Data analysis with Python for Beginners

Opening Session

Mohan Pudasaini

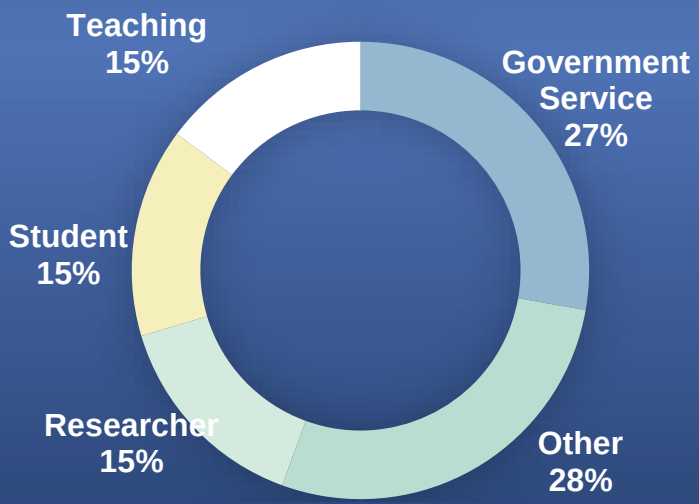
2079-11-14

Introduction

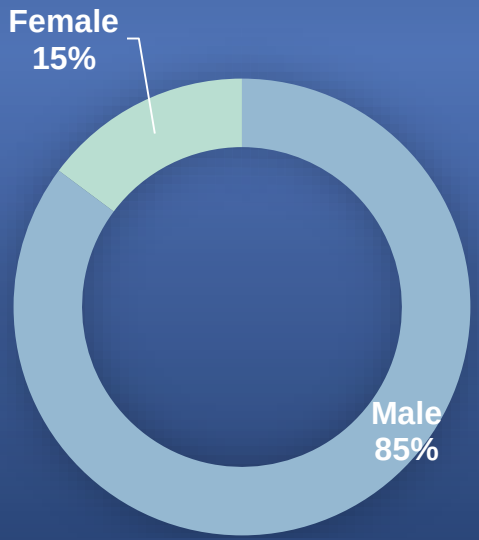
About the Participants

Total Participants: 54

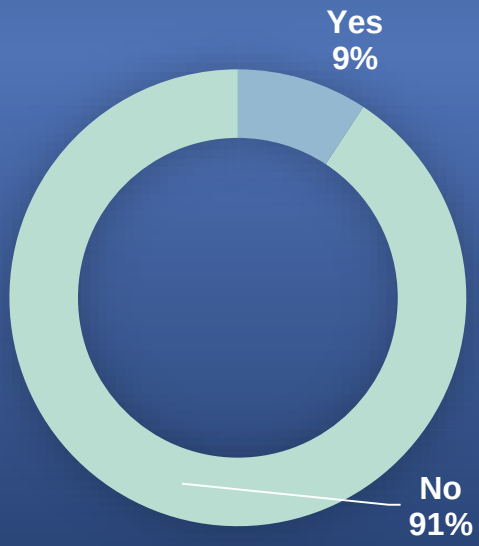
Service wise



Gender Wise



Python Experience



Scope of the workshop

Scope of the workshop

Before workshop

Problem

After workshop

Cleaning

Summarizing

Visualization

ID	TOTAL ACTIONS	ACTION 1	ACTION 2	TOTAL TIME
10	120	80	40	0:50:05
11	255	130	125	1:40:03
12	180	100	80	1:20:19
13	305	205	100	1:58:58
14	71	50	21	0:35:41
15	418	310	108	2:08:18
16	222	150	72	1:32:58

summary	age
count	9
mean	49.22222222222222
stddev	20.897235330168545
min	21
max	79

