Note: This is a draft syllabus, but will give you an idea of the topics covered.

DS-UA 0111: Data Science for Everyone

Credits: 4 credits

Course description and impact

Data Science for Everyone is a foundational course that prepares students to participate in the data-driven world that we are all experiencing. It assumes only that students have a sound knowledge of high-school Algebra. It develops an intuition about underlying statistical theories such as inference, and develops students’ programming skills in Python so that students can write programs to summarize and compare real-world datasets. Building on these data analysis skills, students will learn how draw conclusions about populations from random samples of those populations. The main aim of the course is for students to integrate data science theories, methods, and techniques into disciplinary questions raised with the humanities, social sciences, and sciences. As part of a proper liberal arts foundation, we will also explore ethical, legal, and privacy issues that arise when implementing data science methods.

The course brings together several trends. One is that we increasingly live in a digital world, and students in many areas of study can benefit from having hands-on knowledge of how computers work. That knowledge is best gained by programming in a general purpose programming language. Hence the course teaches basic programming.

The second trend is that data are increasingly available and that effective conclusions about the insights that can be drawn from data are increasingly important in many fields. This course will teach theory undrepinning inferential procedures accessible to students with no prior experience with statistics or college-level math.

The course is structured into two parts. In the first, students learn theoretical principles behind programming and some computer science topics, and develop practical skill in coding. The other part is computational inference, in which students learn about statistical concepts and techniques and how to program computers to make predictions. Students will apply this knowledge to projects that touch on liberal arts subject matter across the disciplines.

Some departments will decide to offer liberal arts courses that are intended to be taken concurrently with this course. These “connector courses” can discuss the implication of computing and inference in a specific domain, provide domain-specific examples, and extend the techniques taught to also include techniques of specific interest in the domain.

This course would be offered every semester. The only prerequisite is high-school Algebra.
Learning objectives

After successfully completing the course, students are able to:

● Learn the basics of computer programming theory.
● Understand statistical theory to be able to conduct basic data analyses and learn relevant techniques to parse and clean data.

Learn how to form research questions based in the humanities, social sciences, and/or sciences, as they relate to data science and computational statistics. Students will compare multiple datasets on similar subjects and determine differences. They will synthesize information and form conclusions.

● Design and test basic statistic hypotheses using p-values. For example, determine how significant the difference in means are in two populations.
● Learn relevant modeling approaches and techniques. Conduct computational models according to theoretical, disciplinary questions.
● Predict discrete target variables using clustering techniques.

Office hours

The instructor will be available two hours a week for meeting with students.

[Specifics added here for finalized syllabus]

Topics covered

The course is structured into a sequence of lectures and accompanying assignments.

The assignment consists of labs and homeworks.

● Labs (10 in total) are short exercises done in class and submitted in class.
● Homeworks (8 in total) are longer exercises designed to take a week. Some of these build on previous homeworks.

The labs are held in-class on the regular classroom schedule.

In addition to the instructor, support for students is provided by Teaching Assistants and Tutors. Once the course has been offered, both of these roles will be filled by students who did well in the class or by data science majors. Tutoring involves meeting with students to help them with the material. Small group sessions with four students will be scheduled and students will be able to sign up for a time that meets their needs. One on one sessions will also be available. Teaching Assistants will provide in-class assistance and grading and guidance to students on homework and projects.
The class will have a computing infrastructure. Students will learn how to program in Python using Jupyter notebooks. Student and instructor notebooks will be hosted on a cloud provided internally to NYU. The use of a cloud avoids students having to install software on their personal devices and provides an access point to automate some of the grading.

The textbook, *Computational and Inferential Thinking: The Foundations of Data Science* is provided online (and free) at: https://www.inferentialthinking.com/chapters/intro. The material is technical and some students will find it challenging. Hence the reading assignments are shorter than the guidelines. The textbook is provided on-line as web pages, so we report below our estimate of the equivalent page numbers is a printed book.

There are no stand-alone essay-type assignments. Instead, writing is embedded in the labs and homeworks. During the labs, students write computer programs. The computer programs are complex to write, as often each part must be cohesive with the others. During the homeworks, students also write computer programs, and in addition, they provide short essays on the interpretation of the data and the implications for the data around some decision or problem that the data inform. The writing length would typically be two to three pages for these essays.

Below we provide a week by week schedule. This schedule does not reflect academic holidays that typically occur during a semester and would need to be adjusted to reflect those holidays in an actual offering of the course. The entries under a week correspond to the lectures contain “TOPIC; SECTIONS IN TEXTBOOK TO READ; ASSIGNMENTS IF ANY”.

