

Syllabus

Introduction to Python for Data Science

(4 ECTS)

Iuliia Naidenova, lecturer (e-mail: naidenovayn@gmail.com, web-page: <https://www.hse.ru/en/org/persons/14538835>)

School of Economics and Finance

Meeting Minute # ___ dated _____ 20_

1. Course Description

a) Pre-requisites

English

Statistics

Econometrics

Familiarity with other programming language (such as R) is advantageous.

b) Abstract

The course provides students with wide general overview of Python – a general-purpose programming language that is becoming ever more popular for data science. The focus is on the application of Python specifically for data science. The course is about ways to import, store and manipulate data, and helpful data science tools to conducting data analyses. The course is intended for students with little programming background. The learning process is facilitated with DataCamp platform.

2. Learning Objectives

The main objective of the course is to provide students with the basic concepts of Python, its syntax, functions and packages to enable them to write scripts for data manipulation and analysis. The course develops skills of writing and running a code using Python. The course covers various variables types and their features, basic operators and statements, loops, as well as the main packages for data science: NumPy, Pandas, Matplotlib.

At the end of the course, students should be able to write short scripts to import, prepare and analyze data.

3. Learning Outcomes

By the end of the course, a student develops the following competencies, where

GK – general competence

GPK – general professional competence

Code	Competencies
------	--------------

GK-3	Is able to solve problems in professional activities based on analysis and synthesis
GK-5	Is able to work with information: find, evaluate and use information from various sources, necessary to solve scientific and professional problems (including on the basis of a systematic approach)
GPK-3	Is able to analyze socially significant problems and processes and predict their possible development in the future
GPK-11	Is able to collect, analyze and process statistical data, information, scientific and analytical materials needed to solve economic problems
GPK-12	Is able to choose tools for processing economic data in accordance with the task, analyze the results of calculations and substantiate the findings
GPK-17	Is able to use modern hardware/software and information technologies for solving analytical and research problems.

4. Course Plan

№	Topics	All hours	Lecture hours	Seminar hours	Self-study hours
Topic 1. Introduction to Python for Data Science					
1	Introduction to Python for Data Science	30	-	-	20
Topic 2. Intermediate Python for Data Science					
2	Data types and operators	20	-	-	20
3	Cleaning and merging data	20			20
Topic 3. Python DataFrames					
4	Analysis, selection, and visualization techniques with Pandas DataFrames	20	-	-	20
5	Extracting and transforming DataFrames	20	-	-	20
6	Merging DataFrames	20			20
Topic 4. Importing Data in Python					
7	Importing Data in Python	20			20
Topic 5. Environment for scientific programming in Python					
8	Jupyter Notebook	12	2		10
Sum		152	2	0	150

Topic 1. Introduction to Python for Data Science

1. Introduction to Python for Data Science.

An introduction to the basic concepts of Python. Learn how to use Python interactively and by using a script. Create variables and acquaint with Python's basic data types. Learn to store, access, and manipulate data in lists. Learn how to use functions, methods, and packages. NumPy is a fundamental Python package to efficiently practice data science. Learn to work with powerful tools in the NumPy array, and get started with data exploration.

<https://www.datacamp.com/courses/intro-to-python-for-data-science>

Topic 2. Intermediate Python for Data Science

2. Data types and operators.

Building of various types of plots, and customizing them to be visually appealing and interpretable. Learn about the dictionary, an alternative to the Python list, and the pandas DataFrame. Creating and manipulating datasets, access the information from these data

structures. Boolean logic is the foundation of decision-making in Python programs. Learn about different comparison operators, how to combine them with Boolean operators, and how to use the Boolean outcomes in control structures. Filtering data in pandas DataFrames using logic. While and for loops.

<https://www.datacamp.com/courses/intermediate-python-for-data-science>

3. Cleaning and merging data.

Cleaning and merging data, diagnosing issues such as outliers, missing values, and duplicate rows.

<https://www.datacamp.com/courses/cleaning-data-in-python>

Topic 3. Python DataFrames

4. Analysis, selection, and visualization techniques with Pandas DataFrames.

<https://www.datacamp.com/courses/pandas-foundations>

5. Extracting and transforming DataFrames, advanced indexing, rearranging and reshaping data.

<https://www.datacamp.com/courses/manipulating-dataframes-with-pandas>

6. Merging DataFrames with pandas

<https://www.datacamp.com/courses/merging-dataframes-with-pandas>

Topic 4. Importing Data in Python

7. Importing Data in Python. The ways to import data into Python: from flat files such as .txt and .csv; from files native to other software such as Excel spreadsheets, Stata, SAS, and MATLAB files; and from relational databases such as SQLite and PostgreSQL.

<https://www.datacamp.com/courses/importing-data-in-python-part-1>

Topic 5. Environment for scientific programming in Python

8. Jupyter Notebook as an environment for scientific programming in Python, its structure and features.

5. Reading List

a) Required

Vanderplas JT. Python Data Science Handbook : Essential Tools for Working with Data. First edition. Sebastopol, CA: O'Reilly Media; 2016. Available from: search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1425081

b) Optional

Thomas, Seemon. Basic Statistics, Alpha Science Internation, 2014. ProQuest Ebook Central. Available from: <https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=5190782>

6. Grading System

The teacher assesses a self-study students' work as fulfilment of tasks in DataCamp. Each fulfillment task before deadline gives 2 point, each fulfillment task after deadline gives 1 point. The whole grade (O_{DC}) is calculated as a percentage of maximum possible grade and then converted to 10 point scale.

And the resulting course grade is calculated as follows:

$$O_{result} = 0,5 * O_{DC} + 0,5 * O_{exam}$$

with O_{exam} – a grade for the exam on a 10 point scale, O_{DC} – a grade for fulfilment of tasks in DataCamp (10 point scale).

7. Examination Type

Final student assessment (exam) is a project that is performed in a team of no more than 2 people. Each team uses provided dataset and apply one or a combination of the learnt methods of data analysis in Python. As a result of the project each team write down the report and prepare working file. The grade for the exam includes the grade for the report, grade for the working file and the grade for answering questions.

Example of exam tasks:

1. Import the file “data_for_exam.xlsx”
2. Describe the variable X1 – its type, number of observations, descriptive statistics.
3. Create a dataframe named df1 from the dataset
4. Create a new dataframe named df2 with public companies only
5. Create a plot to show the relationship between company performance and of other variables
6. Combine dataframe df1 with another dataset df3 from the file “more_data.xlsx” by the index. Drop all observations with missings in the new variables. Create a plot using one or more variables from the df3. Interpret the plot.

8. Methods of Instruction

At the start of the course, students receive the schedule for DataCamp courses fulfillment. At the end of the course, the lecture that combines the material of the course and a discussion of particular aspects take place.

9. Special Equipment and Software Support (if required)

Software that is essential for the course is:

- Access to DataCamp
- Anaconda
- Microsoft office
- Chrome or any other web browser.