**![A black sign with white text

Description automatically generated]()**

**Fall 2020 Syllabus**

**Python for Data Science Bootcamp (PDSB)**

**Prof. Luis E. Morales**

**Title: Executive in Residence**

**Affiliation: Pratt School of Engineering**

**Contact:** [**luis.morales@duke**](mailto:luis.morales@duke)**.edu, 919-522-0081**

**Course Description**

This course is designed to help students build proficiency in the use of Python for Data Science. It covers foundational concepts and it also provides hands on experience with critical skills including loading, cleaning, manipulating, visualizing, analyzing and interpreting data. Upon completion of this course, students will be in a position to use their skills to identify, formulate and solve some practical data analysis problems. There are two entry points into this course, the first one is for students with little or no Python knowledge. At this level the course is open to everyone, and there are no specific prerequisites. The second one is for students that meet a minimum proficiency standard and don’t need to learn basic Python programming.

For students starting at the first entry point, the first two weeks will provide an introduction to basic Python. Topics covered during the first two weeks include: variables, expressions, statements, functions, loops and iterations, strings, lists, dictionaries, tuples, lambda functions and regular expressions. At the start of week 3 (second entry point), all students will come together to complete four additional weeks of instruction. For 3 weeks, the class will focus on building proficiency with three critical for Data Science Python libraries (Numpy, Pandas, Matplotlib). Finally, the last week will focus on a review of probability and statistics with an emphasis on simulation of chance experiments. Topics for this last module will include discrete distributions, continuous density functions, combinatorics, conditional probabilities and a final review of important densities and functions.

**What background knowledge do I need before taking this course?**

At entry point one, the course is open to everyone. For those looking to start at entry point 2, a basic understanding of Python programming is required that includes the following pre-requisites:

* Completion of online course “Python for everybody, Exploring Data Using Python 3”, Charles R. Severance, free open book at <https://www.py4e.com/> or equivalent\*.
* Experience creating and using Jupyter notebooks within an Anaconda Navigator

\*Specific Python experience must include practitioner level mastery of the following concepts: Variables, Expressions, and Statements (Chapter 2 – PY4E), Conditional Execution (Chapter 3 – PY4E), Functions (Chapter 4 – PY4E), Loops and Iteration (Chapter 5 – PY4E), Strings (Chapter 6 – PY4E), Lists (Chapter 8 – PY4E), Dictionaries (Chapter 9 – PY4E), Files (Chapter 7 – PY4E), Tuples (Chapter 10 – PY4E), Lambda function, Format method & Print function, Introduction to Regular Expressions (Chapter 11 – PY4E), Findall, Search, Split and Sub functions

**What will I learn in this course?**

By the end of this course, you will be able to:

* Manage fundamental python programming techniques
* Employ frequently used libraries, such as numpy, pandas, matplotlib in your own codes
* Design project topics and collect valid data for that topic
* Apply data analysis methods on real data including loading, cleaning, manipulating, visualizing and running basic statistical analysis
* Evaluate whether the selected methods and experimental results are reasonable
* Conduct simulations of chance experiments with discrete or continuous outcomes

**What will I do in this course?**

Activities:

* + Lecture Sessions:
    - Delivery: Asynchronous online (including notes and video)
    - Total: 23; 4 per week (3 on week 6)
    - Each lecture is on average 35 minutes long
  + Checkpoint Sessions:
    - Delivery: Synchronous, recorded
    - Total: 6; 1 per week (Thu)
    - Each session is 30 mins long
  + Assessments:
    - Assignments
      * **Total:** 6; 1 per week
      * **Released and collected:** Released Monday due Friday
      * **Time to complete and submit:** 5 days
    - Mini-Projects
      * **Total:** 3; 1 every other week
      * **Released and collected:** Released Monday due following Monday

**How will my performance in this course be assessed?**

Formative Assessments:

* Assignments are used to evaluate ability to apply concepts and techniques covered in class.
* Mini projects will evaluate whether students can program python codes to load data, clean data, manipulate data, visualize data, and perform basic statistical analysis in an incremental stage by stage manner.

