Winter 12-2019

Python Foundations: Data Science for All

Leslie J. Albert

Esperanza Huerta

Scott Jensen

Python Foundations: Data Science for All

Tutorial

Abstract

The Python Foundations seminar emphasizes a hands-on approach and is designed to reach a diverse student population to provide them with the entry-level programming skills required in data science. The creation of the seminar materials has been funded by a three-year federal grant and has followed a systematic approach to enhance the student learning experience and raise student awareness of data science. The seminar materials—structured in three parts: pre-seminar, live seminar, and post-seminar—are licensed under the Creative Commons Attribution-ShareAlike 4.0 International license to encourage adoption by any institution. To foster interest in the seminar, students have the possibility to earn a digital badge attesting to their newly acquired skills.

Keywords

Tutorial, Python, Data Analytics, Data Science.

Introduction

Funded by a three-year federal grant[1], Python Foundations is one of nine seminars developed to introduce data science topics to undergraduate students from any academic background. The seminars are designed to be offered as extracurricular activities in a stress-free environment in which students can explore data science topics. The tutorials are intended to pique the students' curiosity about data science and boost their confidence in their analytic skills. Each seminar, including Python Foundations, is designed to be a stand-alone seminar that does not require students to have any prior programming or data science knowledge.

The grant funded the creation of a series of seminars to increase and diversify the pool of undergraduate and community college students who are aware of data science and have skills to gather, wrangle, and cleanse data. Each seminar is structured in three parts: pre-seminar, live seminar, and post-seminar. Pre-seminar materials provide a gentle introduction to the topic, encouraging students (but not requiring them) to explore the topic before the live seminar. The live seminar provides hands-on activities in which students can play around with real data, guided by an instructor. The post-seminar provides assignments for students to complete on their own to earn a digital badge attesting to their new skills (Hurst 2015; Reeves et al. 2017).

The creation of the seminar materials, including the Python Foundations seminar, has followed a methodology to ensure the materials are useful in developing the desired data science skills. The topics of the seminars were selected with the guidance of an Advisory Board to ensure the relevance of the topic for data science. The development of the seminar materials is planned in three phases: 1) initial development to create a beta version in year one, 2) refinement of the materials in year two based on the feedback from year one (the current phase), and 3) evaluation of the seminar materials by an independent evaluator in year three (to be conducted in 2020).

Although the grant is currently in year two, and the seminar materials will be modified before their final version, the beta version (version 0.1) is available and shared under the Creative Commons Attribution-Share Alike 4.0 International license to encourage the adoption of the materials by any institution. The teaching materials available for download include: slides, Python scripts (or Jupyter notebooks), exercises, teaching notes, datasets, Canvas course packages, test materials, and guidelines for setting up digital badges.

[1] Details about the grant and web addresses have been omitted, as required, to preserve the anonymity of the review and will be provided, if accepted, in the final version of the paper.

Tutorial development

The topics for the tutorials were selected by an Advisory Board created to guide the efforts of the faculty awarded the grant. During the first two years of the grant, the Advisory Board has included four to six members, who are highly recognized for their expertise in data science and their commitment to training undergraduate students. Given that data science relies heavily on programming, the creation of an introductory seminar in programming was deemed necessary. Python was selected as the language of choice because of its relevance in data science. Although there are some resources available online to learn Python, these resources are mainly focused on explaining Python syntax, rather than on explaining programming concepts transferable to other programming languages. In addition, current free online resources do not provide support material for faculty wanting to introduce Python to students who have no programming experience.

The seminars, including Python Foundations, are organized in three parts. The pre-seminar includes materials that provide general information about the seminar and introduces the topic, including instructions for students to install the programming environment in their own computers. The live-seminar material includes slides for a 2:45 minutes session conducted in a computer lab, along with programs and data sets for hands-on exercises. The post-seminar includes additional programs and data sets for additional practice, and a final test that must be passed to earn a digital badge.

The seminar was designed to allow students to attend the live seminar without working on the pre- and post-seminar materials and still gain from the experience in a stress-free environment. However, to encourage students to further develop their skills, digital badges are awarded to students who work on the post-seminar exercises and are able to pass a test demonstrating their skills. The seminar materials are organized in a student bundle (includes readings, Python programs, data sets and a practice test) and an instructor bundle (includes teaching notes, solutions to exercises and testing materials). The entire instructor bundle is also accessible as a Canvas module, so it can be imported directly to Canvas or other learning management system.

The creation of the materials was planned in three phases. During the first phase a beta version (0.1) of the material was created. The Python Foundations seminar was offered twice during Spring 2019 to a total of 55 business students, and once during Fall 2019 to a total of 69 students, all from a large public American University. In the second phase, the beta version is being modified based on the feedback obtained from the students and the instructor's observation of the students' understanding of the material. The revised version (version 1) of the materials will be used in the next phase-two during Spring 2020, to be offered to undergraduate and community college students from any academic background. The third and final phase will be conducted during Fall 2020 in which an independent evaluator will assess the materials and provide feedback to improve the materials. The material will be revised based on this feedback for a final release (version 2) in 2021.