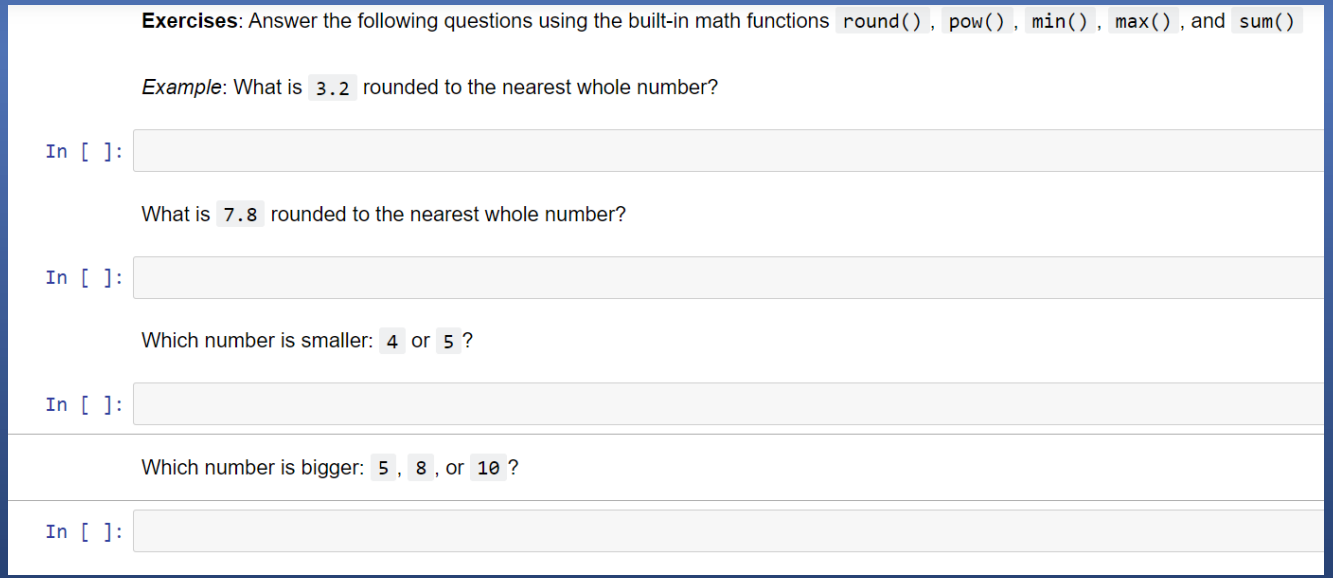
Workshop Objective

With no prior coding skills assumed, participants should be able to:

- Understanding the fundamentals of Python including data types, basic operations, and syntax.
- Learning how to use Jupyter Notebook and IDEs for Python
- Familiarizing with popular Python libraries and packages for data manipulation, analysis, and visualization, such as Numpy, Pandas, Matplotlib, and Seaborn.
- Learning how to use loops, functions, and conditional statements in Python to create efficient and reusable code.

Main Feature of the Workshop

workshop includes live exercises that are designed to help participants gain practical experience and develop skills in Python programming, data analysis, and visualization



Exercises: Answer the following questions using the built-in math functions `round()`, `pow()`, `min()`, `max()`, and `sum()`

Example: What is `3.2` rounded to the nearest whole number?

In [ ]:

What is `7.8` rounded to the nearest whole number?

In [ ]:

Which number is smaller: `4` or `5`?

In [ ]:

Which number is bigger: `5`, `8`, or `10`?

In [ ]:

What is Python

- Python is a general-purpose, open-source programming language.
- It has a simple syntax and is used for various tasks, such as web development, game development, scientific computing, and automation.
- Python is highly versatile and can be used for data science tasks like data cleaning, analysis, and visualization, as well as artificial intelligence tasks like natural language processing and computer vision.
- Python has a large community of developers continually improving and updating its libraries and tools.
- Many tech giants, like Google, Facebook, and Instagram, use Python for various applications beyond data science, such as server-side web development, testing, and automation.

