National Research University Higher School of Economics
University of London Parallel Degree Program

Undergraduate Program in International Relations

Data Analysis in Python Syllabus
(Winter-Spring 2019/20)

Abstract
In this course students are introduced to the rapidly growing field of data analytics with the specific focus on Python programming language. Students will learn concepts, techniques and tools they need to make meaningful inferences from data. Students will be exposed to a real-world data sets to gain practical skills in data manipulations. Each week will involve seminars and coding simulations. In the final project students will build a working code that can be readily applied for exploratory data analysis in their own (future) research domain.
To prepare for the class students will have to take the "Programming for Everybody (Getting Started with Python" course of the University of Michigan at https://www.edx.org/course/programming-for-everybody-getting-started-with-python

Pre-requisites

- basic courses in math, statistics and probability theory;

Course objectives
To provide a hands-on introduction to Python and its basic applications in the field of data science.

Learning Outcomes

- using Python to carry out basic statistical modeling and analysis.
- mastering basic Python libraries for data science: numpy, scipy, matplotlib;
- application of efficient data wrangling algorithms;
- application of basic tools (plots, graphs, summary statistics) to carry out exploratory data analysis.

Duration
Winter - spring 2019/20 (Modules 3, 4).

Course outline

Week 1. Introduction to the field of data science. Examples of data science approaches applied in economics and political science. Course information on grading, prerequisites and expectations.
Week 2. Introduction to Python. Review of the environment setup process. Anaconda IDE. NumPy, SciPy, Jupyter notebooks.
Week 3. Importing data to Python. Various data sources: text files, web, APIs. Raw and processed data. Working with dates. Pandas library. Merging DataFrames. The principles of tidy data. Data Quality: inaccurate data; sparse data; missing data; insufficient data; imbalanced data.
Week 4. Descriptive statistics. Measures of location: mean, median, mode. Measures of spread: standard deviation, interquartile range, range. Percentiles. Robust statistics. Data transformations.
Week 5. In-class assignment 1.
Week 6. Visualizing data in Python. Matplotlib library. Scatter plot. Line chart. Histogram. Bar chart. Categorical, times series and statistical data graphics.
Week 7. Interactive plots in Python. Introduction to plotly. Finding suitable representation of the data.
Week 8. In-class assignment 2.
Week 9. Basics of probability theory. Distributions, sampling, t-tests. Introductory hypotheses testing and statistical inference.
Week 10. Introduction to linear regression. Estimation techniques. Evaluating the quality of the regression model. Model interpretation.
Week 11. Drawbacks of the linear regression approach. Stability of the coefficients across different parts of the dataset. Rolling estimations.
Week 12. Overfitting. Occam's Razor principle. In-sample and out-of-sample model evaluation. Measuring predictive accuracy of the model.
Week 13. In-class assignment 3.
Week 14. Class wrap-up and discussion of the group project.

Required readings

1. Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".
2. 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.
3. Cady, F. (2017). The data science handbook. John Wiley & Sons.
4. Bowles, M. (2015). Machine learning in Python: essential techniques for predictive analysis. John Wiley & Sons.
5. Attewell, P., Monaghan, D., & Kwong, D. (2015). Data mining for the social sciences: An introduction. Univ of California Press.

Optional readings

6. Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

7. Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine, 66(3), 411-421.
8. Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. Annual Review of Political Science, 20, 529-544.
9. Chang, A., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say'usually not'. Available at SSRN 2669564.
10. Focardi, S. M., & Fabozzi, F. J. (2012). What's wrong with today's economics? The current crisis calls for an approach to economics rooted more on data than on rationality. The Journal of Portfolio Management, 38(3), 104-119.

Grading system

0.5 * ((1/3)*(in-class assignment 1) + (1/3)*(in-class assignment 2) + (1/3)*(in-class assignment 3)) + 0.5*(group project).

In-class assignments (10 pts/each) – week 5, 8, 11

An in-class assignment will be given three times during the course. In-class assignments are problem sets that are to be solved in Python. Each problem set concerns a particular topic. Problem set 1 deals with the basic descriptive statistics, problem set 2 involves working with matplotlib library, problem set 3 is based on linear regression models.

Sample problems:

- Consider the daily oil prices and the USDRUB daily exchange rate. Compute the sample average, standard deviation of daily returns over the entire sample period. Test if mean values are significantly different from zero. Test if mean values significantly differ from each other. State explicitly your null and alternative hypotheses in each case. Plot histograms of the null distributions.
- Consider the daily oil prices and the USDRUB daily exchange rate. Set the length of the rolling window to 50 days ($W = 50$) and compute P-values and R^2 of rolling regressions of USDRUB returns on BRENT returns. Let $\widehat{usdrub}_t$ be your predicted value of USDRUB at timestep t:

  $$\widehat{usdrub}_t = \hat{\alpha}_{t-1} + \hat{\beta}_{t-1} brent_{t-1},$$

  where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of the regression coefficients. Let $usdrub_t$ be the observed value of USDDRUB. The prediction error is then defined as follows:

  $$e_t = \widehat{usdrub}_t - usdrub_t$$

  Plot histogram of prediction errors. Test if the mean prediction error significantly differs from zero. Does your model systematically overestimate or underestimate USDRUB exchange rate?

Group project (10 pts)

Maximum group size: 3 students.

Group project evaluation criteria:

- the purpose of the study is clearly stated (1 point);

- all steps of the research process are described in a clear and concise way (2 points);
- research outcomes are clearly defined (2 points);
- includes intuitive visualizations of research outcomes (2 points);
- all members of the project team are able to explain the code used for computations (1 points);
- code is properly structured (1 point);
- meets submission timeline (1 point).

Methods of instruction

The course is delivered via seminars. Material for this course will be presented using simulations in Python and demonstrations of the source code.

Special Equipment and Software Support

Python, Anaconda IDE

Instructors: Mikhail Vladimirovich Kamrotov (kamrotov@gmail.com)

Office hours: by appointment

Classroom policies:

- Hand-in assignments policy: All home assignments should be submitted electronically via instructor's email on the due date. No deadline extensions are possible.
- Cheating policy: In case of any kind of plagiarism (with the detected source), the assignment is evaluated as zero without the chance to make up for it. In case of two written assignments with the similarity index of 50% and higher from two students, both get a zero for the assignment.