Python Programming for Life Science Students (PyLifeS)
January - April semester, 2024

A course offered to the students of Integrated M.Sc Systems Biology
at the School of Life Sciences, University of Hyderabad

Raghunathan Ramakrishnan
ramakrishnan@tifrh.res.in
Tata Institute of Fundamental Research Hyderabad
Hyderabad, India

Course material: https://github.com/raghurama123/pylifes

tifr HYDERABAD | TATA INSTITUTE OF FUNDAMENTAL RESEARCH

Content

Topic		Slide
General References	...	3
Working with GitHub	...	4
Python basics	...	5
Count Roses: A mini project with Pandas	...	6

References

General References

1. *Understanding Bioinformatics*, Marketa Zvelebil, and Jeremy O. Baum (Garland Science, 2008).

2. *Bioinformatics Algorithms: Design and Implementation in Python*, Miguel Rocha University of Minho, Braga, Portugal Pedro G. Ferreira (Academic Press, 2018).

3. *Computing Skills for Biologists: A toolbox*, Stefano Allesina & Madlen Wilmes (Princeton University Press, 2019).

4. *Python for the Life Sciences: A Gentle Introduction to Python for Life Scientists*, Alexander Lancaster, Gordon Webster (Springer, 2019).

5. William Bo Rothwell, "*Linux for Developers*", Pearson (2018). *See the chapters about GitHub*.

Working with GitHub

This section has been addressed in a practical, hands-on session. If you're unfamiliar with GitHub, allocating approximately 30 minutes to follow through these steps is recommended.

Git and GitHub

- `git` is a commandline program (version control system) which you can access from Linux/Mac terminal.
- GitHub is a platform where you can store and manage your codes/repository
  - interactively on the GitHub website or the GitHub desktop program (*https://desktop.github.com/*) which you can install in your laptop or desktop
  - through `git` commands in a terminal. Check Ref.5 to learn about `git` commands.

Fork the main branch (once)

- Create a GitHub accound and fork the main branch : *https://github.com/raghurama123/pylifes*

Push your changes

- If you have GitHub-deskop installed on a Windows environment, your copy of the repository should be located at `Onedrive/Documents/Github/pylifes`
- You can edit any files present in this folder or create a new file.
- Then, open the GitHub-desktop application, and figure out how to `commit` and `push` your changes

Pull and merge new changes

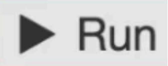
- Before every class, check your copy of the course content at *https://github.com/USER/pylifes*
- Click the message x `commits behind`
- Click `Create pull request` (which you have to click twice) and click `Merge pull request`

Python basics

You are expected to have the Jupyter Notebook platform installed to work with Python codes. Those with a Windows-based laptop, may install Anaconda *https://docs.anaconda.com/free/anaconda/install/windows/*

Your first Jupyter notebook

- The essential aspects of the python language are collected in *notebooks/PyLifeS01*
- You should read the note book, line-by-line, clarify your doubts.
- Run each line input cell by `shift + enter` or by clicking the run button at the top



- You can type-in any new code, text or comments in this notebook.
- You can also copy the input from one cell and paste in another cell, modify it and run it.
- At the end of each session, remember to save your notebook, and commit/merge to your repository.