Python Popularity

PYPL Index

Worldwide, Feb 2023 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	27.7 %	-0.7 %
2		Java	16.79 %	-1.3 %
3		JavaScript	9.65 %	+0.6 %
4	↑	C#	6.97 %	-0.5 %
5	↓	C/C++	6.87 %	-0.6 %
6		PHP	5.23 %	-0.8 %
7		R	4.11 %	-0.1 %
8	↑↑	TypeScript	2.83 %	+0.8 %
9		Swift	2.27 %	+0.3 %
10	↓↓	Objective-C	2.25 %	-0.1 %

https://pypl.github.io/PYPL.html

TIOBE Index

Feb 2023	Feb 2022	Change	Programming Language
1	1		Python
2	2		C
3	4	↑	C++
4	3	↓	Java
5	5		C#
6	6		Visual Basic
7	7		JavaScript
8	10	↑	SQL
9	9		Assembly language
10	8	↓	PHP

https://www.tiobe.com/tiobe-index/

Python Application

Web Development

Game Development

Machine Learning and Artificial Intelligence

Data Science and Data Visualization

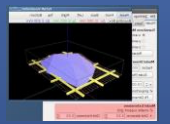
Desktop GUI

Web Scraping Applications

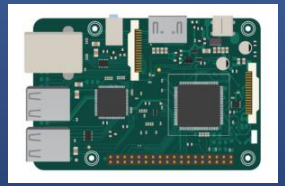
Business Applications

Audio and Video Applications

CAD Applications



Embedded Applications



https://www.edureka.co/blog/python-applications/

Career Scope- Data

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century –2012
Is Data Scientist Still the Sexiest Job of the 21st Century?- 2022

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Salary

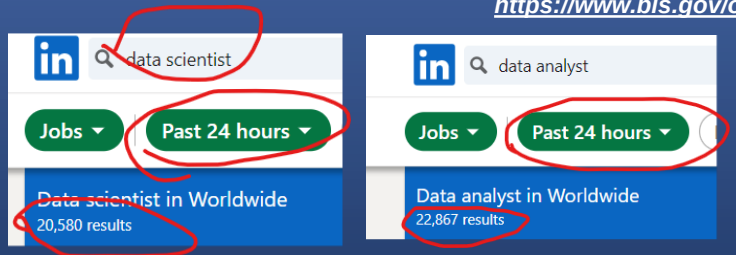
$125,955 / y *https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm*

Employment growth

Employment of data scientists is projected to grow 36 percent from 2021 to 2031, much faster than the average for all occupations.(average growth of all job 5%)

