Master Tutorial

TITLE

Efficient Data Wrangling and Visualization with R

SHORTENED TITLE

Data Wrangling with R

ABSTRACT

The majority of an analyst’s time is spent on cleaning, formatting, and manipulating raw data. This tutorial session will take you through the steps needed to get messy data into R, clean it, and produce useful insights with as little code as possible. Bring your laptop for this interactive session (download session materials here: ).

PRESS PARAGRAPH

Researchers and analysts often spend much of their time simply trying to get raw data into a usable state. Proficiency in data manipulation is especially important for those who wish to harness the power, efficiency, and cost-effectiveness of writing ones own scripts in R. In this hand-on tutorial, we will show how to efficiently transform seemingly unusable data from different sources (including SQL databases) into something easily described, summarized, and visualized, all without leaving the R console.

WORD COUNT: 1157

**Efficient Data Wrangling and Visualization with R**

R (R Core Team, 2018) is an open-source programming language that is designed for statistical computing (Hornik, 2013). R can perform anything from standard data analysis (e.g. Multiple Regression, Hierarchical Linear Modeling, or Structural Equation Modeling) to highly specialized computations that may be unique to a scientific field. R is a programming *language* and not just a statistical analysis package. Many data scientists and practitioners can contribute to R by writing new and unique software, called “packages” in R. At the time of this writing, the Comprehensive R Archive Network (CRAN) contains 12,959 available packages, including packages to read data in varying formats (e.g., readr, open.xlsx, haven, rjson, officer), access databases (e.g., DBI, odbc, RSQLite), clean data (e.g., dplyr, tidyr, stringr, reshape2), perform data analyses and machine learning (e.g., caret, xgboost, randomForest, caret, survival), visualize results (e.g., shiny, ggplot2), and interface with other programming languages (e.g., Rcpp, reticulate, RJava). These packages, just like R itself, are free of charge.

This Master Tutorial will teach attendees how to leverage R and several R packages to efficiently extract, combine, and clean data from various sources (including SQL databases) for applying standard analytical techniques and visualization packages, in particular ggplot2. Users should have a basic understanding of R syntax, types (e.g. numeric and character), and data structures (in particular data.frames). Given a basic understanding of R, we will teach users how to efficiently manipulate, join, rearrange, and clean data. we will then show users how to quickly and effectively extract meaning from data with functional syntax.

*Proposed Session*

The proposed tutorial is a continuation of the authors’ ongoing attempt to make R more accessible to the I/O community, encourage good coding standards, and lower the entry barrier for scientists and practitioners. R’s user base is quickly increasing. Due to its flexibility, power, and freedom from costly and intrusive licenses, R usage has surpassed the most popular packages like SPSS and SAS in the academic realm (Muenchen, 2015). R has recently been ranked as fifth among the top ten programming languages (Smith, 2016), and recent polls (Piatetsky, 2015) indicated that R is the most popular analysis software among data scientists. Furthermore, universities are increasingly turning to R as the tool of choice in their advanced statistics and research methods courses, arguably making proficiency with R a requirement for quantitative researchers in general. Given the increasing importance of R for data analysis, the need for people who are able to independently perform their data wrangling in R is higher than ever.

Note that the tutorial will be interactive. Interested attendees will be encouraged in advance of the conference to download the R software and packages. Materials for the tutorial will be provided using either SIOP’s mySIOP.org file repository or a link to a file repository from which the users can download.

Audience members are strongly encouraged to bring laptops and to have downloaded the materials ahead of time. For those who wish to follow along, we will make available all materials and R scripts at . We request 80 minutes for the tutorial, with the approximate time for each topic as well as additional information provided below.

**Topic #1: Reading In and Cleaning Data (30 minutes)**

After an initial setup, we will demonstrate how to import data into R from flat files, Excel spreadsheets, and remote SQL databases. Once data have been read into R, we will outline common issues with data entry and formatting as well as the simple (and sometimes esoteric) ways to remedy them in R. From there we will introduce piping (a useful functional coding practice) to quickly and expressively apply the extremely popular dplyr (Wickham, …) package to filtering rows of the data, selecting columns, renaming variables, and creating variables from preexisting variables (such as standardization or sten scoring).

**Topic #2: Data Manipulation (20 minutes)**

Once foundational skills have been established, we will move onto the more advanced topics of reshaping data frames and joining different datasets by common variables. In particular we will focus on the differences between long and wide data as well as the key differences between the most common join operations. By the end of this section, participants will have been exposed to the core skills required for data wrangling.

**Topic #3: Getting insights (20 minutes)**

Finally, we will show participants how to quickly and effectively gain insights from their data by grouping cleaned data on categorical variables, like SES or country, and summarizing measured variables, such as test scores or simple tabulations, within each of the groups. By grouping on required variables, summarizing within groups, and reshaping data into various formats, one can easily construct complex, fine grained summary tables. Additionally, we will show participants how to pipe their summarized data into ggplot2, a package that while having a reputation for being difficult, becomes significantly easier when data are formatted a specific way.