1. Week 1
   a. Introduction; Reading is 1, 1.1, 1.2, 1.3 in the textbook (about equivalent to 12 pages); Lab 01: Expressions
   b. Cause and effect; 2 (about 11 pages); Homework 01
   c. Tables; 3 (about 31 pages)
2. Week 2
   a. Data types; 4 (about 8 pages), 5 (about 11 pages); Lab 02: Types & Sequences
   b. Building tables; 6.1, 6.2 (about 18 pages); Homework 02
   c. Charts; census; 7, 7.1 (about 13 pages), 6.3, 6.4 (about 9 pages); Lab 03: Arrays & Tables
3. Week 3
   a. Census (continued); 6.3, 6.4 (about 9 pages); Homework 03
   b. Histograms; 7.2, 7.3 (about 15 pages)
   c. Functions; 8, 8.1 (about 13 pages); Lab 04: Histograms & Functions
4. Week 4
   a. Groups; 8.2, 8.3 (about 12 pages); Homework 04
   b. Joins: 8.4 (about 3 pages)
   c. Table examples; 8.5 (about 11 pages); Lab on Project 1
5. Week 5
6. Week 6
   a. Models; 10.3 (about 5 pages); Homework 06 Part a
   b. Comparing distributions; 11, 11.1, 11.2 (about 17 pages)
   c. Decisions and uncertainty; 11.3 (about 8 pages); Lab 06: Assessing Models

7. Week 7
   a. A/B testing: 12, 12.1 (about 9 pages); Homework 06 (World Progress)
   b. Causality; 12.3 (about 10 pages)
   c. Examples; 12.2 (about 8 pages); Lab: Midterm Review

8. Week 8
   a. Midterm exam
   b. Confidence intervals; 13, 13.1, 13.2 (about 16 pages)
   c. Interpreting confidence; 13.3, 13.4 (about 15 pages); Lab 07: Bootstrap

9. Week 9
   a. Center and spread; 14, 14.1, 14.2 (about 15 pages); Homework 07 (Diet and Disease)
   b. The Normal Distribution; 14.3, 14.4 (about 10 pages)
   c. Sample means; 14.5 (about 5 pages)

10. Week 10
    a. Designing experiments; 14.6 (about 4 pages);
    b. Correlation; 15, 15.1 (about 14 pages)
    c. Linear regression; 15.2 (about 10 pages); Lab 08: Correlation

11. Week 11
    a. Least squares; 15.3, 15.4 (about 12 pages); Homework 08 (Movie Classification)
    b. Residuals: 15.5, 15.6 (about 10 pages)
    c. Regression inference: 16 (about 15 pages); Lab 09: Regression

12. Week 12
    a. Classification; 17, 17.1, 17.2, 17.3 (about 17 pages); Homework 08 Part b
       Classifying Movies)
    b. Classifiers: 17.4, 17.5, 17.6 (about 21 pages); Lab: Homework support
    c. Decisions: 18 (about 10 pages)

13. Week 13
    a. Privacy and ethics: readings to be determined when an instance of the course is
designed, but covering contemporary developments such developments as the
GDPR, Facebook data sharing practices, how bias creeps into predictive models,
and other issues; reading are probably from the web
    b. Case Studies on privacy and ethics, at discretion of instructor; intended to be on
contemporary issues in the news and prepared by the Teaching Assistant; Lab
session
    c. Review: Lab 10: Decisions

14. Week 14: at discretion of instructor; possibly catch up on syllabus or new topics
15. Week 15: at discretion of instructor; possibly catch up on syllabus or new topics

Course assessment

All assignments (labs and homeworks) must entirely be the student's own submissions. Any sharing or copying of assignments is considered cheating and will result in an F in the course. A second cheating incident will, by CAS rules, result in a one-semester suspension from the College.

Students accumulate up to 100 points during the course.
- Up to 10 points for completing labs
- Up to 12 points for completing homeworks on time
- Up to 32 points for the midterm
- Up to 46 points for the final.

Grades will be determined using this scale:

<table>
<thead>
<tr>
<th>Grade in Course</th>
<th>Points Earned</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>94 - 100</td>
</tr>
<tr>
<td>A-</td>
<td>90 - 93</td>
</tr>
<tr>
<td>B+</td>
<td>87 - 89</td>
</tr>
<tr>
<td>B</td>
<td>84 - 86</td>
</tr>
<tr>
<td>B-</td>
<td>80 - 83</td>
</tr>
<tr>
<td>C+</td>
<td>76 - 79</td>
</tr>
<tr>
<td>C</td>
<td>72 - 75</td>
</tr>
<tr>
<td>C-</td>
<td>70 - 71</td>
</tr>
<tr>
<td>D+</td>
<td>66 - 69</td>
</tr>
<tr>
<td>D</td>
<td>62 - 65</td>
</tr>
<tr>
<td>D-</td>
<td>60 - 61</td>
</tr>
<tr>
<td>F</td>
<td>Less than 60</td>
</tr>
</tbody>
</table>
Moses statement

Disability Disclosure Statement: Academic accommodations are available for students with disabilities. The Moses Center website is www.nyu.edu.csd. Please contact the Moses Center for Students with Disabilities (212-998-4980 or mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.