https://www.bls.gov/ooh/math/data-scientists.htm

New job post in 24 hours





Compare Python and Other Tools in Data Science

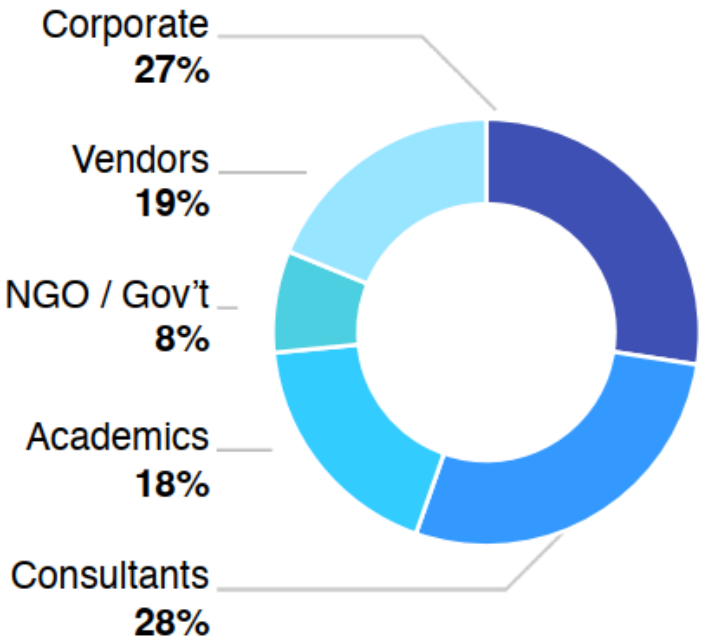
2020 Data Science Survey Highlights

Karl Rexer, PhD
President, Rexer Analytics

www.RexerAnalytics.com

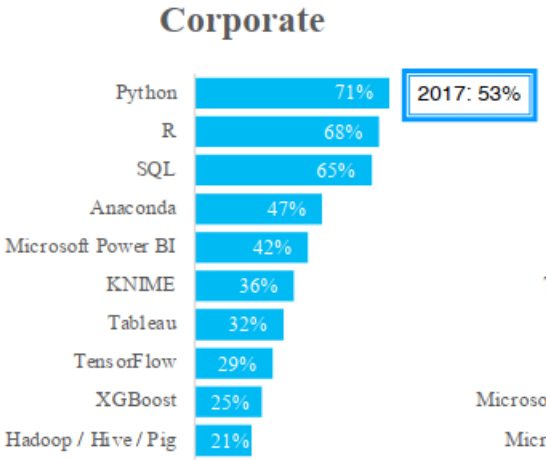
2020 Data Science Survey: Overview

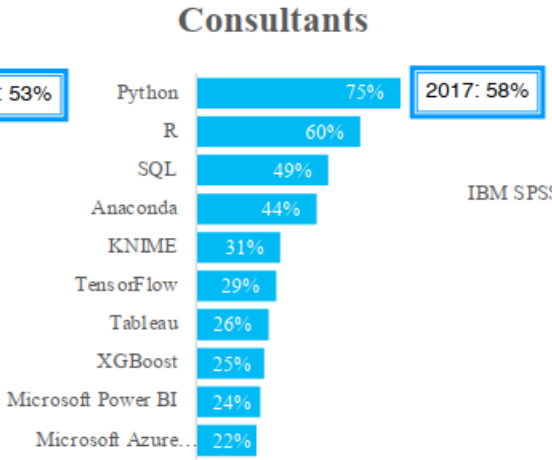
- 9th survey since 2007
- 49 questions
- 10,000+ invitations emailed & promoted by newsgroups, vendors and bloggers
- Respondents: 579 analytic professionals from 71 countries
- Data collected in May to July 2020

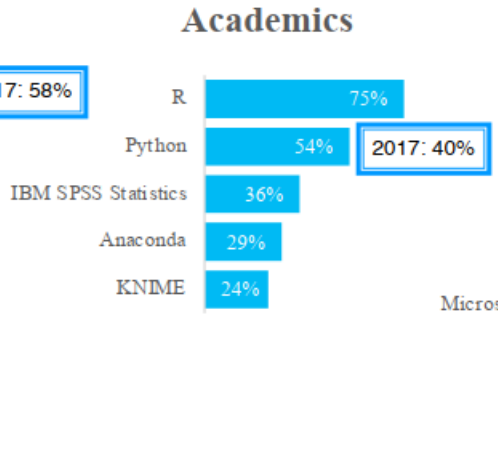


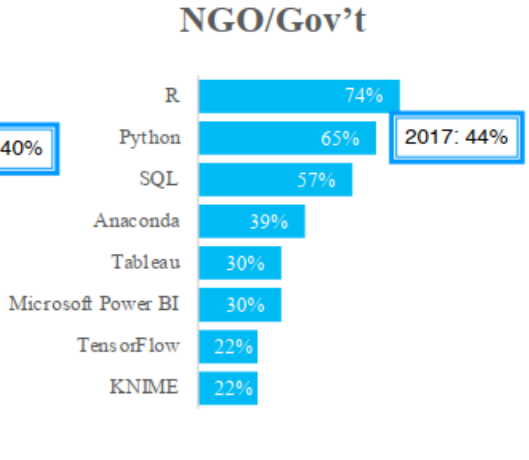
Most Data Scientists use Multiple Tools

What data science / analytic tools, technologies, and languages did you use in the past year?