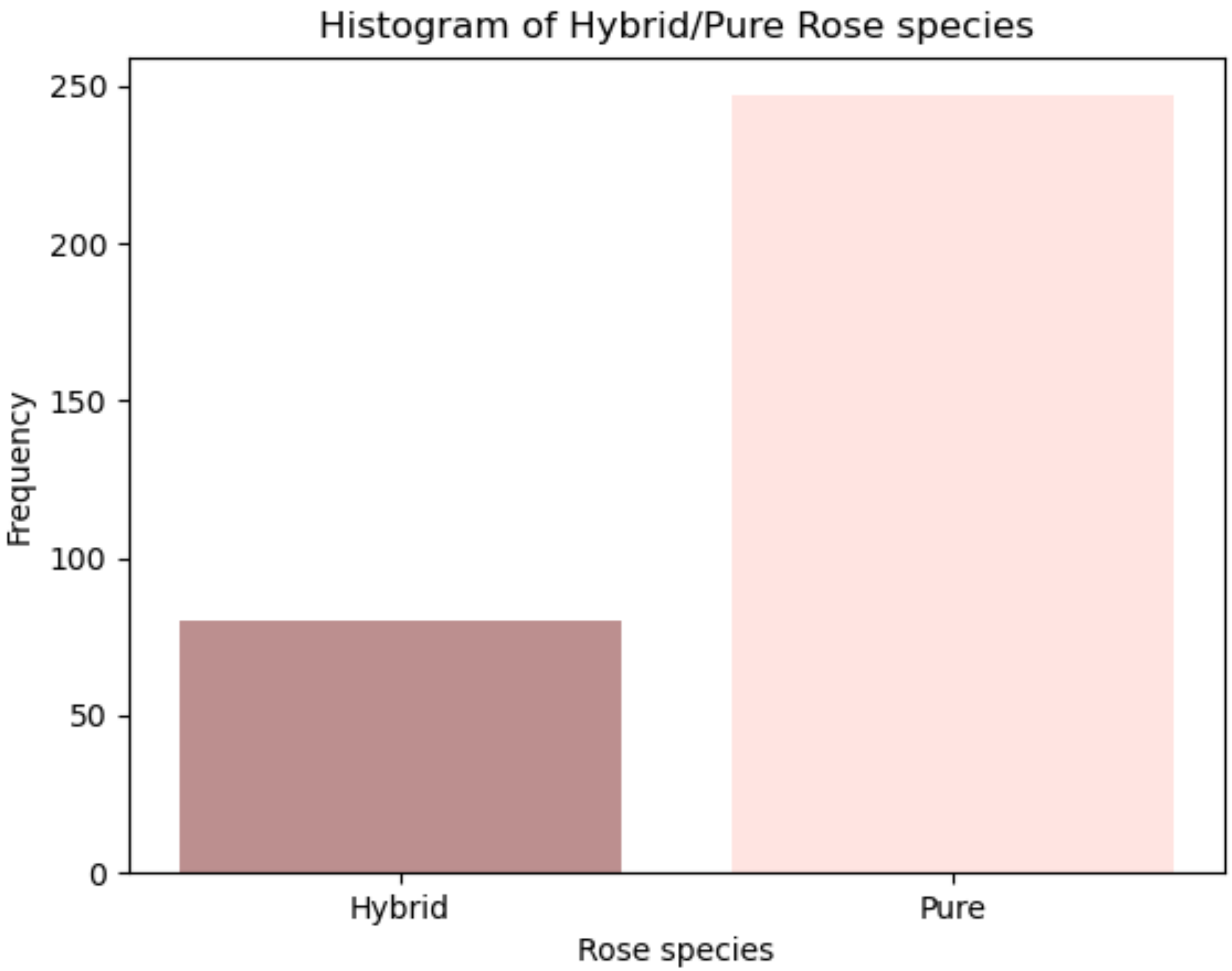
Topics covered

1. Basic datatypes in Python (int, float, complex, string)
2. String manipulation (splitting, splicing)
3. Python list (more commonly encountered datatype during string processing)
4. String concatenation and list appending (building larger strings)
5. For loop (to iterate over a list)
6. While loop (when the number of cycles is unknown apriori)
7. Break vs. continue (gracefully exiting a loop)
8. Functions (writing custom functions, doc string, help)
9. Module (importing modules, math, numpy, numpy arrays.)
10. Input/output (Read from/Write in a file, formatted printing)

Count Roses: A mini project with Pandas

Load the Rose dataset and perform basic statistics

- Genus and species names of all the known species of Rose plant are collected in the file `datasets/Rose.csv`. Let's load this file using the library Pandas and count the number of hybrid species where the Genus and species names are separated by 'x' and pure species.
- We will learn to use some functions in Pandas functions, and attributes of objects created using Pandas, and finally use the library matplotlib to plot a histogram as shown below.
- The necessary python codes are given in *notebooks/PyLifeS02_Rosa*