**How can I prepare for the class sessions to be successful?**

* You are required to use your own laptop for programming activities after setting up your development environment.
* You are expected to complete selected textbook readings before watching lecture videos and notebooks.
* You are expected to review the posted class videos online and work through the posted class notebooks before corresponding checkpoint sessions
* You are expected to attend (or review the recording of) checkpoint sessions on Thursday
* You are expected to track due dates and milestones associated with assessments (assignments, mini-project) to ensure your work is submitted on time.

**What required texts, materials, and equipment will I need?**

Required Textbooks:

1. Python for everybody, Exploring Data Using Python 3, Charles R. Severance, free open book at <https://www.py4e.com/> (Entry level 1 students only)
2. Python Data Science Handbook: Essential Tools for Working with Data, by Jake VanderPlas, O'Reilly Media; 1 edition (December 10, 2016), ISBN-13: 978-1491912058, full text and code freely available at https://jakevdp.github.io/PythonDataScienceHandbook/.
3. Introduction to Probability by Charles M. Grinstead and J. Laurie Snell, free open book at www.dartmouth.edu/~chance/teaching\_aids/books\_articles/probability\_book/book.html

Required free software: Python 3, Anaconda, Jupyter Notebook

**How will my grade be determined?**

This is a pass/fail course. If an assessment (assignment, mini-project) fails, student is expected to resubmit.

**What is the schedule of topics by week?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Timepage6image3177232** | **Class topic/unit name** | **Pre-class work for students** | **Planned activitiespage6image3165584** | **Assessments Released** |
| **Week 1** | Course introduction, Basic Python programming I | Donoho D. 50 Years of Data Science. Journal of Computational & Graphical Statistics, 2017, 26(4): 745- 766 https://www.py4e.com/Chapters 1-6 | lectures, checkpoint | Assignment 1 |
| **Week 2** | Basic Python programming II, | https://www.py4e.com/Chapters 7-11 | lectures, checkpoint | Assignment 2 Mini-project 1 |
| **Week 3** | Numpy, Load and manipulate data using the panda's library | PDSH reading chapters 2 and 3 (up to section 3.04) | lectures, checkpoint | Assignment 3 |
| **Week 4** | Load and manipulate data using the panda's library (cont.) | PDSH reading chapters 3 (sections 3.05, 3.06, 3.07, 3.08 and 3.10) | lectures, checkpoint | Assignment 4 Mini-project 2 |
| **Week 5** | Plotting, charting & data visualization using the matplotlib library; Probability and statistics | PDSH reading chapter 4 (sections 4.01, 4.02, 4.05, 4.14); Introduction to Probability, chapter 1 | lectures, checkpoint | Assignment 5 |
| **Week 6** | Probability and statistics | Introduction to Probability, chapters 2-5 | lectures, checkpoint | Assignment 6 Mini-project 3 |

**What is the schedule of topics by lecture?**

**Week 1 (PY4E – Chapters 1-6)**

* + Lecture 1 - Class Introduction, Why Program? (Chapter 1 – PY4E)
  + Lecture 2 - Variables, Expressions, and Statements (Chapter 2 – PY4E), Conditional Execution (Chapter 3 – PY4E)
  + Lecture 3 - Functions (Chapter 4 – PY4E), Loops and Iteration (Chapter 5 – PY4E)
  + Lecture 4 - Strings (Chapter 6 – PY4E), ,

**Week 2 (PY4E – Chapters 7-11)**

* Lecture 5 - Lists (Chapter 8 – PY4E), Dictionaries (Chapter 9 – PY4E)
* Lecture 6 - Files (Chapter 7 – PY4E), Tuples (Chapter 10 – PY4E)
* Lecture 7 - Lambda function, Format method & Print
* Lecture 8 - Introduction to Regular Expressions (Chapter 11 – PY4E), Findall, Search, Split and Sub functions

**Week 3 (PDSH, Chapters 2 and 3, up to section 3.02**):

* Lecture 9 - Introduction to Numpy, Understanding Data Types in Python, The basics of NumPy Arrays
* Lecture 10 - Computation on Arrays – Ufuncs, Aggregates, Broadcasting
* Lecture 11 – Boolean Arrays and Masks, Fancy Indexing, Sorting
* Lecture 12 – Introduction to Pandas, Panda Objects, Data Indexing and Selection