**Topic #4: Wrap-up (10 minutes)**

The presenter will try to leave a 10 minute buffer at the end of the session. This 10 minute buffer will be used to answer audience questions and help with technical problems. The presenters will point participants to sections in the syntax script that could serve as good starting points for further self-study. By the end of this section, users should be more confident when importing data, exploring data, and constructing descriptive results. Skills learned should reduce the time spent beginning R projects, result in more efficient and cleaner code, and into time savings and lesson R-induced stress.

**Learning Objectives**

By the end of this workshop, you should be able to:

1. Understand common flat file formats and how to read/import them into R.
2. Send basic SQL queries to a remote database.
3. Fix common data entry problems, such as non-standard missing values
4. Use pipelines and efficient coding to connect your data preparation process.
5. Perform advanced column selection and row filtering.
6. Reshape data into a tidy format.
7. Understand and use left, right, and inner joins to combine different datasets on common variables.
8. Perform complex, aggregated summaries of key variables within groups.
9. Pipe data into ggplot2 in the (long) format the package expects.
10. Write cleaner, more concise R code.

**Presenter Information**

Jeff Jones

Director, Talent Analytics and Data Sys

Korn Ferry

33 South Sixth Street

Suite 4900

Minneapolis, MN 55402

Phone: 612-373-3418

Email: jeff.jones@kornferry.com

Membership Status: Member

Steven Nydick

Data Scientist Programmer

Korn Ferry

33 South Sixth Street

Suite 4900

Minneapolis, MN 55402

Phone: 612-373-3548

Email: steven.nydick@kornferry.com

Benjamin Wiseman

Data Scientist Programmer

Korn Ferry

33 South Sixth Street

Suite 4900

Minneapolis, MN 55402

Phone: 612-373-3445

Email: benjamin.wiseman@kornferry.com

**Presenter Bios**

Steven Nydick is a Data Scientist Developer at the Korn Ferry Institute, where he designs R-based tools and scoring algorithms. He is the lead author and maintainer of the catIrt R package as well as several internal R packages helping with everything from plotting to powerpoint generation to interfacing with servers. He has contributed to developing psychometric models and corresponding estimation algorithms that have been published in *Applied Psychological Methods* and the *Journal of Educational and Behavioral Statistics*. Steven received his Ph.D at the University of Minnesota in Psychometrics and Quantitative Psychology, where he primarily studied IRT-based adaptive test for selection and classification. He also has an M.S. in Statistics from the University of Minnesota.

Ben Wiseman is a Data Science developer at the Korn Ferry Institute responsible for maintaining and developing R-based automation tools, models, reports, and user interfaces. He has publications in entomology, ecology, and molecular evolution and has worked with and trained numerous clients in the military, public, and private sectors on a wide range of applications. Ben received his MSc from Lincoln University (New Zealand) in applied statistical modelling where he developed a user-facing geospatial AI platform for DOCs predator monitoring and control systems.

[[I don’t know how much detail about non-psychology stuff is relevant]]

**Appendix**

CV Jeff Jones

CV Steven Nydick

CV Benjamin Wiseman

Steven Nydick

Email: Steven.Nydick@KornFerry.com

**Education:**

PhD, University of Minnesota, Psychometrics/Quantitative Psychology, 2013.

Advisor: Niels Waller

MA, University of Minnesota, Psychometrics/Quantitative Psychology, 2012.

Advisor: Niels Waller

MS, University of Minnesota, Statistics, 2011.

Advisor: Sanford Weisberg

BS, Syracuse University, Mathematics and Psychology, 2006.

**Professional Experience:**

Data Scientist Developer, Korn Ferry, 2018 – Present.

Senior Psychometrician, Pearson VUE, 2016 – 2018.

Psychometrician, Pearson VUE, 2013 – 2016.

Research Assistant, University of Minnesota, 2013 – Present.

Intern in Psychometrics, ARRT, 2012 – 2013.

Intern in Psychometrics, ACT, 2011.

Graduate Instructor/Section Leader, University of Minnesota, 2007 – 2013.

**Awards:**

Doctoral Dissertation Fellowship, 2013

Graduate Research Partnership Program, 2010

Archimedes Prize in Mathematics, 2006

**Manuscripts Published and In Press:**

Wang, C. & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-

compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*, 119-134.

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of*

*Educational and Behavioral Statistics*, *39*, 203-230.

**Software:**

Nydick, S. W. (2014). catIrt: An R package for simulating computerized adaptive tests. R package version

0.5-0).