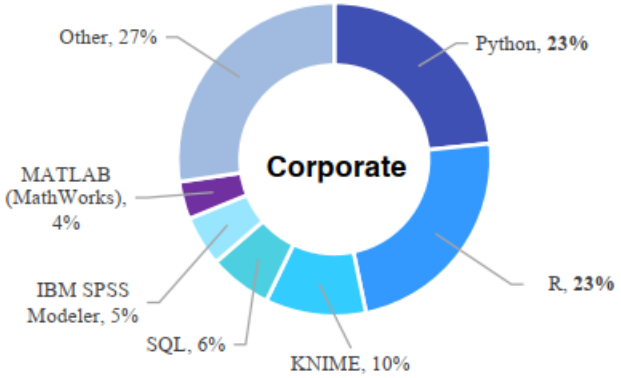


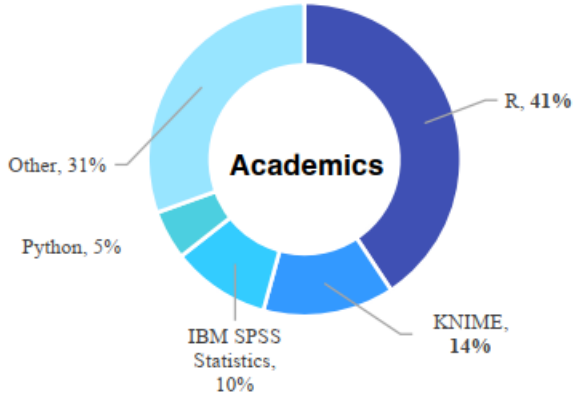
Across our four primary professional categories, Python and R are unsurprisingly the dominant analytics tools used. Python, in particular, has shown substantial growth in adoption in the past 3-5 years,

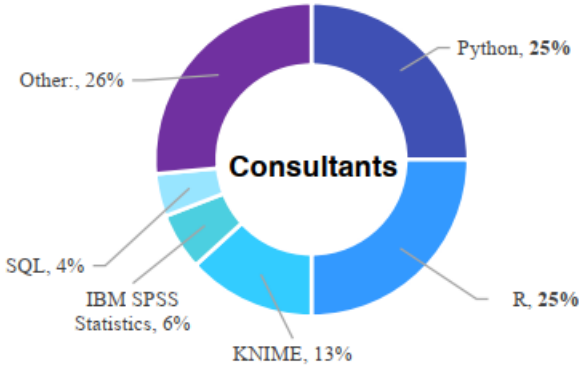
All tools used by more than 20% of a group are shown

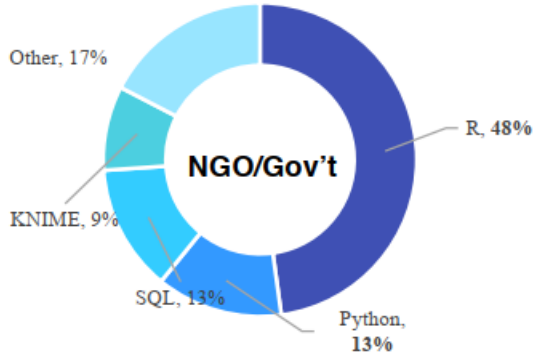
Primary Tools

What one data science tool, technology or language do you use most frequently?









Python vs R

Use Case: Data Analysis

Usage

Python is generally used when the data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database.

Since it's a full-fledged programming language, Python is a good tool to implement algorithms for use in production.

R is mainly used when the data analysis tasks require standalone computing or analysis on individual servers.

For exploratory work, R is easier for beginners. Statistical models can be written with a few lines of code.

https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

Python vs R

Advantages

- General-purpose programming languages are useful beyond just data analysis.
- Has gained popularity for its code readability, speed, and many functionalities.
- Great for mathematical computation and learning how algorithms work.
- Has high ease of deployment and reproducibility.

- Widely considered the best tool for making beautiful graphs and visualizations.
- Has many functionalities for data analysis.
- Great for statistical analysis.
- Built around a command line, but the majority of R users work inside of RStudio, an environment that includes a data editor, debugging support, and a window to hold graphics as well.

https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

Python vs R

Disadvantages

- Python doesn't have as many libraries for data science as R.
- Python requires rigorous testing as errors show up in runtime.
- Visualizations are more convoluted in Python than in R, and results are not as eye-pleasing or informative.

  Python packages for data visualization:

  - **seaborn**: Library based on Matplotlib
  - **Bokeh**: Interactive visualization library
  - **Pygal**: Create dynamic dynamic svg charts

- For people with no software engineering experience, base R can be more difficult to learn because it was developed by statisticians, not to make coding easier. But R has a set of packages known as the Tidyverse, which provides powerful yet easy-to-learn tools for importing, manipulating, visualizing, and reporting on data.
- Finding the right packages to use in R may be time consuming.
- There are many dependencies between R libraries.
- R can be considered slow if code is written poorly.
- Not as popular as Python for deep learning and NLP.

Python install

Jupyter notebook Install