Schedule and Literature:

Day	Topic(s)
1	Getting started with R and Rstudio, how to get help; understanding error messages, how to find out what an error means, data import Use Case: Data Processing of GESIS Panel data
	Suggested reading: http://www.r-bloggers.com/why-use-r/ Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. Social Science Computer Review, 36(1), 103-115.
2	Basic data analysis - t-tests, chi-square-tests; and the use of the survey package. General regression (linear, logit, multilevel, lasso etc.); clustering (kmeans)
	Suggested reading: Faraway, J. (2005): Linear Models with R. CRC Press. https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf Finch, W., Bolin, E., and Kelley, K. (2015) Multilevel Modeling Using R. Journal of Statistical Software, 62(1).
3	Graphics: basic and lattice plots as well as the use of the ggplot2 package.
	Suggested reading: https://ggplot2.tidyverse.org/
4	Web Scrapping and Data cleaning: Scraping data off of a webpage, the rvest package, basic regular expressions and an explanation of APIs. The use of tidyr, plyr, and reshape2 packages.
	Suggested reading: https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/ http://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/
5	Loops, Functions and packages; the advantages of the functional programming paradigm (debugging/parallelization), and recursion. Use of the "with" function (instead of "attach"). Specific Topics: The basics of plotting geolocation data.
	Suggested reading: http://adv-r.had.co.nz/Functional-programming.html https://geocompr.robinlovelace.net/

Preparatory Reading:

No preparatory reading is necessary as this is an introductory course on R.

Additional Recommended Literature:

- Bradnam, K., & Korf, I. (2012). UNIX and Perl to the rescue!: a field guide for the life sciences (and other data-rich pursuits). Cambridge University Press.
- Burns, P. (2011). The R inferno. http://www.burns-stat.com/documents/books/the-r-inferno/
- Matloff, N. (2011). The art of R programming: A tour of statistical software design. No Starch Press.
- Muenchen, B. (2011)- R for SPSS and SAS Users. Springer Science & Business Media.
- Schutt, R., O'Neil, C. (2013). Doing data science: Straight talk from the frontline. " O'Reilly Meida, Inc.".

- Shalizi, C. (2013). Advanced data analysis from an elementary point of view.
- http://www.stat.cmu.edu/%7Ecshalizi/ADAfaEPoV/
- Spector, P. (2008). Data manipulation with R. Springer Science & Business Media.
- Teetor, P. (2011). R Cookbook: Proven recipes for data analysis, statistics, and graphics. " O'Reilly Media, Inc.".
- Wickham, H. (2011). The split-apply-combine strategy for data analysis. Journal of Statistical Software, 40(1), 1-29.
- Google's R Style Guide. https://google.github.io/styleguide/Rguide.xml