# Advanced Data Analysis

M. Fuat Kına

fuatkina@gmail.com

fuat.kina@marmara.edu.tr

Welcome to "Advanced Data Analysis," a comprehensive and demanding course designed to bridge the gap between the rich, dynamic world of social sciences and the transformative power of computing technologies. This course is meticulously structured to introduce you to a range of computational tools and programming fundamentals, specifically tailored for applications in social science research. Starting with the basics of setting up essential software like Anaconda Navigator and Google Colab, the course progresses through the nuances of Python programming, data manipulation using Pandas, and dives into the realms of statistical analysis, and machine learning. Each week, we'll explore critical themes, from web scraping and spatial analysis to cutting-edge topics in machine learning like Artificial Neural Networks. The course is enriched with a carefully curated selection of readings, providing both theoretical backdrops and practical insights into each topic. By the end of this journey, you'll be equipped with the skills and knowledge to harness computational techniques in social science research, opening doors to innovative methodologies and insightful analyses in your future endeavors.

## Grading

- Assignments: %45
    - Assignment 6 will be graded out of 10 points, while all other assignments will be out of 7 points.
- Midterm report: %5
- Final project: %30
    - Class presentation, written report (approximately 2000 words), a python script that you used, and a Github repo which includes all. You are expected to submit only the Github link.
- Attendance: %20

Detailed explanations for **the assignment** are shared in the public GitHub page of the course. For this course, you are expected to create a GitHub repository. Please name your repo as “Python_Course” or something like that. You are responsible to learn how to use and navigate a GitHub repo. Don’t worry! If you don’t have a GitHub account, don’t worry! It is effortless to learn. Infinitely many online sources are ready to guide you. This repo will be your submission place during this course. So, you will add me and the teaching assistant as your collaborators. Then each week you are going to upload your assignments (python scripts, written reports as pdf files, and so on) to this repository with a relevant title, 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!

The final version of **your project** is too expected to be uploaded to your GitHub repo. 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. Make sure that the format details of your

report are as follows: Times New Roman, 12 punto, single-spaced. Always upload pdf files instead of inconsistent word documents.

The possibilities for final projects are quite extensive. I encourage you to select a problem that genuinely intrigues you. This could involve generating data that doesn't currently exist using machine learning models or delving into and explaining a factor within the realm of your research interests. You may tune a regresion model, or a spatial analysis as well. As an illustration, in a similar course years ago, I engaged in spatial regression. In this project, I constructed spatial lags to control for spatial autocorrelation, and demonstrated the bias in the OLS model. While you have the option to employ the techniques covered in the course, feel free to introduce a new approach that can be implemented using Python. The crucial aspect is the utilization of Python in your project. Ultimately, your evaluation will be based on the level of effort you invest but ensure that your efforts are meaningful for your future research and the computational social sciences community.

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*: Programming.

*Assignment 2*: Pandas.

*Assignment 3*: Scraping.

*Assignment 4*: Statistics.

*Assignment 5*: Spatial statistics.

*Assignment 6:* Machine Learning. For this last assignment, 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*
    - Setting up Anaconda Navigator and Google Colab:
        - https://docs.anaconda.com/anaconda/install/
        - https://colab.research.google.com/?utm_source=scs-index

- https://github.com/socialcomquant/summer-school-2022/tree/main/Software_Installation_Guidelines

**PROGRAMMING**

2. *Fundamentals of Python Programming*
   - Data types, lists, sets, dictionaries, basic operations, if statements
   - **Reading:** Lazer, D., Brewer, D., Christakis, N., Fowler, J., & King, G. (2009). "Life in the network: the coming age of computational social science". Science, 323(5915), 721-723. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2745217/
   - **Reading:** Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher Bail. (2020) "Computational Social Science and Sociology." Annual Review of Sociology. URL: https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-121919-054621

3. *Fundamentals of Python Programming*
   - Functions, and loops
   - Introduction to Numpy
   - **Technical reading:** Shaw, Exercises 0-40.
   - **Another option**: The Official Docs Python Tutorial

