Data Analysis –
Advanced Statistics with Python

HS 2021

Dr. Julia Jerke
jerke@soziologie.uzh.ch
Thursday, 12.15 -13.45 Uhr
Location: AND 2.46

Course summary:

In 2012, the Harvard Business Review headlined "Data Scientist: The Sexiest Job of the 21st Century". Today, every human produces on average more than one giga byte of data – each day. The amount of data collected and stored is constantly increasing. Analyzing data, identifying relationships and patterns, making predictions, therefore, has become essential in the information age. As sociologists, you are well prepared for these challenges, with a broad statistical background and analytical thinking.
This course will, consequently, build on the statistics foundations taught in the bachelor program and further your knowledge of advanced data analytical methods. It is meant to be an overview to provide you with the proper tools and to enable you to further specialize in these areas in the future.
For the practical part of this course, we will draw on the programing language Python. Python is a very powerful and user-friendly programming language with a steep learning curve. According to the TIOBE index, Python is currently the second most popular programming language. It is a general-purpose language and, therefore, can be applied not only for statistical computations.

Course organization:

This course is part of the Sociology master's program. Students from other programs are warmly welcome as well. Previous knowledge of Python is not required. However, previous experience with the logic of statistical programming languages such as R or Stata are of advantage. Knowledge of statistical foundation is compulsory (e.g., univariate and bivariate sample description, hypothesis testing, linear regression).

The sessions will be composed of an introduction to the respective session topic and a practical part with Python. Due to the hands-on nature of the course, it is important that you bring your laptop to each session.

Exam

Upon successful completion of this course, 6 ECTS are credited. Students are required to actively contribute to the sessions. Throughout the course, 8 exercise sheets will be distributed that cover the topics of the sessions. To pass the course you need to meet the following two requirements:

(1) Successfully complete 6 out of the 8 exercise sheets (> 50% of the points)
(2) Obtain at least 60% of the total score across all exercise sheets

The exercise sheets must be submitted electronically 7 days after publication.

Course overview

Date	Session
23.09.2021	Introduction to Python • Course overview • Why Python? • Python installation • Python workflow • Getting to know the IDE *Spyder*
30.09.2021	Getting and exploring data *** • Getting to know the *Pandas* library • Importing data sets • Univariate sample statistics • Bivariate sample statistics • Getting to know *Jupyter Notebook*
07.10.2021	Linear regression *** • Repetition of the OLS principle • Prerequisites and interpretation of results
14.10.2021	Logistic regression *** • Introduction to logistic regression • Difference between linear and logistic regression • Interpretation of the coefficients
21.10.2021	Advanced regression methods • Overview of other regression methods • Brief introduction to some of these other regression methods • Empirical application examples
28.10.2021	Data visualization I *** • Getting to know the *matplotlib* library • Plotting univariate graphs of discrete and continuous data • Plotting bivariate data (e.g. scatter plots)
04.11.2021	*Special topic:* Git and Github • Getting to know the Git-workflow • Creating repositories • Pushing and pulling changes
11.11.2021	Cluster analysis *** • Introduction to cluster analysis • K-means clustering method • Hierarchical clustering method
18.11.2021	Principal component analysis (PCA) *** • Introduction to PCA • Running a simple PCA
25.11.2021	Data visualization II • Getting to know other Python graphic packages • Creating more advanced charts • Creating interactive charts

02.12.2021	*Special topic*: Introduction to machine learning • What is machine learning? • Supervised learning • Unsupervised learning • Application examples
09.12.2021	*Special topic:* Natural language processing *** • Introduction to NLP • Basic NLP with Python • Application examples
16.12.2021	*Special topic:* Databases and SQL *** • Where to get your data from • Working with databases • Requesting data sets using SQL

*** Exercise sheets are planned for these sessions. However, the assignment of the exercise sheets to the sessions may still change.