I am Bilingual - Python and R

Priyanga Dilini Talagala Thiyanga S. Talagala

2021-04-27

Contents

Preface

WIP!!

This book is still in progress in various draft forms.

Chapter 1

Introduction to R and Python

R Vs Python: What's the Difference? https://www.guru99.com/r-vs-python.html

1.1 About R and Python

1.1.1 R

R is an object oriented, open source programming **language** and **environment** for statistical computing and graphics. R is not a statistics system but an environment within which statistical techniques are implemented. Further, R gains more capabilities via packages, its fundamental shareable units that bundle together R functions, code, data, documentation, and tests etc. (R Core Team, 2020).

1.1.2 Python

Python is an object-oriented, interpreted, and interactive programming language. The motto of Python language is "Batteries included" as the functionality of the language can be performed via its comprehensive standard in built Libraries (Wikipedia contributors, 2020a).

1.2 History of R and Python

1.2.1 R

R is an implementation of the S programming language which was created by John Chambers in 1976. In 1991, an alternative implementation of the basic S language was developed by Ross Ihaka and Robert Gentleman, University of Auckland, New Zealand. It was published in 1993 (Wikipedia contributors, 2020b).

1.2.2 Python

In 1989, Guido van Rossum at Centrum Wiskunde & Informatica (CWI) in the Netherlands started the implementation of Python as a successor to ABC programming language. Python 2.0 was released in 2000. Python 3.0, a major revision of the language that is not completely backward-compatible was released in 2008 (Wikipedia contributors, 2020a) . Today many developers create libraries strictly for the use with Python 3.

1.3 Story behind their names

1.3.1 R

R was introduced by **R**oss Ihaka and **R**obert Gentleman and it was named after the first names of the two authors. The name of the "S" language also had some influence on the selection of its name and it was selected partly as a play on the name of S (Wikipedia contributors, 2020b).

1.3.2 Python

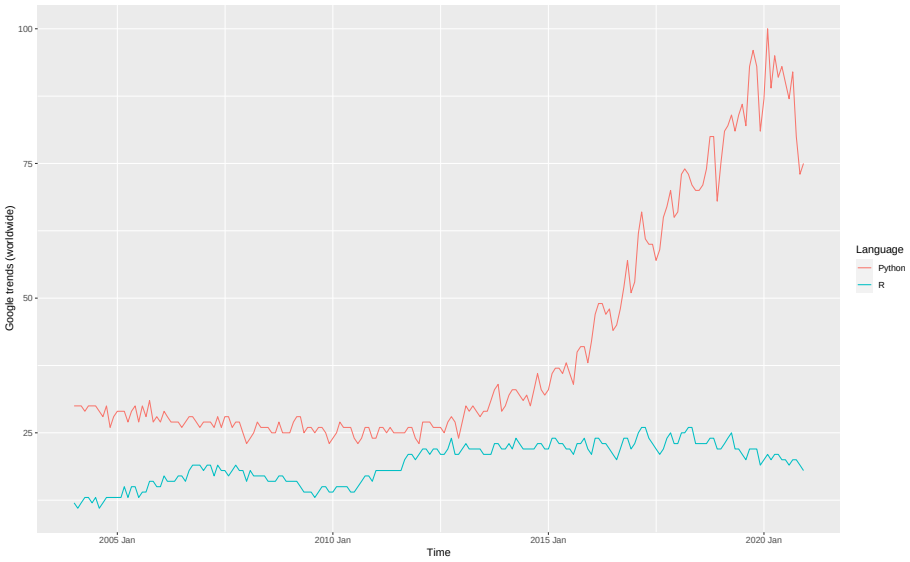
Python was named after a famous TV show 'Monty Python's Flying Circus'. Guido van Rossum, the creater of Python was a big fan of the TV show. He wanted to name his invention with a short, unique and slightly mysterious name and chose Python as a working title for his ongoing project.

1.4 Logo

Figure 1.1: Retrieved from: https://www.r-project.org/logo/

Figure 1.2: Retrieved from: https://www.python.org/community/logos/

1.5 Worldwide Google Trends



1.6 Installation

1.6.1 Python

Ref: https://www.w3schools.com/python/python_getstarted.asp

Many PCs and Macs will have python already installed.