**Assignment 1**

4. *Python for Data Science: Numpy and Matplotlib*
   - From lists to datasets and basics of data visualization
   - Numpy and Matplotlib
   - Problem-solving in Python: Exercises
   - **Reading**: King, G. (2011). "Ensuring the data-rich future of the social sciences". Science 331 (6018): 719–721. doi:10.1126/science.1197872. URL: https://gking.harvard.edu/files/datarich.pdf
   - **Reading**: Lazer, D. (2015). The rise of the social algorithm. Science, 348(6239), 1090-1091. URL: https://science.sciencemag.org/content/348/6239/1090

**FUNDEMENTALS OF DATA**

5. *Mastering Data Manipulation with Pandas - Part 1*
   - Introduction to Pandas, indexing and selection, operating on data, handling missing data
   - **Reading**: Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076. URL: https://science.sciencemag.org/content/350/6264/1073
   - **Reading:** Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.
   - **Technical reading**: VanderPlas, Chapter 3 (1/2)
   - **Another option**: https://pandas.pydata.org/docs/user_guide/index.html#user-guide

6. *Advanced Data Techniques with Pandas - Part 2*
    - Complex data operations: merging and grouping in Pandas
    - **Reading**: Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, 13. URL: https://www.frontiersin.org/articles/10.3389/fdata.2019.00013/full
    - **Reading:** Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. Public Opinion Quarterly, 85(S1), 399-422.
    - **Technical reading**: VanderPlas, Chapter 3 (2/2)

**Assignment 2**

7. *Exploring Web Scraping*
    - Techniques and tools for extracting Web data
    - **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.
    - **Reading**: "Chapter 10: Social Data APIs" from "Schäfer, M. T., & Van Es, K. F. (2017). The datafied society: Studying culture through data. Amsterdam University Press. URL: https://dataschool.nl/research/the-datafied-society-mirko-tobias-schafer/?lang=en"

**Assignment 3**

## STATISTICS

8. *Introduction to Statistical Analysis in Python*
    - From descriptive to inferential statistics in Python
    - **Reading**: Carmichael, I., & Marron, J. S. (2018). Data science vs. statistics: two cultures?. Japanese Journal of Statistics and Data Science, 1(1), 117-138. URL: https://link.springer.com/article/10.1007/s42081-018-0009-3
    - **Reading**: Weihs, C., & Ickstadt, K. (2018). Data Science: the impact of statistics. International Journal of Data Science and Analytics, 6(3), 189-194. URL: https://link.springer.com/article/10.1007/s41060-018-0102-5

**Midterm report**

9. *Introduction to Statistical Analysis in Python*
    - Derivation of OLS models
    - Introduction to Spatial Analysis
    - **Optional technical reading**: The Official Docs Python Tutorial for Statistics. Available at https://docs.python.org/3/library/statistics.html

**Assignment 4**

10. *Spatial Analysis Fundamentals*
    - Spatial statistics, spatial dependence, spatial weight matrices
    - Running a spatial regression
    - Introduction to GeoPandas
    - **Reading**: LeSage, J., & Pace, R. K. (2009). Chapter 1 (pp.1-25). In "Introduction to Spatial Econometrics." Chapman and Hall/CRC.
    - **Reading**: Anselin, L., Gallo, J. L., & Jayet, H. (2008). Spatial panel econometrics. In The econometrics of panel data: Fundamentals and recent developments in theory and practice (pp. 625-660). Berlin, Heidelberg: Springer Berlin Heidelberg.

**Assignment 5**

## MACHINE LEARNING

*11. Machine Learning Essentials*
- Understanding the basics of Machine Learning and model types
- Supervised and unsupervised Machine Learning models
- **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

*12. Advanced Machine Learning: Model Tuning and Evaluation*
- Model evaluation, and hyperparameter tuning
- Various algorithms:
    - Classification, regression, and clustering
    - K-nearest neighbors, penalized linear regressions (ridge and lasso)
- **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.
- **Technical reading**: VanderPlas, Chapter 5 (1/2)

*13. Artificial Neural Networks*
- Exploring advanced models: Naive Bayes, SVM, Random Forest, PCA
- Artificial Neural Networks
- **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
- **Technical reading**: VanderPlas, Chapter 5 (2/2)

**Assignment 6**

*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

- 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/
- McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc."
- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2016). Big data and social science: A practical guide to methods and tools. Chapman and Hall/CRC.