Although the seminar materials are expected to improve as a result of these phases, the beta version is available to be used by any interested instructor or institution. We hope that instructors adopting the seminar materials will provide us with feedback from their experience using the materials. We are planning to incorporate their suggestions in upcoming versions of the materials. Considering that the amount of feedback we can receive from instructors depends on the amount of exposure of the seminar materials, a critical aspect of this phase is to disseminate widely the existence of this material among data science instructors.

Challenges in the development of the tutorials

Although the faculty involved in the seminars are experienced instructors, creating the seminar materials posed several challenges. First, each seminar is required to be a stand-alone seminar that does not require students to have any prior programming or data science knowledge. Second, each seminar needs to be approachable to students from any academic background. That is, examples and data should not be tied to a particular academic discipline, but should be understandable and interesting to millennials and gen Z students from any discipline. Third, each seminar must provide all the support instructors need to use the seminar materials as a plug-in application. That is, it should contain all the teaching notes, exercises, and tests necessary to offer the seminar without additional effort. Fourth, each seminar is required to provide

material accessible at no cost to students and instructors. We briefly discuss these challenges and the efforts made to solve them in the Python Foundations seminar.

Stand-alone seminar

Creating stand-alone seminars was required to minimize the commitment from students attending the seminars and the instructors delivering the seminars. Students should be able to chose any number of seminars without being concerned with prerequisites or necessary background knowledge. One of the goals of the seminars is to raise awareness of data science on students who are not in disciplines in which these topics are covered. For this reason, the seminars do not require students to have any prior programming or data science experience; all the necessary knowledge necessary to understand the topic of the seminar should be included within the seminar itself. As a result, the Python Foundations seminars first provides a brief introduction of how computers work, emphasizing the four elements computer have (input devices, processing unit, output devices, and memory) and distinguishing between input devices and input information (and output devices and output information). The seminar also distinguishes between word processors, text editors, and integrated development environments. Because the seminar is geared towards students who have no programming background, a special emphasis is given to make clear concepts commonly confused by novice programmers, such as distinguishing numbers and digits.

General interest

The seminars are required to appeal a broad variety of students, regardless of their academic background. For this reason, the hands-on and follow up activities of the seminar were required to be understandable and appealing to millennials and gen z students. The Python programs were designed to exclude specialized formulas and only include basic arithmetic operations and data comparisons not tied to a specific discipline. A key concern for millennials and gen z is climate change and life style, including origins of food and debt management (Francis and Hoefel 2018). To appeal to these generations, the examples in the Python Foundation seminar deal with electric cars, interest calculation, and avocado price estimation. The programs about electric cars and avocado price include real data.

Complete solution

The seminars have been designed to serve as a "one-stop" answer to the question of how to bring a diverse student population up to speed on data science concepts, techniques, and tools. While no single answer could address all possible aspects of data science, the one provided by these seminars covers most of the important ones. Instructors who makes use of the seminars will not need to replicate the investigation of curricular needs, nor will they have to reinvent the solution.

Each seminar provides everything instructors need to utilize its materials. An instructor can use the teaching notes, exercises, and tests included with the seminar to offer the course without having to recreate them. This reduces non-productive work for the instructor, freeing her to focus on the specific teaching demands that will arise with each group of students.

Free distribution

A high priority for this project has been that all materials included would be free to use by students and instructors. A problem instructors and students face with many of the materials currently available to them is that such resources are copyrighted. Sharing or reuse of copyrighted original materials, such as those found in libraries worldwide, is frequently restricted. This limits the immediate use of such items, making it more difficult for the potential teachers and learners to apply them.

As stated above, the nine seminars and their materials are released under the Creative Commons Attribution-ShareAlike 4.0 International license. This license gives users the right to "copy and redistribute the material in any medium or format," and "remix, transform, and build upon the material for any purpose, even commercially" (Creative Commons undated).

None of the nine of the seminars in this project include copyrighted materials. Instead they incorporate summaries of original readings, and citations of the original sources. The student bundle of materials is

freely available; anyone can access it. The instructor bundle, on the other hand, is only available to certified instructors. Certification to access the instructor bundle involves demonstration by teachers of their instructor status at an educational institution. Once such demonstration has occurred, instructors are able to access the instructor bundle. The restrictions imposed on the instructor bundle are due to a need to limit access to testing materials.

Tutorial materials

After students complete the Python Foundations seminar, they are able to explain the role of programming in data science, and can interpret, modify, and create basic programs in Python. The Python Foundation seminar is planned to take approximately 6 hours to complete. The pre-seminar material includes a note on programming in data science, the summary of an article on the history of Python, and a technical note to install Python. Depending on how fast students read, the pre-seminar is designed to last between 20 and 30 minutes. The live seminar material includes slides, narratives and solutions to programs using data about electric cars, and narratives to programs to calculate interests. The live seminar is designed to last 2:45 hours. The post-seminar material includes solutions to the narratives to calculate interests, narratives, data, and solutions to estimate avocado prices, a guide to additional resources, a practice test, and a final test. Table 1 lists the materials available in the beta version (version 0.1) of the instructor bundle for the Python Foundations seminar.

