International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 12 | Dec - 2020 ISSN: 2582-3930

A Study on Data Science Basics with Python concepts

Mrs.R.PRABA

Assistant Professor, Department of Information Technology,

Dr.N.G.P. Arts and Science College, Coimbatore.

Abstract

Data Science deals with the huge volume of data, which includes data collection, analyzing the data and finally the decision making and also the finding patterns in data through analysis and making the future predictions. Data science is an Emerging technology in all the fields such as Banking, Health care, Consultancy and Manufacturing for finding the better results. This paper gives an overview of Data Science and python basics.

Keyword- data science, python, libraries,applications

Introduction

Data Science Emerging Technology for the purpose of Data Collection and the data analysis and the Future predictions. The programming language used for the Data Science is Python and R. Mostly Python is used because of its good functionalities such as Open Source, Interpreted Language and so on.

Banking, Consultancy, Manufacturing and Healthcare Industries mostly using the Data Science for the best planning, the analysis of health benefits training and the future predictions (ex: Who will win the Elections).Data Science can be applied every businesses where the data is available. The one who mastering in the Machine Learning, Statistics, R Programming (or) Python Programming, Mathematics and the Data Bases can become a Data Scientist. Data Scientists are used to understand the business problems, data collection, data extraction, data cleaning, data normalization, future predictions and the result representations.

Python is a popular programming language used by the Data Scientist. Python was created by Guido van Rossum, and released in 1991. Python has built-in mathematical functions and the libraries such as Pandas, Numpy, Matplotlib and Scipy to calculate the mathematical problems and for the data analysis.

Most of the Data Science work can be carried out by using the Python Programming (or)

R-Programming

Python and its Libraries

1. Pandas

Pandas are an Open Source Library used for the Data Manipulation. Pandas are best for the time series Data. It has functions for analyzing, cleaning, exploring and manipulations. Pandas allow the user to analyze the big data and make the conclusions based on the statistical theories.

Data frames

Panda's data frames are rows and columns. It is like two dimensional arrays,

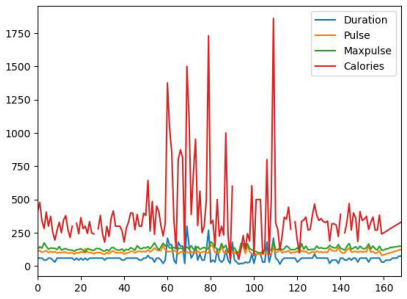
```
import pandas as pd

data = {"Roll No": [101, 102, 103], "Marks": [50, 75, 85]}

#load data into a DataFrame object

df = pd.DataFrame(data)

print(df)
```

OUTPUT

```
Roll    No     Marks
0       101      50
1       102      75
2       103      85
```

Data cleaning

Data cleaning can be done in pandas for the bad data like missing values, wrong format, wrong data and the duplicates.



Pandas-Example

2. Numpy

Numpy stands for Numerical Python consists of multi dimensional array objects, used for the purpose of processing the arrays.

Numpy Data types

- `i` - integer
- `b` - boolean
- `u` - unsigned integer
- `f` - float
- `c` - complex float
- `m` - timedelta
- `M` - datetime
- `O` - object
- `S` - string
- `U` - unicode string
- `V` - fixed chunk of memory for other type (void)

Example

```
import numpy as np

arr = np.array([[[1, 2, 3], [4, 5, 6]], [[7, 8, 9], [1, 2, 3]]])

print(arr)
```

OUTPUT

```
[[[1 2 3]
  [4 5 6]]
 [[7 8 9]
  [1 2 3]]]
```

3.Matplotlib

It is the multiplatform Data visualization library. Matplotlib (Most popular python plotting library) consists of several plots like line, bar, scatter and histogram.

```
#Three lines to make our compiler able to draw:

import sys

import matplotlib

matplotlib.use('Agg')
```

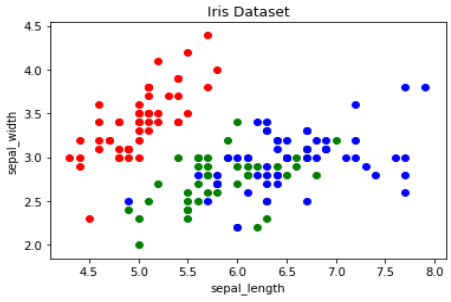
```
import matplotlib.pyplot as plt

import numpy as np

xpoints = np.array([10, 80])

ypoints = np.array([30, 10])

plt.plot(xpoints, ypoints)

plt.show()

#Two  lines to make our compiler able to draw:

plt.savefig(sys.stdout.buffer)

sys.stdout.flush()
```

Data Visualization

It is the graphical representation of the data. By using the data visualization (summarizing and presenting the data) people can understand easily.

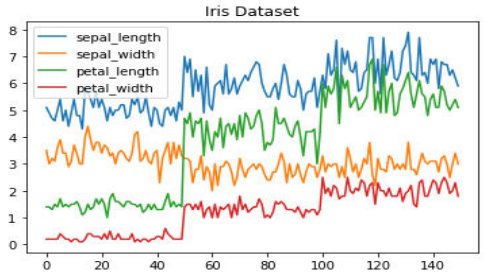
Scatter plot

Each value of the data set is represented by a dot.



