Advanced Data Analysis with Python

M. Fuat Kına

mkina@ku.edu.tr

This course, broadly speaking, is designed to familiarize the student with Python 3 and advanced data analysis techniques. We will cover core programming concepts using Python, which generally apply to programming. These include syntax, data types, functions, loops, and recursion. We will then cover database management, creation, manipulation, and visualization. A brief overview of Statistics, Causal Inference, Spatial Analysis, and Bayesian statistics with an emphasis on practical use will be followed by introductions to the most common machine learning methods. This is a demanding course, with the ultimate goal of a final project with an original analysis testing one or several hypotheses. The course consists of informative lectures on advanced data analysis techniques and helpful hands-on exercises.

Grading

- Three assignments: %15 x 3
- Midterm report: %5
- Final project: %40 (presentation, Github page, written report, and python code)
- Presentation of a weekly reading: %10

Detailed explanations for **the assignment** will be shared in the relevant weeks. You are expected to share your python work file via e-mail before the deadline. The readability of your code will be considered while grading. Therefore, please add inline comments to explain your codes if necessary.

The midterm report should not exceed one page. You are expected to define a research purpose, describe which methods might work, list possible sources you may utilize, and explain how you plan to access your data. You may think of this report as a brief research proposal. But please be precise! Your midterm reports will be graded when the feedback meeting is done. Do not hesitate to write me to schedule the appointment anytime.

The final version of **your project** is expected to be uploaded to your own GitHub page. Just send me the Github link of the project. If you don't have a GitHub account, don't worry! It is effortless

to learn. Unsurprisingly, there will be both code and writing in the project. In the written report, please include the research purpose and hypothesis. Further, please do not forget to describe the analytical model, report the findings, and cite data sources and previous publications.

Reading weekly papers, or book chapters, is a must, if not noted as "optional." And each student is expected to present one paper during the class. The presentations should be less than 20 minutes.

Assignments

Assignment 1: You are expected to solve assigned problems, which require some programming skills. Group study is totally welcome, but you have to submit YOUR OWN solutions at the end.

Assignment 2: There are two options for this assignment. First, you may use relevant web scraping techniques that you have learnt, in order to build a pandas dataframe, which may be part of your final project in this course, or another significant task of interest. Or, second, you may write your own function for linear regression. There are common linear regression tools of various libraries in python. However, you have to build your own function, exploiting derivation of OLS parameters within matrix notation.

Assignment 3: You are expected to develop an analytical framework through machine learning, including preprocessing, model selection, training and testing, hyperparameter tuning, and presenting a group of outputs. This assignment is designed like a shared task. The higher scores will be awarded. Details for the task will be announced.

Materials

- VanderPlas, Jake. 2016. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/
- Shaw, Zed A. 2017. Learn Python 3 the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code (Zed Shaw's Hard Way Series). 1st Edition. Addison-Wesley. Available at: https://learnpythonthehardway.org/python3/
- The Official Documentation for Python. Available at: https://docs.python.org/3/
- Weekly readings and links for some online sources and Python scripts
- Two more books:

- Angrist, J.D., & Pischke, J.S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion, Princeton: Princeton University Press.
- LeSage, J., & Pace, R. K. (2009). Introduction to Spatial Econometrics. Chapman and Hall/CRC.

Weekly Schedule

1. *Introduction*
    - Install Anaconda Navigator: https://docs.anaconda.com/anaconda/install/
    - Inspect Google Colab: https://colab.research.google.com/?utm_source=scs-index
    - https://github.com/socialcomquant/summer-school-2022/tree/main/Software_Installation_Guidelines

2. *Python basics - 1*
    - Data types, lists, sets, dictionaries, basic operations, if statements, functions, and loops
    - **Reading:** Shaw, Exercises 0-44.
    - **Another option**: The Official Docs Python Tutorial

3. *Python basics – 2*
    a. From lists to datasets and basics of data visualization
        - Numpy and Matplotlib
    b. Problem-solving in Python: Exercises

Assignment 1 (due next week)

4. *Data frames of Pandas - 1*
    - Introduction to Pandas, indexing and selection, operating on data, handling missing data
    - **Reading**: VanderPlas, Chapter 3 (1/2)

- **Another option**: https://pandas.pydata.org/docs/user_guide/index.html#user-guide

5. *Data frames of Pandas - 2*
    - Further data manipulation, merging, grouping
    - **Reading**: VanderPlas, Chapter 3 (2/2)

Midterm report

6. *Automated data collection*
    - Readymade sources (administrational data, surveys)
    - APIs, web scraping
    - How to extract data from a webpage?