To check if you have python installed on a Windows PC, search in the start bar for Python or run the following on the Command Line (cmd.exe):

```
C:\Users\Your Name>python --version
```

To check if you have python installed on a Linux or Mac, then on linux open the command line or on Mac open the Terminal and type:

```
python --version
```

If you find that you do not have python installed on your computer, then you can download it for free from the following website: https://www.python.org/

Install PyCharm IDE

PyCharm is a cross-platform IDE that provides consistent experience on the Windows, macOS, and Linux operating systems.

PyCharm is available in three editions: Professional, Community, and Edu. The Community and Edu editions are open-source projects and they are free, but they have fewer features. PyCharm Edu provides courses and helps you learn programming with Python. The Professional edition is commercial, and provides an outstanding set of tools and features. For details, see the editions comparison matrix.

1.6.2 R

You can download it for free from the following websites: - R (https://cran.r-project.org/)

Install Rstudio IDE

RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser.

- RStudio (https://www.rstudio.com/products/rstudio/download/#download).

1.7 Install and Load Libraries

1.7.1 R

R Packages: A Beginner's Guide https://www.datacamp.com/community/tutorials/r-packages-guide?utm_source=adwords_ppc&utm_campaignid=1655852085&utm_adgroupid=61045434222&utm_device=c&utm_keyword=%2Bload%20%2Bpackage%20%2Br&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=469789579329&utm_targetid=aud-522010995285:kwd-589281898774&utm_loc_interest_ms=9071445&utm_loc_physical_ms=1009919&gclid=Cj0KCQjwyZmEBhCpARIsALIzmnKGh4ZVHa4OxhLq0JUzpoBMMRhQvCGEmvscwcB

An **R package** is a way to organize your own work and share it with others. Typically, a package contains code, documentation for the package and the functions inside, some tests to check everything works as it should, and data sets.

Three of the most popular repositories for R packages are: CRAN, Bioconductor and Github.

1.7.1.1 Installing Packages From CRAN

```
install.packages("package_name")
```

Example

```
install.packages("tidyverse")
```

After running this, some messages will be diplayed on the console. They will depend on what operating system you are using, the dependencies, and if the package was successfully installed.

To install more than a package at the same time, we can use a character vector

```
install.packages(c("vioplot", "MASS"))
```

The function `install.packages` will download the source code from on the CRAN mirrors and install the package (and any dependencies) locally on your computer.

1.7.1.2 Load Packages

After a package is installed, you are ready to use its functionalities.

If you just need a sporadic use of a few functions or data inside a package you can access them with the notation

`packagename::functionname()`.

If you will make a more intensive use of the package, then maybe is worth to load it into memory. The simplest way to do this is with the `library()` command.

Please note that the input of install.packages() is a character vector and requires the name to be in quotes, while library() accepts either character or name and makes it possible for you to write the name of the package without quotes.

Once you have the package installed, you can load the library into your R session for use. Any of the functions that are specific to that package will be available for you to use by simply calling the function as you would for any of the base functions. Note that quotations are not required here.

```
library(tidyverse)
```

1.7.2 Python

Use 'import module' or 'from module import'? https://stackoverflow.com/questions/710551/use-import-module-or-from-module-import

1.8 Ranked:15Python packages

for Data Science

http://blog.thedataincubator.com/wp-content/uploads/2017/04/Ranked-15-Python-Packages-for-Data-Science.pdf

Chapter 2

Variables, expressions, and statements

2.1 Basic Exmaple

This is a test code

2.1.1 R code

```
# This is an R code
x <- 1
y <- 3
print(x+y)
```

```
## [1] 4
```

2.1.2 Python Code

The 'python' engine in knitr requires the `reticulate` package.

```
library(reticulate)
```

```
# This is a Python code
x = 1
y = 3
print(x+y)
```

```
## 4
```

