Paper 184-2023

From Data Access to Exploratory Data Analysis – My Journey into the World of Python

Leon Rod Davoody

ABSTRACT

Kids who learn to code with Python can improve their critical and logical thinking and problem-solving skills. They can better understand everything by breaking complex tasks into smaller steps. Also, by building something from scratch, they can exercise their creativity and see firsthand the result of their efforts. In this paper I will share with you my journey through the world of Python and the basic statistics using Python.

INTRODUCTION

What is Python? Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.

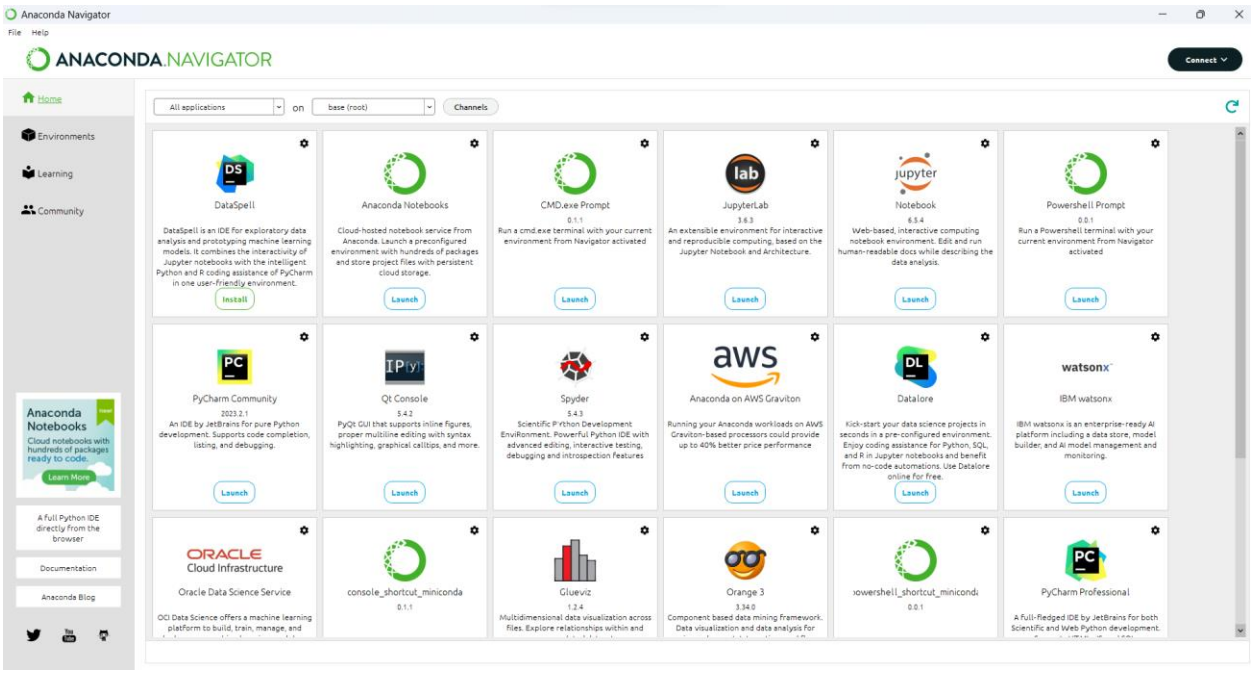
Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to

Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

ANACONDA

I learned it is easier to run Python through Anaconda.

What is the use of Anaconda in Python?



Anaconda software helps you create an environment for many different versions of Python and package versions. Anaconda is also used to install, remove, and upgrade packages in your project environments. Furthermore, you may use Anaconda to deploy any required project with a few mouse clicks.

You can go to this link to download Anaconda; Anaconda Links: [Anaconda Links]

When you install and open Anaconda you will see many subprograms inside Anaconda including: PyCharm, Spyder, Jupyter, etc. My observation was that Spyder was easier and less difficult to master.

What is Spyder?

Spyder is a free and open-source scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It features a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

Spyder is the Scientific Python Development Environment:

• A powerful interactive development environment for the Python language with advanced editing, interactive testing, debugging and introspection features.

• A numerical computing environment thanks to the support of IPython (enhanced interactive Python interpreter) and popular Python libraries such as NumPy (linear algebra), SciPy (signal and image processing) or matplotlib (interactive 2D/3D plotting).

Spyder is inside Anaconda package, or you download it separately from below websites:

- Downloads, bug reports and feature requests: https://github.com/spyder-ide/spyder
- Discussions: http://groups.google.com/group/spyderlib

To learn about some basic statistics using Python, you can download my simplifier version dataset through my account at GitHub: https://github.com/search?q=SummerSports2016&type=repositories. However, you can download the original dataset through website: https://www.kaggle.com/datasets/rio2016/olympic-games?rvi=1

Example #1 – Read and Display SummerSports2016 SAS Dataset:

```
# Import Libraries
import pandas as pd
# Name and Location of SAS Dataset
SAS_FILE = 'D:\Workshop Data\SummerSports2016.sas7bdat'
# Read All Rows from SummerSports2016 SAS Dataset
sasfile= pd.read_sas(SAS_FILE,encoding='latin-1')
# Display Detail SummerSports2016 Data Listing
Sasfile
```