### Textual analysis

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

### Causal inference

- Akbari, K., Winter, S., & Tomko, M. (2023). Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis*, *55*(1), 56-89.
- Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, *40*(S1), S97-S140.
- Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent Developments in Causal Inference and Machine Learning. *Annual Review of Sociology*, *49*.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.
- Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, *90*(3), 967-1027.
- Chu, Z., Huang, J., Li, R., Chu, W., & Li, S. (2023). Causal effect estimation: Recent advances, challenges, and opportunities. *arXiv preprint arXiv:2302.00848*.
- Farbmacher, H., Huber, M., Lafférs, L., Langen, H., & Spindler, M. (2022). Causal mediation analysis with double machine learning. *The Econometrics Journal*, *25*(2), 277-300.
- Huber, M., & Kueck, J. (2022). Testing the identification of causal effects in observational data. *arXiv preprint arXiv:2203.15890*.
- Li, F., Ding, P., & Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220153.
- Oyenubi, A., & Kollamparambil, U. (2022). Does the child support grant incentivise childbirth in South Africa?. *Economic Analysis and Policy*, *73*, 812-825.
- Rina Friedberg, Julie Tibshirani, Susan Athey & Stefan Wager (2021) Local Linear Forests, Journal of Computational and Graphical Statistics, 30:2, 503-517, DOI: 10.1080/10618600.2020.1831930.
- Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., & Gretton, A. (2020). Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*.

### Big data and social sciences

- Althoff, T., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). "Large-scale physical activity data reveal worldwide activity inequality". Nature, 547(7663), 336. URL: https://www.nature.com/articles/nature23018
- Arabas, S., Bareford, M. R., de Silva, L. R., Gent, I. P., Gorman, B. M., Hajiarabderkani, M., ... & McCreesh, C. (2014). Case studies and challenges in reproducibility in the computational sciences. arXiv preprint arXiv:1408.2123.

- Bail C. et al. (2018). “Exposure to opposing views on social media can increase political polarization”. Proceedings of the National Academy of Sciences. doi:10. 1073/pnas.1804840115.
- Bail C. (2014). “The cultural environment: measuring culture with big data”. Theory and Society 43 (3-4): 465–482. doi:10.1007/s11186-014-9216-5.
- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). “Crowd-sourced text analysis: Reproducible and agile production of political data”. American Political Science Review, 110(2), 278-295.
- Birhane, Abeba. "Algorithmic injustice: a relational ethics approach." Patterns 2, no. 2 (2021): 100205.
- Brayne, S., 2017. Big data surveillance: The case of policing. American sociological review, 82(5), pp.977-1008. URL: https://journals.sagepub.com/doi/full/10.1177/0003122417725865
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In Advances in neural information processing systems (pp. 4349-4357). https://arxiv.org/pdf/1607.06520.pdf or https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf
- Boyd, D., & Crawford, K. (2012). “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon”. Information, communication & society, 15(5), 662-679.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). “Semantics derived automatically from language corpora contain human-like biases”. Science, 356(6334), 183-186. URL: https://science.sciencemag.org/content/356/6334/183
- Carter, D., & Sholler, D. (2015). “Data science on the ground: Hype, criticism, and everyday work”. Journal of the Association for Information Science and Technology, (2013).
- Cioffi-Revilla C.. (2010). “Computational Social Science”. WILEY Interdisciplinary Reviews: Computational Statistics, Vol. 2, No. 3, pp. 259-271. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1708051
- Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., ... & Nowak, A. (2012). “Manifesto of computational social science”. The European Physical Journal Special Topics, 214(1), 325-346.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. Behavior Research Methods, 49(3), 803–821. doi:10.3758/s13428-016-0743-z
- Espeland, W., & Yung, V. (2019). Ethical dimensions of quantification. Social Science Information, 58(2), 238-260. URL: https://doi.org/10.1177/0539018419851045
- Foucault Welles, B., & Meirelles, I. (2015). “Visualizing computational social science: The multiple lives of a complex image”. Science Communication, 37(1), 34-58. URL: https://pdfs.semanticscholar.org/c293/9e49b10cd92bdf8b23bb40d7c662ba66ddbf.pdf
- Grimmer, J. (2015). “We are all social scientists now: How big data, machine learning, and causal inference work together”. PS: Political Science & Politics, 48(1), 80-83.
- Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now?. Big data, 6(3), 176-190. URL: https://www.liebertpub.com/doi/pdf/10.1089/big.2018.0083
- Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. The Annals of the American Academy of Political & Social Science,659, 63-76.
- Howison J., Wiggins A., and Crowston K. (2011). “Validity Issues in the Use of Social Network Analysis with Digital Trace Data”. Journal of the Association for Information Systems 12 (12): 767–797.

