Data Analysis in Python
module 1, academic year 2022–2023

Sergei Golovan
New Economic School
sgolovan@nes.ru

TA: Gennadiy Ivanov (givanov@nes.ru)

Course description

The course "Data analysis in Python" is an introduction to statistical data analysis based on open source software ecosystem of Python. Modern industry is overwhelmed by the amount of data it can collect. At the same time the tools that are used to process, analyze, and visualize the data are expensive and outdated. These days data crunching becomes increasingly the domain of free open source programming languages such as Python. Hence, the goal of this course is to give the students tools to process large amount of data efficiently, summarize it, visualize it, and make informative decisions based on that. Students should learn how to clean imperfectly collected data, how to aggregate it, dissect it, and present the results for efficient communication using state-of-the-art graphing capabilities of Python.

The course is mandatory. It consists of 14 lectures and 7 seminars.

Course requirements, grading, and attendance policies

The course doesn't have any special prerequisites except for the standard probability courses.

There will be 5 home assignments which will constitute 50% of the final grade. The final take-home exam will account for the remaining 50%.

Course contents

1. Code and data organization
   (a) Introduction to version control systems
   (b) Storing and organizing code and data
2. Python basics
   (a) Datatypes: lists, tuples, dictionaries, strings, numbers, booleans
   (b) Comprehensions
   (c) Control flow
   (d) Functions, classes

(e) Reusing code

3. Data processing

   (a) Loading data from the web, csv, Excel
   (b) Data structures: Series, DataFrame, Panel
   (c) Merging several data sets
   (d) Indexing and selecting data
   (e) Computational tools for data modification
   (f) Grouping and aggregating data
   (g) Reshaping data
   (h) Time and date functionality

4. Data visualization

   (a) Creating simple plots: line, scatter, bar
   (b) Plotting several data sources
   (c) Fine tuning aesthetics
   (d) Additional libraries for interactive exploration
   (e) Distributing visualizations

Description of course methodology

Lectures will proceed from motivating examples and sample models in economics to general principles of statistical modeling and data visualization. Also, a number of computer exercises will be distributed in order to give students an opportunity to practice the data analysis techniques.

Sample tasks for course evaluation

1. Source of the data: https://www.seattle.gov/transportation/projects-and-programs/programs/bike-program/bike-share

   Data file link: https://s3.amazonaws.com/pronto-data/open_data_year_two.zip

   Example of analyzing this dataset: https://jakevdp.github.io/blog/2015/10/17/analyzing-pronto-cycleshare-data-with-python-and-pandas/

   (a) Setup the environment.
   (b) Show contents of the zipfile `cycle_share.zip`.
   (c) Import `2016_trip_data.csv` directly from the zipfile into Pandas DataFrame. Make sure that *starttime* and *stoptime* are of `datetime64[ns]` type, while *usertype* and *gender* are categorical. Print variable types. Print first five rows of the following columns: *trip_id*, *Date*, *starttime*, *tripduration*, *gender*. Print value counts for the two categorical variables in the dataset. The data is saved in `exam_data.hdf`, table *trips*.

(d) Import `2016_weather_data.csv` directly from the zipfile into Pandas DataFrame. Make sure that *Date* is of `datetime64[ns]` type, while *Events* is categorical. Print variable types. Print first five rows of the following columns: *trip_id*, *Date*, *starttime*, *tripduration*, *gender*. Note that the values of *Events* column can be, for example, 'Fog-Rain' and 'Fog , Rain'. Rename the latter to the former. Print value counts for the *Events* column.

(e) Find the most popular station pairs. Note that a trip from A to B is counted the same as from B to A. Also exclude trips that start and finish on the same station. Save the result to an Excel file. Print the names of the two most popular destinations from the list.

(f) Draw density of log trip duration.

(g) Draw density of log trip duration for men and women separately. Exclude value 'Other' from the column *gender*.

(h) Count the number of trips for each date and for each usertype. Draw two time series (for each usertype) on the same plot. Since both are highly seasonal, resample them to weekly frequency.

(i) Compute the average trip duration for each precipitation type (*Events* column in the weather dataset).

(j) Is there any dependence, separated by gender, between log trip duration and two weather variables: average temperature in Celcius and log precipitation? Draw two scatter plots and corresponding regression lines for each explanatory variable and separated by gender. Exclude value 'Other' from the column *gender*. *Note*: $T_F = 1.8 \cdot T_C + 32$.

Course materials

1. Stanislav Khrapov (2016) *Data analysis in Python: lecture notes*, https://dataanalysispython.readthedocs.io/en/latest/
2. Scipy Lecture Notes, http://www.scipy-lectures.org/
3. Mark Pilgrim (2011) *Dive into Python 3*, Apress, https://www.diveintopython3.net/
4. Pandas (Python data analysis library), https://pandas.pydata.org/
5. Matplotlib (plotting library), https://matplotlib.org/
6. Bokeh (interactive plotting), https://bokeh.pydata.org/
7. Plotly (interactive plotting), https://plot.ly/

Academic integrity policy

Cheating, plagiarism, and any other violations of academic ethics at NES are not tolerated.