Chapter 3

Conditional execution

WIP

Chapter 4

Functions

WIP

Chapter 5

Iteration

WIP

Chapter 6

Tidy workflow

Moving from R to Python: The Libraries You Need to Know https://www.kdnuggets.com/2017/02/moving-r-python-libraries.html

glm, knn, randomForest, e1071 -> scikit-learn

One thing that is a blessing and a curse in R is that the machine learning algorithms are generally segmented by package. Meaning instead of having a single (or set) of ML libraries that each implement some common algorithms, each algorithm gets its own package. It's sort of nice because you can find very esoteric, cutting edge implementations of algorithms, but it can be a pain for day-to-day use where you might be switching between algorithms. This pain is something that Python's scikit-learn solves really well. scikit-learn provides a common set of ML algorithms all under the same API. It makes switching between LogisticRegression and GradientBoostingMachines a one-liner.

reshape/reshape2, plyr/dplyr -> pandas

This was actually the subject of one of our first posts. pandas took the best parts of data munging in R and turned it into a Python package. This includes its own implementation of a data frame along with ways to modify and restructure it. Basically it took the best parts of reshape/reshape2 and plyr/dplyr and Pythonified it!

ggplot2 -> ggplot + seaborn + bokeh

One thing that R still does better than Python is plotting. Hands down, R is better in just about every facet. Even so, Python plotting has matured though it's a fractured community. If you like the ggplot-style syntax, then look no further than Yhat's own ggplot. If you're after super statistical and technical plots then reach for seaborn. And if you're in the market for some super slick, great looking interactive plots then try out bokeh.

stringr -> nothing

String manipulation in "base R" is nearly as unintuitive as it is silly. Any time I'm working with strings in R I do 2 things (in order):

briefly nod in appreciation to New Zealand for producing Hadley Wickham import stringr stringr Much obliged, New Zealand stringr is an absolute lifesaver. It's well written, performant (at least I think so), and easy to install (don't overlook this last item. if people can't install your software, there's no sense in making it).

Ok so stringr appreciation monologue complete. So the good news for you is that Python is so great for string manipulation, you don't really need a string library! It has a fantastic built-in regular expressions library, re, and a built-in string meta-libarary appropriately called string. So lucky for you, Python comes with all string-related batteries included!

RStudio -> Rodeo

To many users, RStudio is synonymous with R. And why not? It's a great IDE for data analysis in R. Historically speaking, there haven't been a lot of comparable options for Python. Of course this is no longer the case. We released the very first version of Rodeo just over a year ago and released the 2.0 for Windows, OSX, and Linux about a month ago.

"Ever since we've used RStudio, we've been looking for an IDE like it for Python. We went through IDEs such as Sublime Text and Spyder, none of which suited our likings. We searched and found Rodeo and couldn't have been more pleased with the IDE." -Stephen Hsu, University of California, Berkeley

Download Rodeo! Knitr -> Jupyter

Knitr is a great way to create reproducible and highly visual analysis using R. It's been a staple in RStudio for a while now. In the Python world, the most analagous package is Jupyter. Jupyter notebooks provide an interactive environment for programming in Python (and other languages) that focuses on reproducibility and visualization–it even has a plugin for R!

sqldf -> pandasql

sqldf is a great way for SQL users to comfortably manipulate data in R. I myself used it when I first started learning R. Way back when, Yhat actually built a similar package for Python called pandasql. Same concept: write SQL queries against your data frames, get data frames back! Fast-forward 3 years and pandasql has over 256 stars on GitHub :). Not bad for a library with only 358 lines of code!

Original. Reposted with permission. https://www.kdnuggets.com/2017/02/moving-r-python-libraries.html

Chapter 7

Import

WIP

Chapter 8

Tidy

WIP

Chapter 9

Transform

WIP

Chapter 10

Data Visualization

10.1 Data

The Palmer penguins dataset was introduced by Allison Horst, Alison Hill, and Kristen Gorman provide a great dataset for data exploration and visualization, as an alternative to iris. It was first introduced as an R package. The released version of palmerpenguins can be instaalled from CRAN with:

R Installation `install.packages("palmerpenguins")`

Using `palmerpenguins` python package you can easily load the Palmer penguins into your python environment.