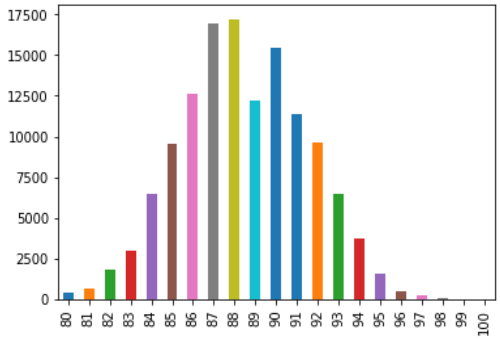
Line Chart

Line charts are represented as follows



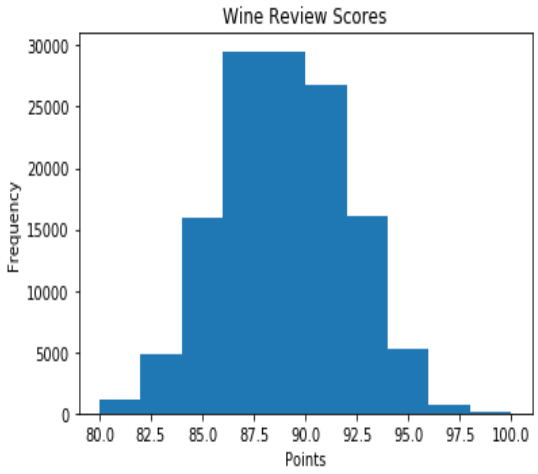
Bar Chart

Bar chart is used for the categorical data.



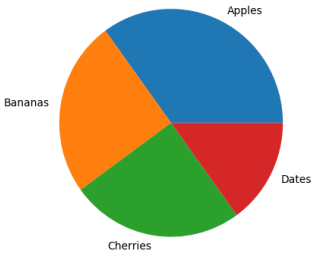
Histograms

Histograms showing the frequency distributions



Pie Charts

Pie charts are created using pie function



Scipy

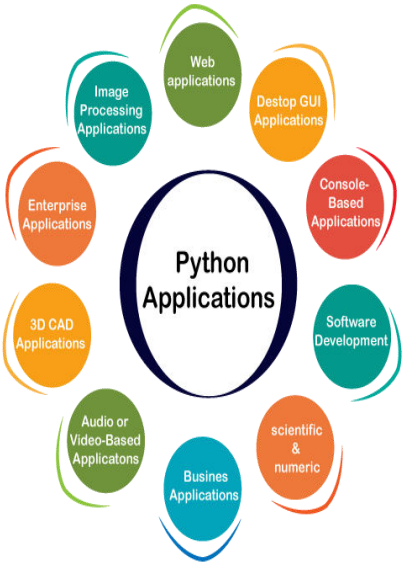
Scipy is to solve the Scientific and Mathematical problems. Scipy is an open source so we can use it freely.

Scipy unit categories

- Metric
- Binary
- Mass
- Angle
- Time
- Length
- Pressure
- Volume
- Speed
- Temperature
- Energy
- Power
- Force

Python applications

Python is the fastest growing programming language and can create any of the applications in all the fields.



1. **Web applications**
   Web applications can be created by using Python. It provides the libraries to handle the internet protocols like html, xml, email processing and the request.

2. **Desktop GUI Applications**
   Graphical User Interface (GUI) used for the smooth interaction in any of the applications. Tk GUI Library is used in python for the user interface.
3. **Console-based Applications**
   Command based Applications which runs on the command line or shell.
4. **Software Development**
   Software can be created by using python.
5. **Scientific and Numeric**
   Python is the most suitable language for the Artificial Intelligence and Machine Learning
6. **Business Applications**
   E-commerce and ERP applications can be created.
7. **Audio or Video based Applications**
   Multimedia Applications can be created with the help of Timplayer, cplay

8. **3D CAD Applications**
   Python can create 3D CAD Applications for the design engineering related architecture
9. **Enterprise Applications**
   Applications that are related to the Organizations can be created.
10. **Image Processing Applications**
    Most of the python libraries are used to work with the images.
    Examples: OpenCV, Pillow.

Conclusion

In this paper, Basic concepts of the Data Science are explained. Python programming and its libraries are used for the Data analysis, Visualization and the Future predictions are discussed with the examples and graphical representations. So this study is helpful for the better understanding of the basic Data Science concepts.

References

1. T.Giri Babu and Dr.G.Anjan Babu, " A Survey on Data Science Technologies &Big Data Analytics" Volume 6, Issue 2,February 2016 ISSN: 2277 128X
2. Michel Dumontier and Tobias Kuhn, "Data Science – Methods, infrastructure, and applications" Data Science 1 (2017) 1–51DOI 10.3233/DS-170013IOS Press
3. http://math.ecnu.edu.cn/~lfzhou/seminar/[Joel_Grus]_Data_Science_from_Scratch_First_Princ.pdf
4. https://www.guru99.com/data-science-tutorial.html
5. https://www.programmer-books.com/introducing-data-science-pdf/