    Reading: Jünger, J. (2018): Mapping the Field of Automated Data Collection on the Web. Data Types, Collection Approaches and their Research Logic. In: Stützer , Cathleen / Welker, Martin / Egger, Marc (Hg). Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications. Köln: Halem Verlag, S. 104 130.

7. *Statistics*
    - Basic statistical analyses: from descriptive to inferential
    - **Reading**: The Official Docs Python Tutorial for Statistics. Available at https://docs.python.org/3/library/statistics.html

Assignment 2 (due next week)

8. *Causal inference*
    - How to measure the actual treatment effect?
    - Difference-in-differences, regression discontinuity, instrumental variable analysis
    - Use case: instrumental variable analysis

    Reading: Angrist & Pischke (2009), Chapter 2 – The Experimental Ideal (pp.9-18).

Reading: Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434), 444-455.

Optional reading: Yörük, E., Burak Gürel & M. Fuat Kına. Under review at *American Journal of Sociology*. Demobilization by Substitution: Containing Rural Unrest Through the Bolsa Familia Program in Brazil (pp.27-31 + Appendix, Part B, pp.9-13).

9. *A brief overview of Spatial analysis*

- Spatial econometrics, spatial dependence, spatial weight matrices
- Running a spatial regression
- GeoPandas

Reading: LeSage & Pace (2009), Chapter 1 (pp.1-25).

Optional reading: Yörük, E., Burak Gürel & M. Fuat Kına. Under review at *American Journal of Sociology*. Demobilization by Substitution: Containing Rural Unrest Through the Bolsa Familia Program in Brazil (Appendix, Section A.2, pp.4-8).

10. Bayesian statistics - 1

- Introduction to Bayes
- Running a Bayesian regression model
- Making predictions

Reading: https://www.analyticsvidhya.com/blog/2016/06/bayesian-statistics-beginners-simple-english/

Reading: Van de Schoot, Rens, David Kaplan, Jaap Denissen, Jens B. Asendorpf, Franz J. Neyer, and Marcel AG van Aken. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. Child development, 85(3), 842-860.

11. Bayesian statistics - 2

- Bayesian models
- Use case: Bayesian multilevel regression and post-stratification (MRP)
    - Sampling bias and how to tackle it.

- **Reading:** Park, David K., Andrew Gelman, and Joseph Bafumi. (2004). "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." Political Analysis, 12(4), 375-385.

Chitchat on your final projects

12. Machine learning - 1

- Basics for ML
- Supervised and unsupervised ML models
- Classification, regression, and clustering
- Hyperparameter tuning and evaluation
- **Reading**: VanderPlas, Chapter 5 (1/2)

Reading: Molina, M., & Garip, F. (2019). "Machine learning for sociology". Annual Review of Sociology. Available at: https://www.annualreviews.org/doi/full/10.1146/annurev-soc-073117-041106

13. Machine learning - 2

- K-nearest neighbors, penalized linear regressions (ridge and lasso), naive bayes, support vector machine, random forest, principle component analysis.
- Artificial Neural Networks

Reading: Hindman, M. (2015). "Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences". The ANNALS of the American Academy of Political and Social Science, 659(1), 48–62.

Reading: Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. Available at: https://arxiv.org/pdf/2108.02497.pdf

Assignment 3 (will be submitted two weeks later)

14. Final project presentations

- **For your Github page**: Gandrud, C. (2013). GitHub: A tool for social data set development and verification in the cloud. Available at SSRN 2199367. URL: https://dx.doi.org/10.2139/ssrn.2199367

Additional sources on Python and Data Science

- Ani Adhikari, John DeNero, David Wagner. "Computational and Inferential Thinking: The Foundations of Data Science." 2nd Edition. Available at: https://inferentialthinking.com/chapters/intro.html?utm_source=pocket_mylist
- Chris Bail. "Data Science & Society" Available at: https://dssoc.github.io/schedule/

For textual analysis:

- https://github.com/cltl/python-for-text-analysis/tree/master/Chapters
- https://nlp-css-201-tutorials.github.io/nlp-css-201-tutorials/