Python Installation `pip install palmerpenguins`

The palmerpenguins package contains two datasets : `penguins` and `penguins_raw`. `penguins` is a simplified version of the `penguins_raw` data.

10.2 R

Load data

```
# Load Palmer Archipelago (Antarctica) Penguin Data
library(palmerpenguins)
# Return the first part of the dataset
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
```

```
##   <fct>   <fct>           <dbl>         <dbl>             <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7               181        3750 male
## 2 Adelie  Torge~           39.5          17.4               186        3800 fema~
## 3 Adelie  Torge~           40.3          18                 195        3250 fema~
## 4 Adelie  Torge~           NA            NA                  NA          NA <NA>
## 5 Adelie  Torge~           36.7          19.3               193        3450 fema~
## 6 Adelie  Torge~           39.3          20.6               190        3650 male
## # ... with 1 more variable: year <int>
```

```r
# Retrieve column names
colnames(penguins)
```

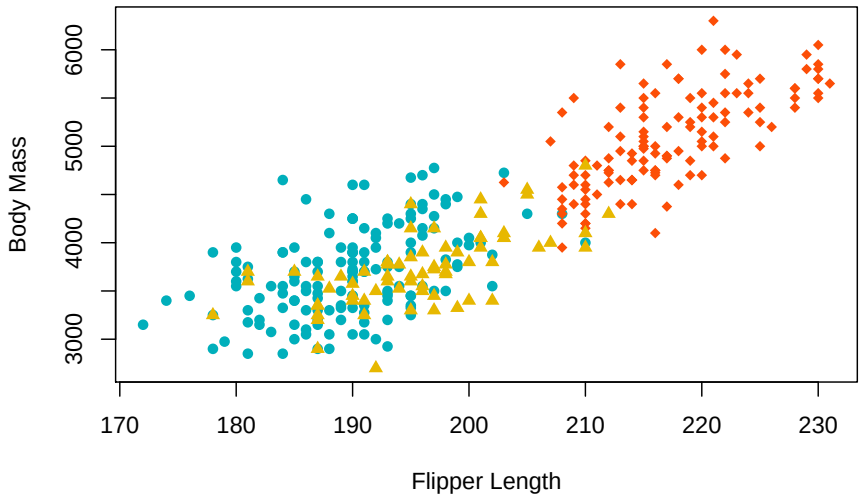
```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

10.2.1 base R package

```r
# Define color for each of the 3 penguine species
colors <- c("#00AFBB", "#E7B800", "#FC4E07")
colors <- colors[as.numeric(penguins$species)]

# Define shapes
shapes = c(16, 17, 18)
shapes <- shapes[as.numeric(penguins$species)]

plot(x = penguins$flipper_length_mm,
          y = penguins$body_mass_g,
          col = colors,
          pch = shapes,
          xlab = "Flipper Length",
          ylab = "Body Mass" )
```

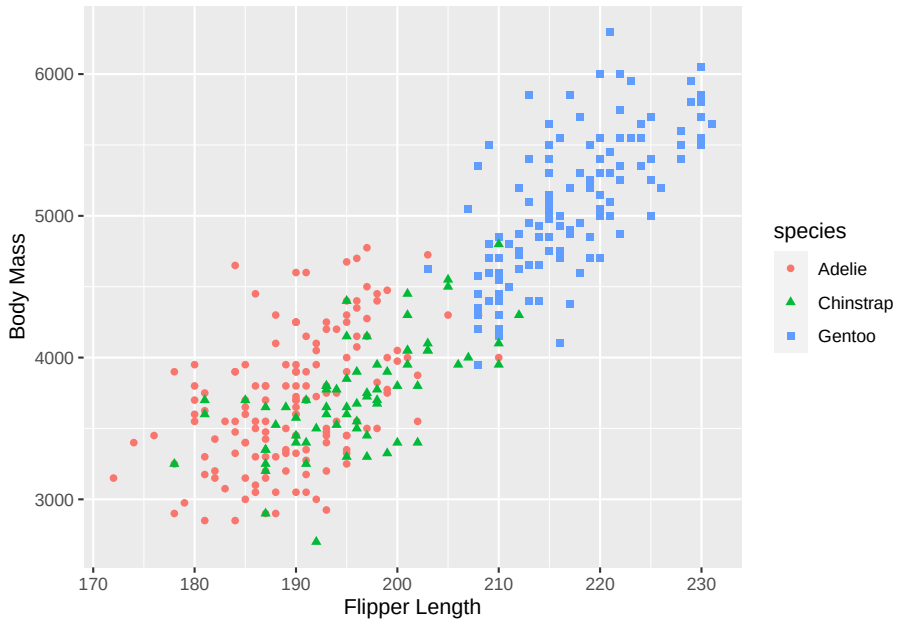


10.2.2 gggplot2 Package

ggplot2 is an R package dedicated to data visualization which is based on The Grammar of Graphics (Wilkinson, 2012).

