Data People: Learn Python!

Roxanne Johnson
@roxLjohnson
www.roxjohnson.com
john3718@umn.edu

View this document online: http://bit.ly/1XCi2GJ

0 Build on what you already know

What tools do you currently use to work with data?

People who work with data have an advantage when learning programming. Proficiency or familiarity with other data tools can provide a conceptual foundation you can build on instead of starting at the very beginning.

The path laid out in orange is a sample of what my journey could have looked like…

Excel/ Spreadsheets	Using formulas in Excel can be similar to programming. You can even create named functions in Excel if you wanted to. Using Pivot tables and other features can help you identify analyses you may want to try in Python. You may even identify tasks that are difficult to do in Excel that may inspire you to try Python!
Databases and queries (SQL)	Structured Query Language (SQL) is a programming language, but is in a different programming paradigm so the style is different, which might be confusing. It may be useful to read up on how the languages are structured differently.
Web-based tool with a graphic Interface (Tableau)	Drag-and-drop menus can be a great way to identify the steps and order of an analysis. Once you can visualize what exactly you want to do with your data, you can translate that into Python to see how it could work.
Math/statistical software (SAS, SPSS, STATA, MATLAB)	If you use a domain-specific programming language, you may be able to take an analysis you've already done (or look at someone else's) in the language you know and try it using Python to see how it could work.
GIS and online mapping tools (MapBox, CartoDB)	Similar to the drag-and-drop menu idea, spatial analysis software allows you to visualize what you want to do with your data. Desktop GIS allows you to add Python to automate tasks. Additionally, GIS software builds a database to store geospatial data.
Other Programming language	If you already know another programming language, this tool may not be that useful to you! You are welcome to use it anyway :)

1 Find content of resource by the kind of data you have

What kind of data do you have?

The kind of data, the domain you work in, and the file format you have will affect what kinds of analysis you might want or be able to do. Identifying some keywords to use in searches will help you find the right tools.

Data Keywords	Content Keywords
Tabular	Financial
Time Series	Economic
Categorical	Education
Survey	Manufacturing
Geospatial	Membership
Statistics	Company
Relational Database	
Inventory	

Why This Tool?

I have been a research analyst for about five years, and actively trying to learn Python (and R) for data analysis for about two years. I've found it extremely challenging and I've met many others with similar experiences. Some of the challenges I have identified:

- **Not knowing where to start or what the big picture looks like; no clear view of what's possible or a map of how to get there**
- **The vast number of resources is daunting; it is difficult to find resources that are both at my skill level and relevant to what I want to learn**
- **Not knowing what I don't know, lacking the awareness that I don't understand a core concept, and not being able to articulate questions or search terms**
- **Many resources are intended for people who are or want to become developers or professional programmers, which isn't me**
- **I don't feel like I fit anywhere- other researchers may not know about programming, people who regularly work with data in Python sometimes Excel-shame me, and programmers tell me that learning Python is easy!**

2 Find content of resource by analysis task

What task(s) do you want to do with data?

There are several tasks you may want to do with your data, and this may impact what kinds of tools you want to learn about using the resources you find.

Some Basics

There are a few things you will have to learn to start working with data using Python:

- Conceptual challenge: if you are used to opening a spreadsheet to view your data, accessing it with Python may seem unintuitive.
- Downloading and installing Python. Version 2 or 3? Anaconda distribution?
- Opening and running Python, Integrated Development Environments (IDEs), Jupyter Notebooks
- Navigating file systems, paths, and the command line
- Basic syntax plus built-in data structures: lists, dictionaries, arrays
- Reading in data from a file, writing to a file
- How to identify, obtain, load libraries that are good for the data task you want
- General library use: how do they work? (dot notation)

Gather, Obtain, Collect, Scrape it

Learn about:

- Find existing data in a usable format, open data
- File formats, accessibility, machine readability
- Identify potential issues with readability and format.
- Collecting your own data and storing it in a usable format: data collection methods (survey) and coding (efficiency), survey design
- Data structures
- Scraping data from websites: Regular Expressions

Useful Tools:
Libraries: csv, and csvkit, pandas, NumPy, Beautiful Soup (HTML/XML)
Collecting, storing, using data: Data Maturity Framework: University of Chicago Center for Data Science & Public Policy

Munge, Clean, Wrangle, Prepare, Format it

Learn about:

- Data hygiene/cleanliness. What potential problems exist? Why are they problems? How do you deal with them?
- Formatting/re-formatting data. Why would you want to? What are the benefits of one format over another? Using a tool like Tableau requires that your data be formatted a certain way. You could use Python to format it.
- Many people use Python to prep their data for analysis in another language or with another tool

Useful tools:
Libraries: pandas, csvkit
Concept/skill: Regular Expressions

Analyze, Explore, Look for Patterns, Play with it

Learn about:

- Types of analysis. Statistical, summarizing, grouping, calculating new fields
- Exploratory vs. explanatory analysis
- Modeling: predictive models and machine learning

Useful tools:
Libraries: pandas, NumPy, SciPy, matplotlib (plotting), csvkit (work with csv files), Seaborn (statistical visualization), pprint (pretty print), scikit-Learn (machine learning)
If you use pivot tables in Excel, try groupby in pandas
Kaggle competition: shows a specific data analysis and predictive modeling in Python, R, and Excel

Visualize it with Charts, Graphs, or Maps

Learn about:

- What types of charts are good for the data you have and story you want to tell?
- What data structures are best for what you have and what you want to do?