Type of resource	Name of resource
Pre-seminar readings	Pre-seminar work (html page and word file) Note 1: Programming in data science (html page and word file) Article 1 – pre-seminar (html page and word file) Summary of article "And now for something different" (html page and word file) Note 2: Installing Python (html page and word file) Installing Python on your computer (word and pdf file)
Live-seminar readings	Live seminar (html page and word file) Python foundations slides (pdf file for students, annotated PowerPoint for instructors) Electric cars (html page and word file) Narratives to electric cars (wprd and pdf files) Interests (html page and word file) Interest narratives (word and pdf file)
Live seminar Python programs	electric_1.py, electric_2a.py, electric_2b.py, electric_3a_py, electric_3b.py, electric_3c.py, electric_4.py, electric_5.py, electric_6.py
Live seminar data sets	electric_data.txt notes_amount.txt

Reading post-seminar	Post-seminar work (html page and word file) Solutions to the narratives to calculate interests (html page and word file) Narratives and data to estimate avocado prices (html page and word file) Avocado prices narratives (word and pdf files) Solutions to the narratives to estimate avocado prices (html page and word file) Other resources (html page and word file)
Python programs	Interests files: interestV1.py, interestV2.py, interestV3.py, interestV4.py, interestV5.py, interestV6.py Avocado prices files: avocado_1.py, avocado_2, avocado_3.py, avocado_4.py, avocado_5.py, avocado_6.py
Datasets	Avocado.txt
Test banks	16 test banks with a total of 80 questions

Table 1. List of materials available in the instructor bundle

Figure 1 shows the narrative of a basic program for a post-seminar activity. Considering that the seminar is designed for students with no programming experience, the narratives explicitly indicate the input and output data, and provide examples of the execution of the program. All the narratives include solutions.

The program compares the price of an avocado to the average annual price. The user inputs the price of the avocado and the average annual price through the keyboard. The program reports on the screen "above" when the price is above the average annual price and "on or below" when the price is on or below the average annual price.

Input: price (keyboard), average annual price (keyboard).
Output: status relative to average annual price (screen).

For instance, when the user inputs an avocado price of 1.3 and an average annual price of 2.0, the execution of the program should be similar to the image below.

```
Avocado price? 1.3

Average price? 2.0
The avocado price is on or below the average
```

When the user inputs an avocado price of 2.3 and an average annual price of 2.0, the execution of the program should be similar to the image below.

```
Avocado price? 2.3

Average price? 2.0
The avocado price is above the average
```

When the user inputs an avocado price of 2.0 and an average annual price of 2.0, the execution of the program should be similar to the image below.

```
Avocado price? 2.0

Average price? 2.0
The avocado price is on or below the average
```

Figure 1. Example of a basic narrative to create a Python program

Future material releases

The materials currently available for the Python Foundations seminar are in beta version (0.1). Version 1 is expected to be released in Spring 2020. Version 1 will update the materials based on student feedback and instructor observations. An independent evaluator will assess the materials and provide additional feedback. Version 2, to be released in Spring 2021 will incorporate the suggestions from the independent evaluators. For versions 1 and 2, it is critical to obtain feedback from instructors adopting the materials. Instructors from any institution are encouraged to adopt the material and share their experiences with the creators of the seminars.

Conclusion

As the world faces a growing shortage of data scientists (National Academies of Sciences 2017), expanding the number of college students interested in data science-related fields is critical. Currently however, few community colleges or undergraduate programs provide training in data science techniques to a broad student population. Python Foundations, one of nine seminars developed as part of a federally-funded grant, offers a novel approach to addressing the need for greater data science proficiencies in today's workforce by exposing students to data science programming concepts in a safe and approachable format. This extracurricular seminar focuses on valuable, transferable skills yet requires no prior knowledge or experience in programming and thus appeals to students from a wide range of educational backgrounds. Students interested in building further on the knowledge they gained during the seminar may complete pre and post-seminar activities and assessments to earn a digital badge. Student materials may be freely accessed online through the project's website. Instructional support materials for faculty interested in adopting the seminars, in part or in whole, are also freely available following validation of faculty employment.

References

Creative Commons. undated. "Attribution-Sharealike 4.0 International." 2019, from https://creativecommons.org/licenses/by-sa/4.0/

Francis, T., and Hoefel, F. 2018. "'True Gen': Generation Z and Its Implications for Companies." *McKinsey Quarterly* Retrieved September 1, 2019, from https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/true-gen-generation-z-and-its-implications-for-companies

Hurst, E. J. 2015. "Digital Badges: Beyond Learning Incentives," *Journal of Electronic Resources in Medical Libraries* (12:3), pp. 182-189.

National Academies of Sciences, E., and Medicine. 2017. "Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report."

Reeves, T. D., Tawfik, A. A., Msilu, F., and Şimşek, I. 2017. "What's in It for Me? Incentives, Learning, and Completion in Massive Open Online Courses," *Journal of Research on Technology in Education* (49:3/4), pp. 245-259.