- Jungherr A. et al. (2017). “Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support”. Social Science Computer Review 35 (3): 336–356. doi:10.1177/0894439316631043.
- Jungherr A. (2019). “Normalizing Digital Trace Data”. In Digital Discussions: How Big Data Informs Political Communication, ed. by Natalie Jomini Stroud and Shannon C. McGregor, 9–35. New York, NY: Routledge.
- Kang, D., & Evans, J. (2020). Against method: Exploding the boundary between qualitative and quantitative studies of science. Quantitative Science Studies, 930-944. URL: https://www.mitpressjournals.org/doi/pdfplus/10.1162/qss_a_00056
- Kitchin, R. (2014). “Big Data, new epistemologies and paradigm shifts”. Big Data & Society 1 (1): 1–12. doi:10.1177/2053951714528481 URL: https://doi.org/10.1177/2053951714528481
- Landers, R. N., Brusso, R., Cavanaugh, K., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. Psychological Methods, 21(4), 475-492. doi:10.1037/met0000081
- Lazer, D., & Radford, J. (2017). “Data ex machina: introduction to big data”. Annual Review of Sociology, 43, 19-39. doi:10.1146/annurev-soc-060116-053457. URL: https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-060116-053457
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Jebara, T. (2009). “Computational social science”. Science, 323(5915), 721-723. doi:10.1126/science.1167742.
- Leavy, S., O’Sullivan, B., & Siapera, E. (2020) “Data, Power and Bias in Artificial Intelligence”. AI for Social Good Conference. URL: https://aiforgood2020.github.io/papers/AI4SG_paper_81.pdf
- Olhede, S. C., & Wolfe, P. J. (2018). The future of statistics and data science. Statistics & Probability Letters, 136, 46-50. URL: https://doi.org/10.1016/j.spl.2018.02.042
- Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). “Social data: Biases, methodological pitfalls, and ethical boundaries”. Frontiers in Big Data, 2, 13. URL: https://www.frontiersin.org/articles/10.3389/fdata.2019.00013/full
- Radford, J., & Lazer, D. (2019). “Big Data for Sociological Research”. The Wiley Blackwell Companion to Sociology, 417-443. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119429333.ch24
- Raento, M., Oulasvirta, A., & Eagle, N. (2009). “Smartphones. An emerging tool for social scientists”. Sociological Methods & Research, 37(3), 426–454.
- Salganik J. M. and Watts D. J. (2009). “Web-based experiments for the study of collective social dynamics in cultural markets”. Topics in Cognitive Science 1 (3): 439–468. doi:10.1111/j.1756-8765.2009.01030.x.
- Scott A. Golder and Michael W. Macy. (2014). “Digital Footprints: Opportunities and Challenges for Online Social Research”. Annual Review of Sociology 40:129–152. doi:10.1146/annurev-soc-071913-043145.
- Tamburrini, N., Cinnirella, M., Jansen, V. A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. Social Networks, 40, 84-89. doi: 10.1016/j.socnet.2014.07.004
- Taylor, L. (2017). “What is data justice? The case for connecting digital rights and freedoms globally”. Big Data & Society, 4(2), 2053951717736335. URL: https://journals.sagepub.com/doi/full/10.1177/2053951717736335
- Matthew S. Weber (2018) Methods and Approaches to Using Web Archives in Computational Communication Research, Communication Methods and Measures, 12:2-3,

200-215, DOI: 10.1080/19312458.2018.1447657 URL: https://www.tandfonline.com/doi/full/10.1080/19312458.2018.1447657

- Zhang, J., Wang, W., Xia, F., Lin, Y. R., & Tong, H. (2020). Data-driven Computational Social Science: A Survey. Big Data Research, 100145.