**Week 4 (PDSH, Chapter 3, sections 3. 4, 5, 6, 7, 8, 10)**

* Lecture 13: Operating on Pandas, Handling Missing Data
* Lecture 14: Hierarchical Indexing, Combining Datasets: Concat and Append
* Lecture 15: Combining Datasets: Merge and Join
* Lecture 16: Aggregation and Grouping, Vectorized String Operations

**Week 5 (PDSH, Chapter 4 – sections 1, 2, 5, 14; Introduction to Probability, Chapter 1)**

* Lecture 17: Introduction to Matplotlib, Simple Line Plots, Simple Scatter Plots
* Lecture 18: Histograms and Binnings, Text and Annotation
* Lecture 19: Visualization with Seaborn
* Lecture 20: Simulation of discrete probabilities, chance experiments with a finite number of possible outcomes. Dice rolling, coin flip

**Week 6 – (Introduction to Probability, Chapters 2 – 5)**

* Lecture 21: Continuous Probability Densities, chance experiments with a continuum of possible outcomes. MonteCarlo experiment: estimate quantities whose exact values are difficult or impossible to calculate exactly. Using simulation to estimate areas of plane figures;
* Lecture 22: Combinatorics, counting the number of ways that a particular event can occur. Combinations: Counting, Tree diagrams, Permutations: Factorials, Binomial coefficients, card exercises, Bernoulli trials; Conditional Probability
* Lecture 23:Important Distributions and Densities: Discrete Uniform Distribution, Binomial distribution, Geometric Distribution, Poisson distribution, Hypergeometric Distribution, Continuous Uniform Density