```
#load ggplot2 package to make statistical graphics
library(ggplot2)
p <- ggplot(penguins) +
  geom_point( aes(x = flipper_length_mm,
                  y = body_mass_g,
                  color = species,
                  shape = species)) +
  xlab("Flipper Length")+
  ylab("Body Mass")

print(p)
```



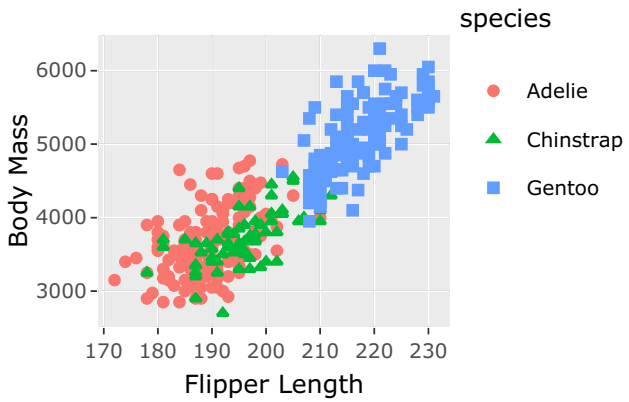
10.2.3 `plotly` R package for interactive data visualization

Interactive visualization focuses on graphic representations of data that improve the way we interact with information

plotly is an R package for creating interactive web-based graphs via the open source JavaScript graphing library plotly.js.

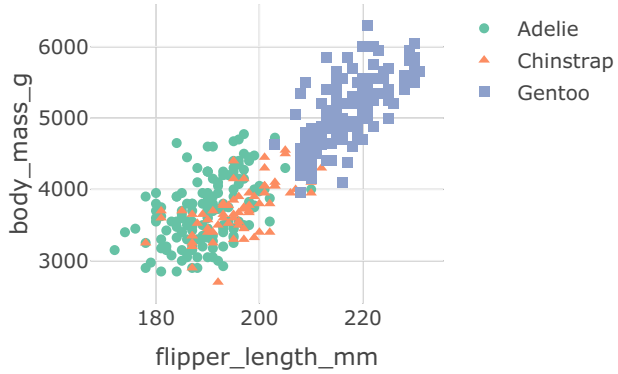
```
library(plotly)
p <- ggplot(penguins) +
  geom_point( aes(x = flipper_length_mm,
                  y = body_mass_g,
                  color = species,
                  shape = species)) +
  xlab("Flipper Length")+
  ylab("Body Mass")

