Advanced Data Analysis in Python

Koç University

Fall 2021

Syllabus

Instructor	**Class Schedule**
David Carlson	Tu-Th 14.30 – 15.40
dcarlson@ku.edu.tr	**Office Hours**
Office: CASE 140	Tu-Th 13.00 – 14.00 (or by appointment)
TA	**Online Access**
Oğuz Bülüt Kök	https://ku.blackboard.com (for grades and readings)
okok20@ku.edu.tr	https://github.com/carlson9/KocPythonFall2021 (for in-class material)

Introduction

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 apply to programming more generally. These include syntax, data types, functions, loops, recursion, and classes and inheritance. We will then cover data base management, creation, manipulation, and visualization. An overview of maximum likelihood estimation and frequentist statistics will be followed by introductions to the most common machine learning methods. This is a demanding course, with the ultimate goal a final project with an original analysis testing one or several hypotheses.

No previous programming experience is assumed. However, a good understanding of linear models is required. Undergraduate prerequisite: MATH 202 (or equivalent); graduate prerequisite: INTL 601 (or equivalent).

Suggested Books

This course contains suggested readings and exercises from two books. Electronic versions of both of these books can be purchased, but can be read for free online. I do not necessarily recommend purchasing them, and I do not recommend purchasing hard versions. As you learn to code, you should also learn how to use online resources that are freely available. Further, you will need to follow along with the readings while coding in Python, making online and soft versions more sensible. In addition to the book readings, assigned articles will be posted on Blackboard, and links for online material are available in the course schedule. The two books, and the associated websites, are:

1) 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/

2) VanderPlas, Jake. 2016. *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/

Requirements and Grading

Grades will not be rounded, these represent strict cut-offs. In the rare event of, for example, exactly a 90, the higher grade will be assigned. Pluses and minuses will be applied at the instructor's discretion and will only be used if there are clear separations within a given grade. **Note that the Koç suggested grades are not followed in this course.**

A	90–100
B	80–90
C	70–80
D	60–70
F	<60

1) *Homework: 50%*

   Both graded and ungraded homework assignments will be assigned throughout the semester. The ungraded assignments will not be checked, but it is essential to complete them for success in the course and to adequately learn the material. Four graded assignments will be checked. They will be posted on Blackboard at least one week before they are due. It is strongly encouraged that you start the homeworks as early as possible. As we will discuss in the first week, all work must be done on git. Working in collaboration with one another is encouraged, but every keystroke must be your own. That is, you can work together, but you must complete your own assignment for grading and review. Late work will not be accepted. The work must be completed on git before the class meets for the week of the due date.

2) *Final project: 50%*

   The final project can be done in groups of two to four students or individually, but the grading will reflect this. More will be expected from groups than from individuals. The final project involves both coding and writing of results. To be more precise, the final project turned in should include a report appropriate for your field or work, and all code, well-organized and commented, for replication. The report should at a minimum explain the hypothesis/hypotheses, discuss in detail the data that is analyzed, discuss in detail the method(s) used and why, and the findings. You will not be in any way down-graded if you find a null result or do not find support for your hypothesis/hypotheses. Exploratory work without any clear hypothesis is generally discouraged, but if it is relevant to your field it is acceptable. Other creative applications are fine, but you should check with the instructor. There are benchmarks to be completed throughout the semester, as reflected in the course schedule. The first benchmark is outlining your hypothesis/hypotheses, expectations, or exploratory goal. This also serves as an opportunity for me to approve or disapprove of a project, give suggestions for improvement, etc. The second benchmark is a detailed data report on the data to be used in the project. The third is a brief description of the modeling choice with a brief justification. The final project must expand on this brief description for readers unfamiliar with the discussed methods. This should include a full model specification, and an intuitive explanation. You are expected to briefly present your work at the end of the semester, but the presentation is not graded.

Disclaimer on Class Recordings

The synchronous sessions are recorded (audiovisual recordings). The students are not required to keep their cameras on during class.

The audiovisual recordings, presentations, readings and any other works offered as the course materials aim to support remote and online learning. They are only for the personal use of the students. Further use of course materials other than the personal and educational purposes as defined in this disclaimer, such as making copies, reproductions, replications, submission and sharing on different platforms including the digital ones or commercial usages are strictly prohibited and illegal.

The persons violating the above-mentioned prohibitions can be subject to the administrative, civil, and criminal sanctions under the Law on Higher Education Nr. 2547, the By-Law on Disciplinary Matters of Higher Education Students, the Law on Intellectual Property Nr. 5846, the Criminal Law Nr. 5237, the Law on Obligations Nr. 6098, and any other relevant legislation.

