GESIS Summer School in Survey Methodology 2023

Syllabus for course:
“Data Science Techniques for Survey Researchers”

Lecturers: Anna-Carolina Haensch
Email: anna-carolina.haensch@stat.uni-muenchen.de
Homepage: https://carohaensch.github.io/

Date: 21.- 25. August 2023
Time: Mo: 10:00-17:00 | Tu-Fr: 09:00-16:00
Venue: GESIS Cologne

About the lecturers
Anna-Carolina Haensch is a post-doctoral researcher at the LMU Munich and an assistant research professor at the University of Maryland (UMD). In 2023, she spent several months at New York University as Visiting Scholar. She was also involved in the Covid-19 Trends and Impact survey run by Facebook in cooperation with the University of Maryland and the Pandemic Recovery Survey. She enjoys teaching quantitative courses to Bachelor and Master students and is interested in Synthetic Data, Multiple Imputation, Big Data in the Social Sciences, and Statistics and Data Science training. From 2017 to 2021, she was a researcher at the GESIS Institute, where she worked on the DFG funded project “HaSpaD – Harmonizing and synthesizing partnership histories from different research data infrastructures”. In 2019, she joined the International Program in Survey and Data Science (IPSDS) Team and the University of Mannheim as a research assistant for a year.

Selected Publications:

Course Description:
A variety of digital data sources are providing new avenues for empirical social science research. In order to effectively utilize these data for answering substantive research questions, a modern methodological toolkit paired with a critical perspective on data quality is needed. This course introduces state-of-the-art data science techniques that are suited for collecting and analyzing digital behavioral data, so-called "big data", and traditional survey data. In addition, aspects of data quality and error frameworks for digital (big) data sources are discussed.

The course will cover the following topics and techniques:
- Overview of Big Data: What is it and why does it matter?
- Total Survey Error for Big Data
- Web Scraping
Data quality for gathered data types
Sampling from online material
(Supervised) Machine Learning for Social Scientists
Working with textual data: Text Mining and Topic Models

After the course, participants will have a profound understanding of important methods from the data science toolkit for collecting and analyzing the data types mentioned. They will be able to apply these methods and techniques in their research using statistical software.

Keywords:
Data Science, Big Data, Machine Learning, Text Mining, Total Survey Error, online

Course Prerequisites:
- General knowledge of statistics and statistical modelling (i.e., regression)
- Prior experiences with syntax-based software (like R, Stata, or Python)

Some basic experience with programming in R is helpful, but not strictly necessary. For those without prior exposure to R, we will ensure everyone is able to execute R markdown files. Students without any R knowledge are encouraged to work through one or more R tutorials prior of the course. Some resources can be found here:
https://rstudio.cloud/learn/primers
http://www.statmethods.net/
https://swirlstats.com/
https://rmarkdown.rstudio.com/lesson-1.html (for R Markdown)

Target Group:
Participants will find the course useful if:
- they are interested in learning some fundamental techniques in data science.
- they want to collect and work with digital behavioral data, be it administrative data or data found online.
- they want to understand what (supervised) machine learning is.

Course and Learning Objectives:
By the end of the course participants will:
- understand the challenges when analyzing digital behavioural data.
- know the promises and benefits of (supervised) machine learning.
- be able to use (supervised) machine learning for data analysis.
- be able to use common routines for analyzing textual data.
- learn some of the metrics used to assess data quality for gathered data types.

Organizational Structure of the Course:
The course is partly theoretical, partly practical. Each topic will be introduced in a lecture. The best way to deepen one’s understanding is with practical hands-on exercises. Files written in R Markdown will be provided to help participants execute the prepared scripts on their own computer and complete the assignments. The teacher will be available to assist and answer questions during the practical sessions.

Software and Hardware Requirements:
Course participants will need to bring a laptop with R (https://cran.r-project.org/) and RStudio installed (https://posit.co/download/rstudio-desktop/). Both programs are free and open source. We will use R for the practical sessions. We will inform you a few days before the course starts about recommended steps to setup your
You should be able to access the internet and install additional packages during the course (Wifi is provided by GESIS).

Day-to-day Schedule and Literature:

<table>
<thead>
<tr>
<th>Day</th>
<th>Topic(s)</th>
</tr>
</thead>
</table>
| 1      | Course introduction  
New data sources in the digital age  
New techniques to analyse such data |
|        | **Suggested reading (suggested, yet do not have to be read before class):**  |
|        |   ▪ Chapter 1 and 10 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020)  
| 1      | Web scraping, html, xml, json, APIs, regular expressions               |
|        | **Suggested reading:**  
| 2      | Data bases and SQL                                                    |
|        | **Suggested reading:**  
   ▪ Chapter 4 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020)  
   ▪ Chapter 15 in Baumer, B.S., Kaplan, D.T., & Horton, N.J. (2021) |
| 2      | Machine Learning I: Introduction                                     |
|        | **Suggested reading:**  
| 3      | Machine Learning II: Supervised learning                             |
|        | **Suggested reading:**  
<table>
<thead>
<tr>
<th>3  (afternoon)</th>
<th>Data quality for gathered data types (like Social Media Data, Sensor etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Machine Learning III</td>
</tr>
</tbody>
</table>
| **Suggested reading:** |  ▪ Chapter 10 and 15 in Hastie, T., Tibshirani, R., and Friedman, J. (2009)  
  ▪ Chapter 7 and 11 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020) |
| 5 | Text mining and Topic modelling |
| **Suggested reading:** |  ▪ Silge, Julia and David Robinson (2017). *Text mining with R: A tidy approach*. O'Reilly |

**Additional Recommended Literature:**

  [https://doi.org/10.1201/9780429324383](https://doi.org/10.1201/9780429324383)
- Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020). *Big Data and Social Science: Data Science Methods and Tools for Research and Practice* (2nd ed.). Chapman and Hall/CRC.  
  [https://doi.org/10.1201/9780429200717](https://doi.org/10.1201/9780429200717)