Example #2 – Read CSV CARS Data File:

```
# Import Libraries
import csv
# Name and Location of CSV File
csvfile= open('D:\Workshop Data\SummerSports2016.csv')
# Read and Process SummerSports2016 CSV File
SummerSports2016CSV = list(csv.reader(csvfile))
print(SummerSports2016CSV)
```

Example #3 – Read Excel Data File:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
path=r"C:\Users\admin\SummerSports2016.xlsx"
data=pd.read_excel(path)
print(data)
```

Let's look at the printout of the dataset:

Output:

	ID	Sex	Age	Height	Weight	Team	NOC	Sport	Medal
0	576	1	23	198	93	Spain	ESP	Basketball	Bronze
1	1466	1	30	192	102	Italy	ITA	Water Polo	Bronze

```
2     1478   2  27    172     74   Italy  ITA  Water Polo  Silver
3     1551   1  20    172     79 Nigeria  NGR    Football  Bronze
4     1716   1  30    187     80 Nigeria  NGR    Football  Bronze
..     ...  ...  ...    ...    ...     ... ...        ...     ...
250  130489   2  34    166     66  Canada  CAN    Football  Bronze
251  130541   1  21    176     65  Brazil  BRA    Football    Gold
252  131981   2  25    180     71   Spain  ESP  Basketball  Silver
253  133648   2  23    172     67  Canada  CAN    Football  Bronze
254  134211   1  22    170     69  Brazil  BRA    Football    Gold
```

Example #4 – Produce Descriptive Statistics:

```
# Import Libraries
import pandas as pd
# Name and Location of SummerSorts2016 SAS Dataset
SASFILE = 'D:\Workshop Data\SummerSports2016.sas7bdat'
# Read SummerSports2016 SAS Dataset
sasfile= pd.read_sas(SASFILE,encoding='latin-1')
# Produce Descriptive Statistics for SummerSports2016 Dataset
sasfile.describe()
# excel data
data.describe()
```

Output:

```
                ID         Sex         Age      Height      Weight
count    255.000000  255.000000  255.000000  255.000000  255.000000
mean   64202.815686    1.505882   25.811765  183.101961   79.129412
std    38745.784540    0.500949    4.539834   12.583477   15.859711
min      576.000000    1.000000   17.000000  153.000000   51.000000
25%    29497.500000    1.000000   22.000000  173.000000   66.000000
50%    64194.000000    2.000000   26.000000  182.000000   76.000000
75%    98163.000000    2.000000   29.000000  193.000000   90.000000
max   134211.000000    2.000000   37.000000  215.000000  130.000000
```

Example #5 – Produce Frequency Table:

```
df=data.groupby('Sport')[['Sport']].count()
df['percentage'] = 100*df.Sport/df.Sport.sum()
df['Valid_percentage'] = 100*df.Sport/df.Sport.sum()- df.Sport.isnull().sum()
df['cumulative_percentage'] = 100*df.Sport.cumsum()/df.Sport.sum()
df.rename(columns = {'Sport':'Frequeny'}, inplace = True)
df
```

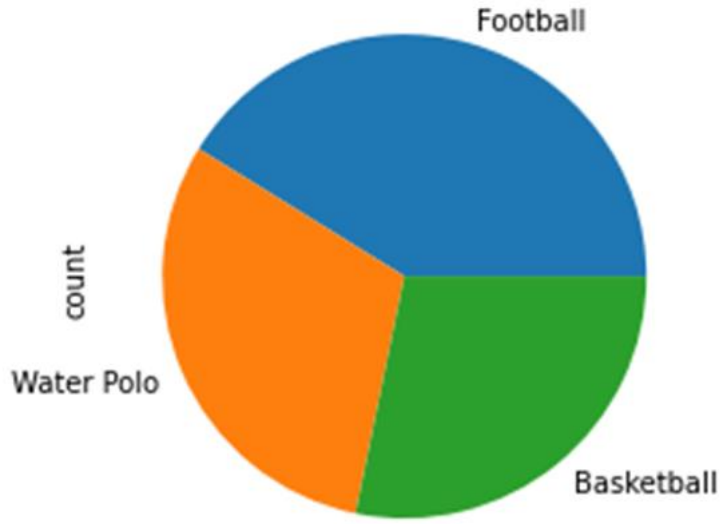
Output:

```
           Frequeny  percentage  Valid_percentage  cumulative_percentage
Sport
Basketball       72   28.235294         28.235294              28.235294
```

Football	105	41.176471	41.176471	69.411765
Water Polo	78	30.588235	30.588235	100.000000

Example #6 - Visual Charts (e.g., Produce Pie chart:)

data['Sport'].value_counts().plot(kind='pie')



"Free" Data Resources to Consider:

11 Websites to Find "Free" and Interesting

90 "Free" Datasets for Your Next Data Science Project

36 Data Analytics Project Ideas and Datasets

Anaconda Links

CONCLUSION

In this paper I shared my journey through the world of Python and the basic statistics using Python. I found out it is easier for me to run Python through Anaconda.

ACKNOWLEDGMENTS

I would like to thank Kirk Paul Lafler for mentoring me to develop this paper.

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS® For Dummies®*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Leon Rod Davoody

lrdavoody@gmail.com

Author Biography:

At the time of this presentation, I am in 6th grade and began using Python two years ago at an educational summer camp. I wrote this paper to show young kids can start at an early age in programming. Also, I like and mainly play water polo. However, I play other sports as well including basketball and football. The dataset I selected shows the outcome of the summer Olympics 2016 competition for the sports water polo, basketball, and football.