**Presentations and Workshops:**

Nydick, S. W. (2016, April). The expected likelihood in computerized classification testing. Paper

presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Nydick, S. W. (2014, April). Multidimensional mastery testing with CAT. Paper presented at the annual

meeting of the National Council on Measurement in Education, Philadelphia, PA.

Nydick, S. W., Wang, C., & Xiong, X. (2014, April). Measuring multidimensional growth—a higher-order

IRT perspective. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, April). Accuracy and efficiency in classifying examinees using

computerized adaptive tests: An application to a large scale test. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, BC.

Nydick, S. W., & Weiss, D. J. (2010, June). Accepting the null: No change in change CAT. Paper presented

at the IACAT conference on CAT, Arnhem, NL.

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure, evaluated for the development of

CATs. In D. J. Weiss (Ed.) *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

**Unpublished Manuscripts:**

Nydick, S. W. (2013). *Intro to R for Psychologists.* Minneapolis, MN: Author.

**Courses Taught:**

Introduction to Data Analysis/Statistics for Undergraduates

Honors Introduction to Data Analysis/Statistics for Undergraduates

Analysis of Psychological Data for Graduate Students

Steven Nydick

Email: Steven.Nydick@KornFerry.com

**Education:**

MS, Lincoln University, Applied Statistics, 2015.

BS, Lincoln University, Biostatistics, 2013.

**Professional Experience:**

Data Scientist Developer, Korn Ferry, 2018 – Present.

Owner, Wiseman Analytics, 2016 – 2018.

Information Services, DHS, 2015 – 2016.

Instructor, Lincoln University, 2013 – 2014.

Research Assistant, Lincoln University, 2011 – 2015.

Research Assistant, Seoul National University, 2011.

Graduate Instructor/Section Leader, University of Minnesota, 2007 – 2013.

**Awards:**

Freemasons university scholarship

Forest and Bird research award

AGLS research scholarship

**Manuscripts Published and In Press:**

Wiseman, BH., Fountain, ED., Bowie, MH. He, S., Cruickshank, RH. 2016. Vivid molecular divergence over volcanic remnants: the phylogeography of Megadromus guerinii on Banks Peninsula, New Zealand. New Zealand Journal of Zoology

Fountain, ED., Pugh, AR., Wiseman, BH., Smith, VR., Cruickshank, RH., and Paterson, AM. 2015. On the captive rearing of Hadramphus tuberculatus (Pascoe 1877) (Coleoptera: Curculionidae: Molytinae):is ex-situ conservation the lesser of two weevils? New Zealand Entomologist.

Gillespie, M., Cruickshank, RH., Wiseman, BH., Wratten, S. 2013. Incongruence between morphological and molecular markers in the butterfly genus Zizina (Lepidoptera: Lycaenidae) in New Zealand.Systematic Entomology 38:151-163.

Fountain, ED., Wiseman, BH., Cruickshank, RH., and Paterson, AM. 2013. The ecology and conservation of Hadramphus tuberculatus (Pascoe 1877) (Coleoptera: Curculionidae: Molytinae). Journal of Insect Conservation 17:737-745.

**Software:**

Wiseman, B. W., Nydick, S.W., Jones, J (2018) roperators: Additional Operators to Help you Write Cleaner R Code. R package version 1.0-1).

Wiseman, B. W. (2015) Neurofriendly: Artificial Neural Networks Made Simple

Wiseman, B. W. (2015) Geofriendly: Easy Spatial Application of Artificial Neural Networks

**Presentations and Workshops:**

Wiseman, B. H. 2017 Data Science with Python. ESRI Developer Summit, Palm Springs, CA.

Wiseman, B. H. 2013 Messy data, messy models and applied statistics. Presented for Bio-Protection seminar, Lincoln University, New Zealand.

Marris, J. and Wiseman, B. H. 2012. Islands in the snow: Ecology, systematics and biogeography of the New Zealand beetle genus Protodendrophagus (Coleoptera:Silvanidae:Brotini). Presented at the New Zealand Ecological Society conference.

Cripps, M., McNeil, M., Patrick, H., Wiseman, B., Nobilly, F., Edwards, G. 2012. Invertebrate abundance and diversity in intensively managed dairy pastures.New Zealand Plant Protection Society Conference.

Wiseman, B. H., Cruickshank, R. H., Bowie, M. H., Fountain, E. D. 2011. Unexpected genetic variation in an endemic ground beetle: The molecular mystery of Megadromus guerinii (Coleoptera: Carabidae). 3rdAnnual Combined Australian and New Zealand Entomological Societies Conference

Wiseman, B. H. (2011). The curious case of Megadromus guerinii: phylogeographic oddities on Bank’s Peninsula. Presented to the Canterbury branch of the New Zealand Entomological Society.

**Courses Taught:**

Research and Analytical Skills

Geospatial Information Systems with Arc GIS

Business Statistics

Intermediate Statistics for Commerce