# The function ggplotly converts a ggplot2::ggplot() object to a plotly object.
plotly::ggplotly(p)
```



Method 2

```
library(plotly)
fig <- plot_ly(penguins,
               x = ~flipper_length_mm,
               y = ~body_mass_g,
               color = ~species,
               symbol = ~species,
               type = "scatter")
fig
```



10.3 Python

Load data

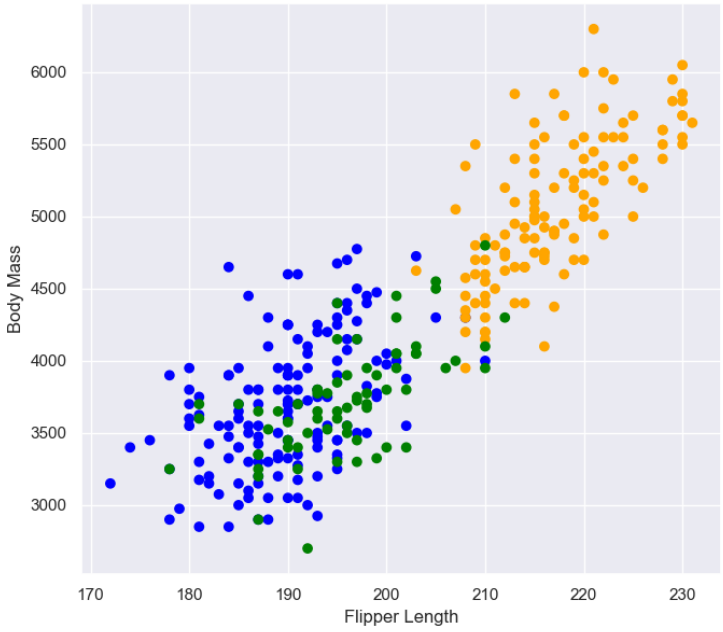
```
#load functions in palmerpenguins package
from palmerpenguins import load_penguins
penguins = load_penguins()
# Return the first part of the dataset
penguins.head()
# Retrieve column names
list(penguins.columns)
```

10.3.1 `Matplotlib` package

Matplotlib is mainly deployed for basic plotting. Visualization using Matplotlib generally consists of bars, pies, lines, scatter plots and so on.

```
# Import matplotlib to make statistical graphics.
# By convention, it is imported with the shorthand sns.
import matplotlib.pyplot as plt
```

```
colors = {'Adelie':'blue', 'Gentoo':'orange', 'Chinstrap':'green'}
plt.scatter(penguins.flipper_length_mm,
penguins.body_mass_g,
c= penguins.species.apply(lambda x: colors[x]))
plt.xlabel('Flipper Length')
plt.ylabel('Body Mass')
```

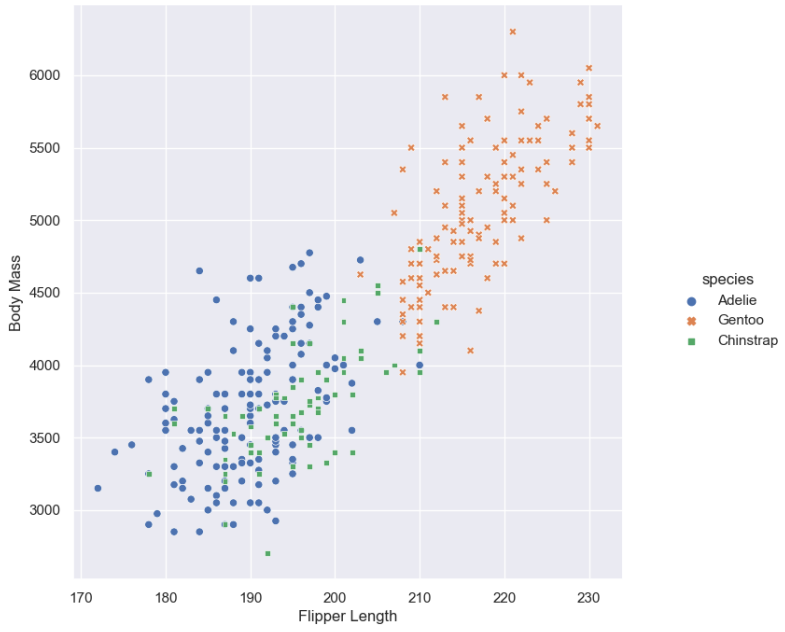


10.3.2 `seaborn` Package

Seaborn provides a variety of visualization patterns. It uses fewer syntax and has easily interesting default themes.

