Getting familiar with R

Oct 1, 2021

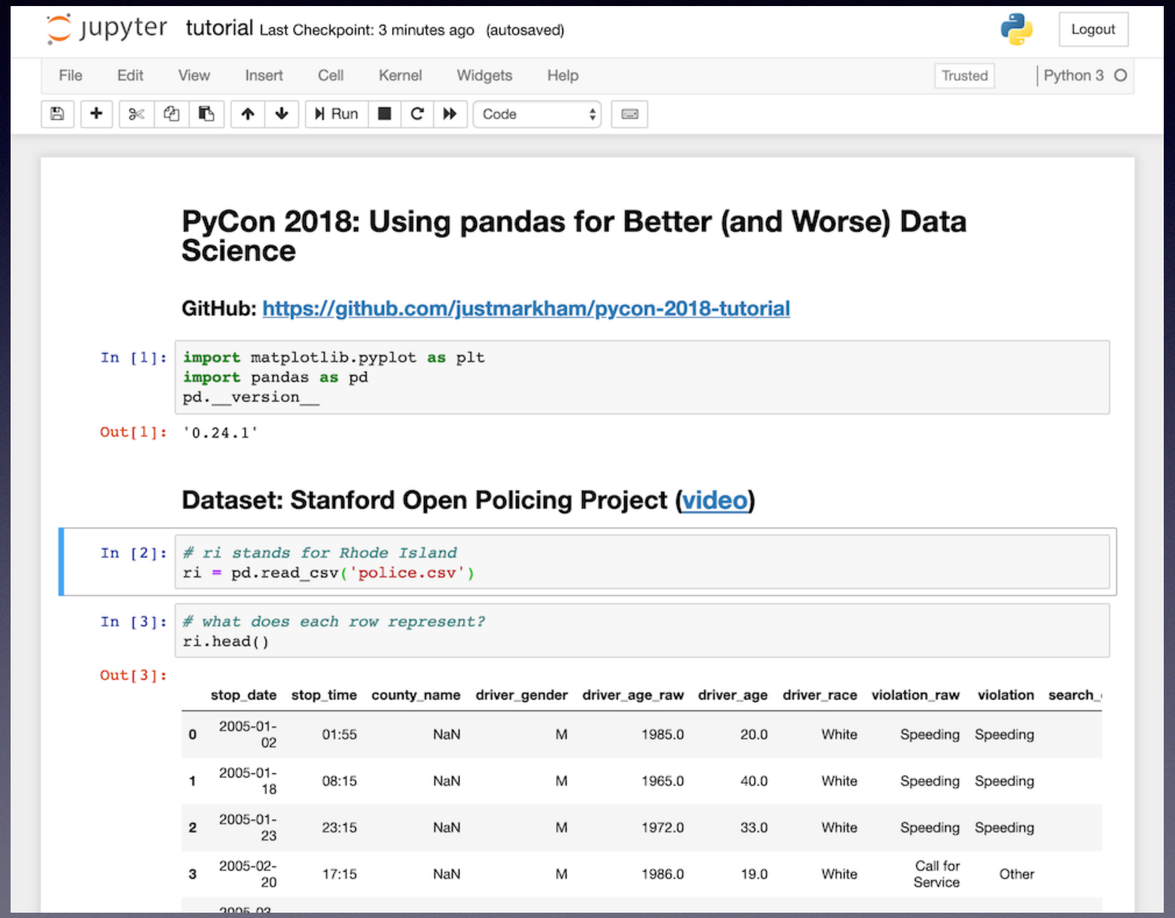
Qingyao
Baudis group

R v.s. Python

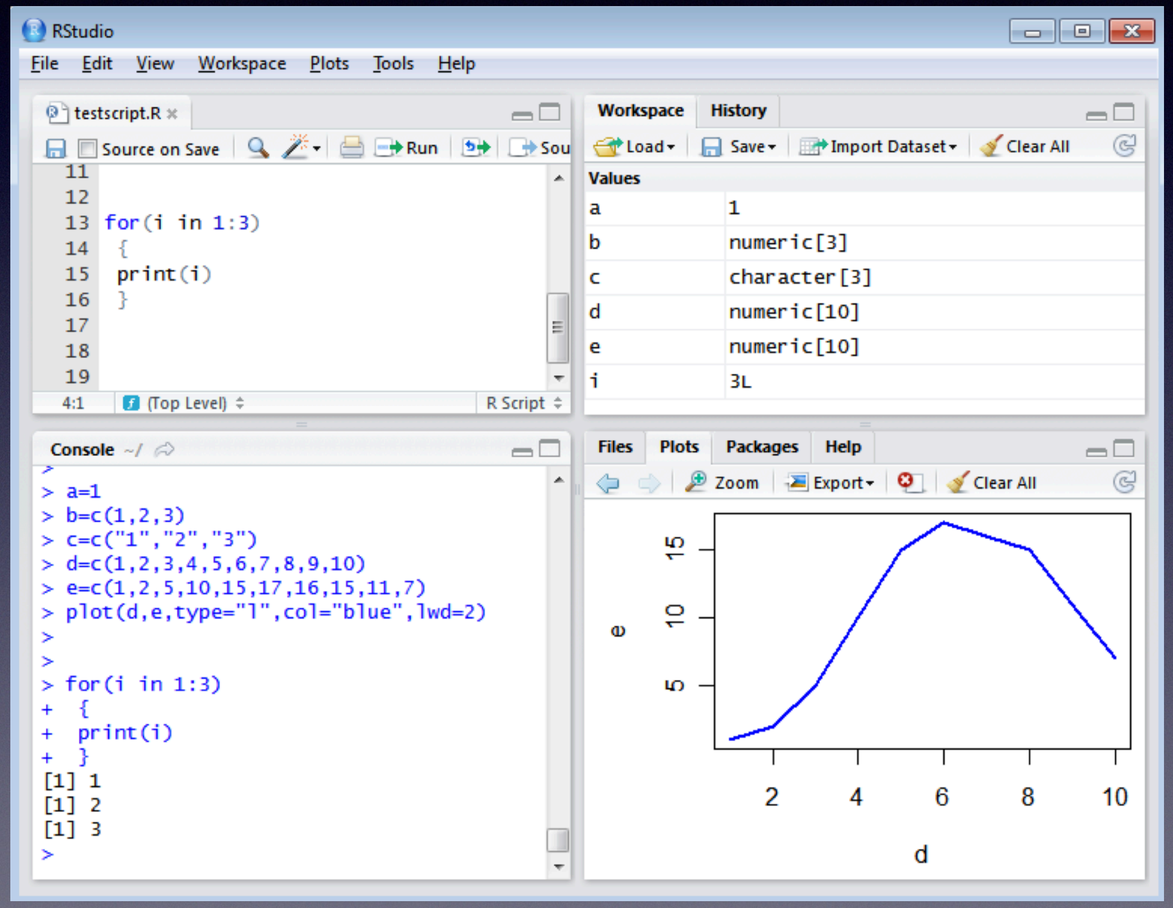
	Python	R
	popular, large community, library support	
Field	General-purpose	Finance, Healthcare
Usage	Web development, Machine learning, Scientific computing	Statistical modeling, Data visualisation