**What material do I need to cover before each assessment? When are they due?**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week 1 (8/10/20-8/14/20)** | | | | | |
| **Assignments** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released |  |  |  |  | 8/10/20 |
| Due |  |  |  |  | 8/14/20 |
| Identifier |  |  |  |  | A1 |
| US Time |  |  |  |  | 9 am |
| Topics Included |  |  |  |  | Up to Lecture #4 |
| **Week 2 (8/17/20-8/21/20)** | | | | | |
| **Assignments** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released |  |  |  |  | Monday 8/17/20 |
| Due |  |  |  |  | 8/21/20 |
| Identifier |  |  |  |  | A2 |
| US Time |  |  |  |  | 9 am |
| Topics Included |  |  |  |  | Up to Lecture #8 |
| **Week 3 (8/24/20-8/28/20)** | | | | | |
| **Assignments** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released |  |  |  |  | Monday 8/24/20 |
| Due |  |  |  |  | 8/28/20 |
| Identifier |  |  |  |  | A3 |
| US Time |  |  |  |  | 9 am |
| Topics Included |  |  |  |  | Up to Lecture #12 |
| **Mini-Project** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released | Monday 8/17/20 |  |  |  |  |
| Due | 8/24/20 |  |  |  |  |
| Identifier | MP1 |  |  |  |  |
| US Time | 9 am |  |  |  |  |
| Topics Included | Up to Lecture #8 |  |  |  |  |
| **Week 4 (8/31/20-9/4/20)** | | | | | |
| **Assignments** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released |  |  |  |  | Monday 8/30/20 |
| Due |  |  |  |  | 9/4/20 |
| Identifier |  |  |  |  | A4 |
| US Time |  |  |  |  | 9 am |
| Topics Included |  |  |  |  | Up to Lecture #16 |
| **Week 5 (9/7/20-9/11/20)** | | | | | |
| **Assignments** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released |  |  |  |  | Monday 9/7/20 |
| Due |  |  |  |  | 9/11/20 |
| Identifier |  |  |  |  | A5 |
| US Time |  |  |  |  | 9 am |
| Topics Included |  |  |  |  | Up to Lecture #20 |
| **Mini-Project** | | | | | |
| Released | Monday 8/31/20 |  |  |  |  |
| Due | 9/7/20 |  |  |  |  |
| Identifier | MP2 |  |  |  |  |
| US Time | 9 am |  |  |  |  |
| Topics Included | Up to Lecture #16 |  |  |  |  |
| **Week 6 (9/14/20-9/18/20)** | | | | | |
| **Assignments** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released |  |  |  |  | Monday 9/14/20 |
| Due |  |  |  |  | 9/18/20 |
| Identifier |  |  |  |  | A6 |
| US Time |  |  |  |  | 9 am |
| Topics Included |  |  |  |  | Up to Lecture #23 |
| **Week 7 (9/20/20)** | | | | | |
| **Mini-Project** | | | | | |
|  | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
| Released | Monday 9/14/20 |  |  |  |  |
| Due | 9/21/20 |  |  |  |  |
| Identifier | MP3 |  |  |  |  |
| US Time | 9 am |  |  |  |  |
| Topics Included | Up to Lecture #23 |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| Python for Data Science Bootcamp | | |
| Prof. Luis E. Morales | | |
| Quick Reference To Online Lectures | | |
| **Description** | **Duration (normal speed)** | **Link to Zoom Video** |
| Lecture #1 | 22 minutes | [PDSB-Lecture#1](https://duke.zoom.us/rec/share/5NBxL-7qpnFLGLfN-XzxX5QlMtrPT6a8hycZ_vIOzU68yrm2lMbgSnXtzLl9dBKf?startTime=1593029544000) |
| Lecture #2 | 1 hour 4 minutes | [PDSB-Lecture#2](https://duke.zoom.us/rec/share/os8vFZer6EROconhshvQVrRxQN6-T6a803Ub8vIEyxmOKgc1Wum-xTLuKMN3EawH) |
| Lecture #3 | 32 minutes | [PDSB-Lecture#3](https://duke.zoom.us/rec/share/vfIqf7jO90RJZdLEq2PfaoJiMY7ueaa80yRMqfsNnU_15_ZWLjRbwH-aOoQrVDE8?startTime=1592505936000) |