Useful tools:
Python libraries: pandas data frame, plotting with matplotlib, Seaborn (statistics), Bokeh (interactive visualization on the web)
Mapping: GIS (ArcGIS, open source QGIS)
Choosing a data visualization to present findings: Flowing Data, Storytelling with Data

3 Find format of resource by project phase

Which phase(s) of the data analysis project are you working on?

There are many types of resources, and some are better than others for the phase you're working on in your project. Some resources are very broad, others very specific. Some help you learn big concepts and ideas, others very specific use of a tool. Identifying where you are in a project can help you look for the most useful resource.

Idea and Exploration
Thinking up a data analysis project: what kind of question you might want to answer or explore, looking for what data is out there, thinking about what tools you have access to use. In this phase you want broad resources that will introduce you to what's possible and exciting.

Scoping and Planning
Come up with a good question or problem statement. You want to make sure it's doable. Do you have the data? Do you already know the tool fairly well? How much time do you have? In this phase you want to narrow down what you can do from what you want to do; resources should help you set an obtainable goal. May require some exploratory analysis.

Implementation and Troubleshooting
Actually doing the data analysis: could include exploratory and/or explanatory analysis, developing and running models. In this phase you are executing your ideas and dealing with challenges as they arise. Good resources will allow you to quickly learn how to use a tool and easily find solutions to specific technical questions and problems.

Presentation, Reporting, and Evaluation
Presenting your findings and results, possibly your methodology. In this phase you want resources to help with formatting, or maybe places to talk about what you did and get feedback on how to share it with others.

General Skill Building
You may want to explore or build skills in Python without a specific project in mind.

4 Pull resources together to make a plan

Planning my First Data Analysis in Python

What tools do I already know?

What kind of data do I have? (file format, domain, keyword)

What task(s) do I want to do?

What phase(s) of the analysis am I working on?

What are some terms I can look up?

What are some Python tools I might learn more about?

What types of resources might I look for?

About this Project

Problem

Non-programmers interested in learning programming as a tool to help them work with data need a way to identify appropriate learning resources because there's so much out there it can be daunting.

Solution

My solution is an interactive tool like a "Choose Your Own Adventure" book that uses information on why a learner wants to learn Python for data analysis and what specific task they want to work on to recommend ways to search for resources. It will also direct the learner to some actual resources for learning Python.

Types of Resources

Google
Google can be really useful if you're just perusing something really broad or finding something really specific.

Tips: take some time to find the right search terms. Copying and pasting error messages into Google can help you with troubleshooting.

One-on-One Help
There are probably others out there who can help you with your project. Asking for help is a great way to connect with other Python people near you!

Tips: Be clear with your ask- what do you want from this person? Write down specific questions to discuss. Remember that they may not remember what it's like to not know a core concept.

Websites, Tech Blogs, and GitHub
Other people out there may have done an analysis similar to yours. They may have written up a guide on how to do it, where they struggled, and they may post their code! You may find others with a similar background and struggles that you can learn from.

Tips: Find people you like, who have the same interests, or who are doing interesting things. Or make your own!

Documentation
The information that comes with libraries and programming tools can show you how to use them and give you details about how they work.

Tip: These documents assume you have a base knowledge of how Python works and can be hard to read, especially if you are a beginner. Look for examples to see how the code needs to be written.

Books
There are a lot of guides, how-to books, and reference books out there.

Tips: Identify what kind of book you are looking for, and evaluate it on content and readability. Is it what you want to learn? Is it way too easy or too hard? Does it walk you through exercises and does it come with code you can download?

Event (Hackathon, Conference)
Events are a great place to get ideas or work as part of a team on a project. They are also a great place to make connections with people who have similar interests.

Tips: Once you know a skill, a hackathon can be a great place to try it out; however it may be difficult to start learning a skill as a team tries to do a project.

Local Python User Group
Your local user group is a great place to meet other Python users and to get feedback on a Python project you did.

Tips: Presenting a project to your local user group is a great way to get feedback on tools you used, hear about other ways to do it, and learn ways to refactor your code and make it readable, Pythonic, and efficient.

Online Class
There are lots of different places to take free online classes. This is a great way to build deeper knowledge and skills but may not be a great resource for the middle of a project you want to get done!

Tips: Look for reviews to find well-executed classes, there are some bad ones out there. Read the syllabus to make sure it's the content you want to learn. Be intentional about your goal for the class. Find others to join you in a cohort and meet up!

In-Person Class, Workshop
Live classes and workshops can be a great way to learn with access to an instructor who is familiar with teaching learners and can answer questions.

Tips: Try to find a class that is the right skill level for you. Look at the class materials or read the syllabus if possible. You want something that is challenging but also manageable. You may want to try and build some basic skills before the workshop to prepare.

My Ongoing Journey

While the orange path is nice and neat, my actual path to learning Python looks like a game of Chutes and Ladders, inspiring me to make this tool. My hope is that it will help other non-programming data people:

- **See what's possible and be inspired to try it**
- **Create a mental map of the landscape that is learning Python for data analysis, so that they can navigate it more easily to get from where they are now to where they want to be**
- **Identify places they might get stuck and help them search for the right resources to get through it**

I hope to take this prototype and feedback from sharing it to inform another iteration of this tool. Stay tuned!

Roxanne Johnson, PyCon 2016