Self-study Task: Explore Pandas

Exercise 1

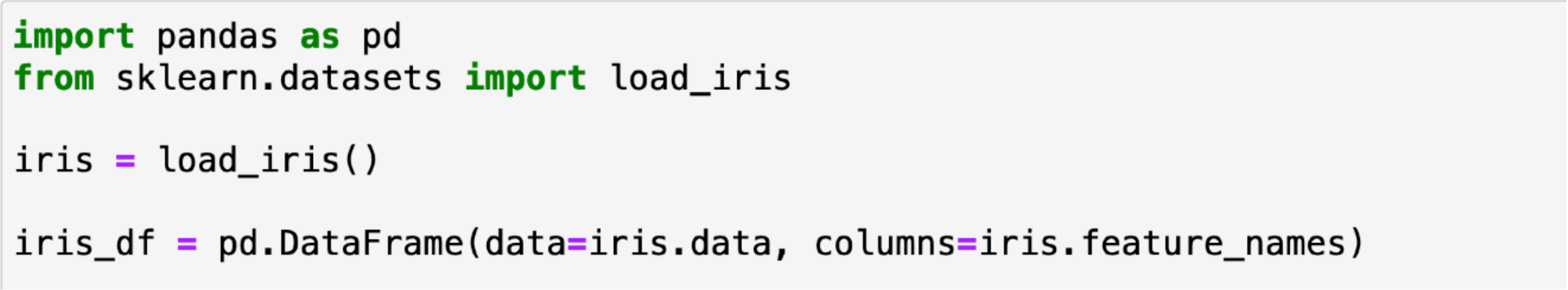
The Pandas function `pd.read_csv` was used to read a csv file as a dataframe. The function `dataframe.to_csv` can be used to write a dataframe in a csv file. Extend *notebooks/PyLifeS02_Rosa* by including Python statements to prepare separate csv files for the scientific names of hybrid and pure species of roses.

Exercise 2

Suppose you have a dataframe named df. Explain the Python statement df_copy=df.copy()

Exercise 3

The iris flower dataset containing measurements of sepal length, sepal width, petal length and petal width can be loaded directly in your code from the scikit-learn library.

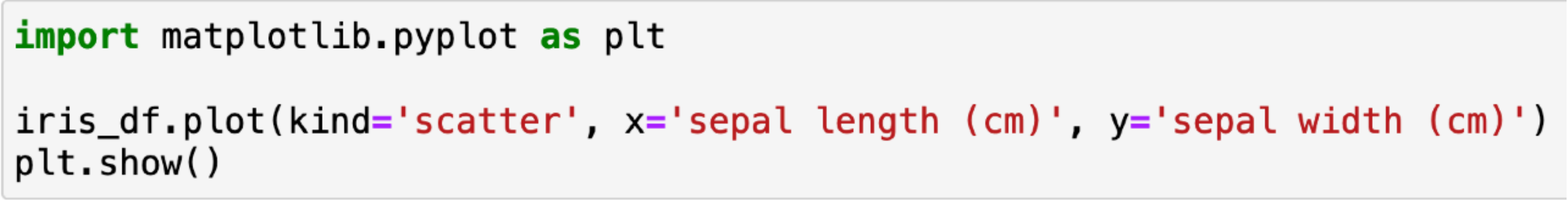


```
import pandas as pd
from sklearn.datasets import load_iris

iris = load_iris()

iris_df = pd.DataFrame(data=iris.data, columns=iris.feature_names)
```

Make a scatterplot of sepal length and sepal width using the following code.



```
import matplotlib.pyplot as plt

iris_df.plot(kind='scatter', x='sepal length (cm)', y='sepal width (cm)')
plt.show()
```

Try the following statements individually and explain the output.

`print(iris_df.describe())`, `print(iris_df.info())`, `print(iris_df.sum())`, `print(iris_df.mean())`, `print(iris_df.median())`, `print(iris_df.max())`, `print(iris_df.min())`