| Lecture #4 | 28 minutes | [PDSB-Lecture#4](https://duke.zoom.us/rec/share/1dFEAfLw12pJe5GOylvAB4QKEqC1aaa80CEb_vdZykimaZi95PxVCzTdLm9Zhl6w?startTime=1592508850000) |
| Sync Session #1 |  |  |
| Lecture #5 | 29 minutes | [PDSB-Lecture#5](https://duke.zoom.us/rec/share/yMp6c-3zrDhLeKvo0XHtR4orG57fT6a8gCFK-vpfn0ugF-Olg9LRcAuloX2ZJ7DG?startTime=1592511619000) |
| Lecture #6 | 33 minutes | [PDSB-Lecture#6](https://duke.zoom.us/rec/share/5-FoFJ395DlLTqf0-k7xXq4eF6S1X6a81SIbq6EMxUeiv9b0EsfpwY7Q4weIUbrs?startTime=1592577518000) |
| Lecture #7 | 26 minutes | [PDSB-Lecture#7](https://duke.zoom.us/rec/share/_ZJMNI7A1npJYoHIzUucaKQFOIrDaaa82yhP_fBYzkZkM1WujDK5TwULHik1cmig?startTime=1592580156000) |
| Lecture #8 | 15 minutes | [PDSB-Lecture#8](https://duke.zoom.us/rec/share/7OksLO7O2SRLSM_nyx3yZOkrEJ7beaa82iYZr6UMyE8NgvRVFqp5Ftks0-2rw9uY?startTime=1592585158000) |
| Sync Session #2 |  |  |
| Lecture #9 | 30 minutes | [PDSB-Lecture#9](https://duke.zoom.us/rec/share/wsglDpbhxjJJQM_tr1rUS4w-F5raaaa81ykZq6VbyBmhlvGqNVL6V_1EYaHht4-l?startTime=1592588576000) |
| Lecture #10 | 48 minutes | [PDSB-Lecture#10](https://duke.zoom.us/rec/share/3-5OFLiurjxOYKvJ5HzZC5MgJp_sX6a8hCAc-_RYmU59rJiHgYuRJdIRtVMTGzKD?startTime=1592679340000) |
| Lecture #11 | 33 minutes | [PDSB-Lecture#11](https://duke.zoom.us/rec/share/-_wvNJ_I-jhLeKf2ykPkB-k9JdnoT6a823AbrqUMzEtkWxzS-Em91BBDINAyt873?startTime=1592683979000) |
| Lecture #12 | 45 minutes | [PDSB-Lecture#12](https://duke.zoom.us/rec/share/usJxKIj-32BJWJ3qzV72RfM_B5Tqaaa80CYb-_sFy0-p0u_6EUhDAPt9xZXHQ3O8?startTime=1592831346000) |
| Sync Session #3 |  |  |
| Lecture #13 | 47 minutes | [PDSB-Lecture#13](https://duke.zoom.us/rec/share/zu5HHoqu02VJeI3v9RzuYfU4HrrEX6a81yJMqKUNnhoNz44QUVwS2PT4ReEJFTWc?startTime=1592837389000) |
| Lecture #14 | 43 minutes | [PDSB-Lecture#14](https://duke.zoom.us/rec/share/_slHKY3970hIYs_M0lP9SosPL4_-T6a81HNN-fQMzk9nQb2Vqt9gPv7O4oCsuujv?startTime=1592843637000) |
| Lecture #15 | 26 minutes | [PDSB-Lecture#15](https://duke.zoom.us/rec/share/28l8Ab7q7UZLXpXyzUXOao84Et_Leaa80ShL86BczU0SCc5ZHm8ZBx3Gp_5iq0hj?startTime=1592852413000) |
| Lecture #16 | 33 minutes | [PDSB-Lecture#16](https://duke.zoom.us/rec/share/wtQ2Da36sUpLSZXpxU7afPF4Lr7Kaaa81SMa_vtcnxuhVUgLlHpYmxUBCG7ytriI?startTime=1592857008000) |
| Sync Session #4 |  |  |
| Lecture #17 | 34 minutes | [PDSB-Lecture#17](https://duke.zoom.us/rec/share/yOVIHqq38FlIZ9LXuWuYA-lmHrW0eaa8hCAf-fcLmR4LiwF5751T4UuOJ49FJPa4?startTime=1592918590000) |
| Lecture #18 | 18 minutes | [PDSB-Lecture#18](https://duke.zoom.us/rec/share/zI97Ie_zx3FJa53h0Uria64bRJS5X6a803cfrPIPmEfijMO_ir1RCkvD6x3UgL-w?startTime=1592923228000) |
| Lecture #19 | 16 minutes | [PDSB-Lecture#19](https://duke.zoom.us/rec/share/vv0kIuCrpyBLGbOUyWLwZPEGL5zpX6a803QeqPAOxB5lYM5fMYM4juu-MRzveICm?startTime=1592928360000) |
| Lecture #20 | 44 minutes | [PDSB-Lecture#20](https://duke.zoom.us/rec/share/otEyD-ji_DtOebf1xFryQasdIKXsaaa80CEXrvcNyRxpBQjbsE7Dz9YjZxq6uN--?startTime=1592930657000) |
| Sync Session #5 |  |  |
| Lecture #21 | 1 hour 5 minutes | [PDSB-Lecture#21](https://duke.zoom.us/rec/share/x-ovfq_B6jxORLPo1FzZV55-IbvuX6a81CYW-PoExRkJ1KCOEHJ4CMV43C4PNO60?startTime=1593008952000) |
| Lecture #22 | 44 minutes | [PDSB-Lecture#22](https://duke.zoom.us/rec/share/45xfC-DA8T9LZ4Xz9UP4SoECLqS9X6a8g3dKq_cJmEkdPIW17LeCpRIPPOINtBIW?startTime=1593018616000) |
| Lecture #23 | 37 minutes | [PDSB-Lecture#23](https://duke.zoom.us/rec/share/ucl3Iq7g2WZOZbfQuHj7CqU6GYHKT6a8hCAd-KULzAHex9a2EEDbWAkrk7pEZUU?startTime=1593024997000) |
| Sync Session #6 |  |  |