```
# Import seaborn to make statistical graphics.
# By convention, it is imported with the shorthand sns.
import seaborn as sns
#load functions in palmerpenguins package
from palmerpenguins import load_penguins
penguins = load_penguins()
```

```
# Apply the default theme
sns.set_theme()
# sns.set_style('whitegrid')
p = sns.relplot(x = 'flipper_length_mm',
            y ='body_mass_g',
            hue = 'species',
            style = 'species',
            data = penguins)
p.set_xlabels('Flipper Length')
p.set_ylabels('Body Mass')
```



The function `relplot()` is named that way because it is designed to visualize many different statistical relationships. The `relplot()` function has a convenient kind parameter that lets you easily switch to this alternate representation: `scatterplot()` with `kind="scatter"`; the default and `lineplot()` with `kind="line"`.

10.3.3 `plotnine` package

https://pypi.org/project/plotnine/

plotnine is an implementation of a grammar of graphics in Python, it is based on ggplot2. The grammar allows users to compose plots by explicitly mapping

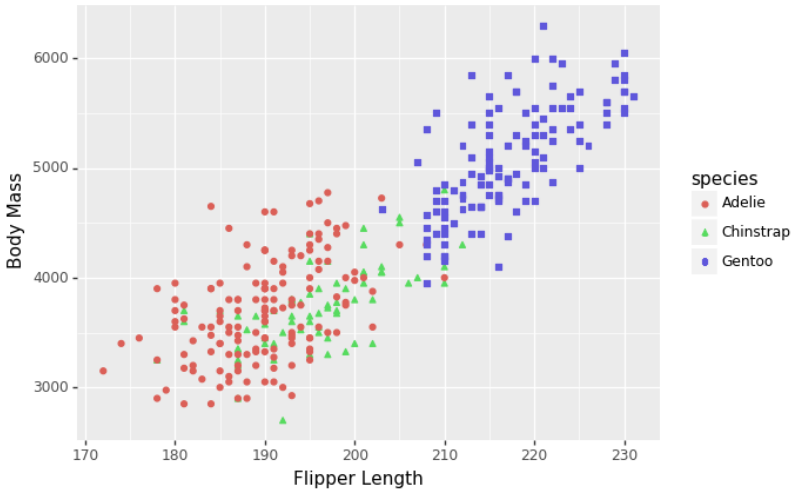
data to the visual objects that make up the plot.

Plotting with a grammar is powerful, it makes custom (and otherwise complex) plots are easy to think about and then create, while the simple plots remain simple.

NOTE: R vs Python Syntax

Unlike in R, now all the variables must be enclosed by single quotes

```
from plotnine import *
# unlike in R, now all the variables must be enclosed by single quotes
(ggplot(penguins) +
  geom_point(aes(x = 'flipper_length_mm',
                 y = 'body_mass_g',
                 color = 'species',
                 shape = 'species')) +
  xlab("Flipper Length")+
  ylab("Body Mass"))
```



10.3.4 `plotly` Python library for interactive data visualization

The `plotly.express` (Plotly Express or PX) module contains functions that can create entire figures at once. It is usually imported as `px`. Plotly Express is a built-in part of the `plotly` library.

```
import plotly.express as px

fig = px.scatter(penguins,
                 x="flipper_length_mm",
                 y="body_mass_g",
                 color= "species",
                 symbol= "species",
                 labels=dict(flipper_length_mm="Flipper Length",
                             body_mass_g="Body Mass"))
fig.show()
```

Chapter 11

Model

WIP

Chapter 12

Communicate

WIP

Chapter 13

Advanced R and Python

WIP

13.1 Time Series Forecasting

R	Python
fable-Forecasting Models for Tidy Time Series	statsmodels- Statistics based models
forecast- Forecasting Functions for Time Series and Linear Models	sktime- A unified framework for machine learning with time series GluonTS- Deep learning-based models.

Bibliography

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wikipedia contributors (2020a). Python (programming language) — Wikipedia, the free encyclopedia. [Online; accessed 25-December-2020].

Wikipedia contributors (2020b). R (programming language) — Wikipedia, the free encyclopedia. [Online; accessed 25-December-2020].

Wilkinson, L. (2012). The grammar of graphics. In *Handbook of computational statistics*, pages 375–414. Springer.