JULY
2021

Week 4

Introduction to Python

Designed for :

EXL

Agenda:

1) Basic python refresher
    - Data types
    - Data Structures
    - Flow control
    - Functions
    - Packages

2) Frequently used libraries
    - Numpy
    - Pandas
    - Matplotlib, Seaborn and Plotly

3) Case Study – Travel insurance data

Data types

1. Numbers – float, int
2. Strings – object, str
3. Boolean – bool

Data Structures

1. List – a mutable collection of potentially heterogenous elements [ ]
2. Tuple – an immutable collection of potentially heterogenous elements ( )
3. Set – a set of elements without duplicates
4. Dictionary – arrangement of data as key-value pairs

Flow Control

1. If/else statements – Used to check and proceed with a condition

2. For-each and while loops – Iterate over the same instructions until a condition is met

Functions

1. Pre-defined functions – functions that come built-in with Python (sum, print, input etc)

2. User Defined functions – custom functions created by user which can be called upon any number of times

3. Lambda functions – a one time use function, defined by the keyword “lambda”

Frequently used libraries

1. Numpy
2. Pandas
3. Matplotlib
4. Seaborn
5. Plotly

Numpy

NumPy is the fundamental package for scientific computing with Python.

It contains -

- a powerful N-dimensional array object - ndarray
- sophisticated (broadcasting) functions
- useful linear algebra, Fourier transform, and random number capabilities etc.

Refer - http://www.numpy.org/

Pandas

- A library written for the Python programming language for data manipulation and analysis
- In particular, it offers data structures and operations for manipulating numerical tables and dataframes

Practical example of the usage of dataframe, series & array on a dataset

Demographic data

Country Name	Birth rate	Internet users	Income Group
Aruba	10.244	78.9	High income
Afghanistan	35.253	5.9	Low income
Angola	45.985	19.1	Upper middle income
Albania	12.877	57.2	Upper middle income
United Arab Emirates	11.044	88	High income

Convert to dataframe

	Country Name	Country Code	Birth rate	Internet users
0	Aruba	ABW	10.244	78.9
1	Afghanistan	AFG	35.253	5.9
2	Angola	AGO	45.985	19.1
3	Albania	ALB	12.877	57.2
4	United Arab Emirates	ARE	11.044	88.0

Extract Birth rate as Pandas Series

```
0    10.244
1    35.253
2    45.985
3    12.877
4    11.044
```

Extract birth rate as numpy array

```
array([10.244, 35.253, 45.985, 12.877, 11.044])
```

Convert dataframe to numpy array

```
array([['Aruba', 'ABW', 10.244000000000002, 78.9],
       ['Afghanistan', 'AFG', 35.253, 5.9],
       ['Angola', 'AGO', 45.985, 19.1],
       ['Albania', 'ALB', 12.877, 57.2],
       ['United Arab Emirates', 'ARE', 11.044, 88.0]], dtype=object)
```

Visualization (Matplotlib, Seaborn and Plotly)

There are five key plots that you need to know well for basic data visualization. They are:

- Line Plot
- Bar Chart
- Histogram Plot
- Box and Whisker Plot
- Scatter Plot

Chart selection

X Variable	Y Variable	Purpose of analysis	Type of chart	Example
Continuous (numerical)	Continuous (numerical)	How Y changes with X	Scatter plot	How cholesterol varies with Age?
Continuous (numerical)	Categorical	How range of X varies for various category levels	Box plot	Cholesterol variation with Men and Women
Categorical	Categorical	What is the number or % of records of X which falls under each category	Stacked bar	How many men have heart disease compared to women?
Continuous	-	Look at the distribution of the values of the X variable	Histogram, boxplot	Distribution of cholesterol ranges
Impact of 2 X variables on a Y variable			Facet_grid()	Distribution of chol across mean and women – compared for people who have and don’t have heart disease

Travel Insurance case study

QUESTIONS ?