The academic expressions, views, and discussions in the course materials including the audiovisual recordings fall within the scope of the freedom of science and art.

Academic Integrity at Koç University

Koç University expects all its students to perform course-related activities in accordance with the rules set forth in the Student Code of Conduct (`http://vpaa.ku.edu.tr/academic/student-code-of-conduct`). Actions considered as academic dishonesty at Koç University include but are not limited to cheating, plagiarism, collusion, and impersonating. This statements goal is to draw attention to cheating and plagiarism related actions deemed unacceptable within the context of Student Code of Conduct:

All individual assignments must be completed by the student himself/herself, and all team assignments must be completed by the members of the team, without the aid of other individuals. If a team member does not contribute to the written documents or participate in the activities of the team, his/her name should not appear on the work submitted for evaluation.

Plagiarism is defined as "borrowing or using someone else's written statements or ideas without giving written acknowledgement to the author." Students are encouraged to conduct research beyond the course material, but they must not use any documents pre-

pared by current or previous students, or notes prepared by instructors at Koç University or other universities without properly citing the source. Furthermore, students are expected to adhere to the Classroom Code of Conduct (`http://vpaa.ku.edu.tr/academic/classroom-code-of-conduct`) and to refrain from all forms of unacceptable behavior during lectures. Failure to adhere to expected behavior may result in disciplinary action.

There are two kinds of plagiarism: Intentional and accidental. Intentional plagiarism (Example: Using a classmate's homework as one's own because the student does not want to spend time working on that homework) is considered intellectual theft, and there is no need to emphasize the wrongfulness of this act. Accidental plagiarism, on the other hand, may be considered as a 'more acceptable' form of plagiarism by some students, which is certainly not how it is perceived by the University administration and faculty. The student is responsible for properly citing a source if he/she is making use of another person's work.

If you are unsure whether the action you will take would be a violation of Koç University's Student Code of Conduct, please consult with your instructor before taking that action.

Course Schedule

Please note this schedule is subject to change.

Week 1: Anaconda set-up, git, shell introduction, basic Python syntax, data types

Readings:

1. Read https://git-scm.com/docs/user-manual.html up until the section "Exploring Git history"

2. Shaw, Appendix

3. Shaw, Exercises 0 – 15.

Ungraded homework:

1. Install Anaconda https://docs.anaconda.com/anaconda/install/

2. Sign up for a free GitHub account https://github.com/

3. Install git https://git-scm.com/downloads

4. Create a public repository called PythonCourse, and add me (carlson9) as a collaborator

Week 2: Functions, loops, recursion, classes

Readings: Shaw, Exercises 16 – 44.

Week 3: Reading from and writing to files, web scraping and APIs

Graded homework 1 due
Readings:

1. VanderPlas, Preface and Chapter 1.

Week 4: Introduction to NumPy

Readings: VanderPlas, Chapter 2.

Week 5: Data manipulation with Pandas

Hypothesis, expectations, or goals due
Readings: VanderPlas, Chapter 3.

Week 6: Visualization with Matplotlib

Graded homework 2 due
Readings: VanderPlas, Chapter 4.

Week 7: Ordinary Least Squares and MLE

Readings: "A Gentle Introduction to Maximum Likelihood Estimation," available at: https://towardsdatas gentle-introduction-to-maximum-likelihood-estimation-9fbff27ea12f

Week 8: GLM I

Data report due

Readings:

1. "Chapter9: The General Linear Model (GLM): A gentle introduction," available at: http://psych.colorado.edu/ carey/qmin/qminChapters/QMIN09-GLMIntro.pdf

2. "Generalized Linear Models," available at: https://www.statsmodels.org/stable/glm.html

Week 9: GLM II

Graded homework 3 due

Readings:

1. "Five Extensions of the General Linear Model," available at: https://www.theanalysisisfactor.com/exte general-linear-model/ (read all the nested links as well).

2. "4.3 GLM, GAM and more," available at: https://christophm.github.io/interpretable-ml-book/extend-lm.html

Week 10: Machine learning I

Readings: VanderPlas, Chapter 5 (first half, to page 445).

Week 11: Machine learning II

Brief model justification due

Readings: VanderPlas, Chapter 5 (second half).

Week 12: Neural networks

Graded homework 4 due

Readings:

1. http://neuralnetworksanddeeplearning.com/chap1.html

2. https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/

Week 13: Final project presentations

Week 14: Final project presentations

Final projects due at end of week