Advantage	Readable (indentation, English syntax) Unstructured data	Data frame! Exploratory visualization
Disadvantage	Slow	Unclear error message Slow

IDE

Python Jupiter notebook

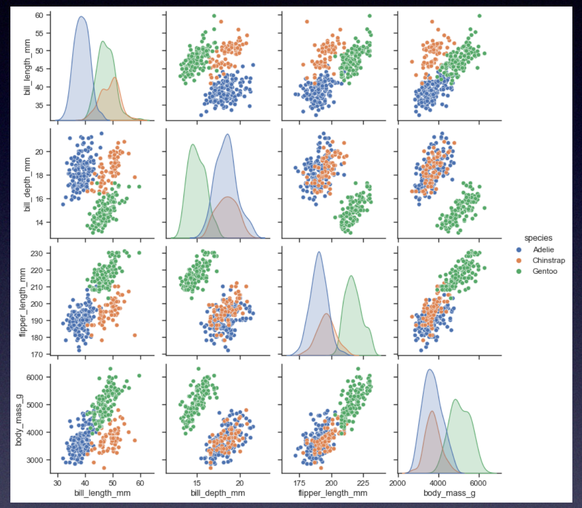


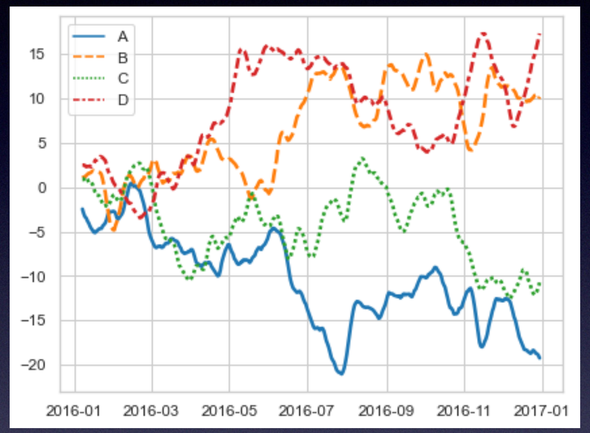
R RStudio

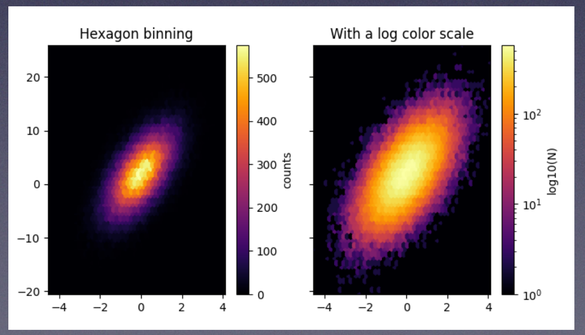


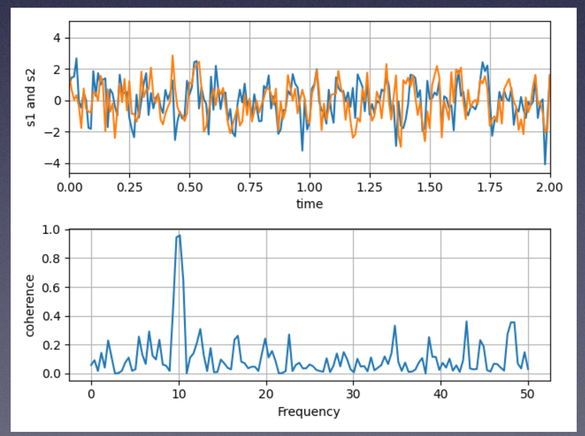
Data visualisation

Python seaborn, Matplotlib







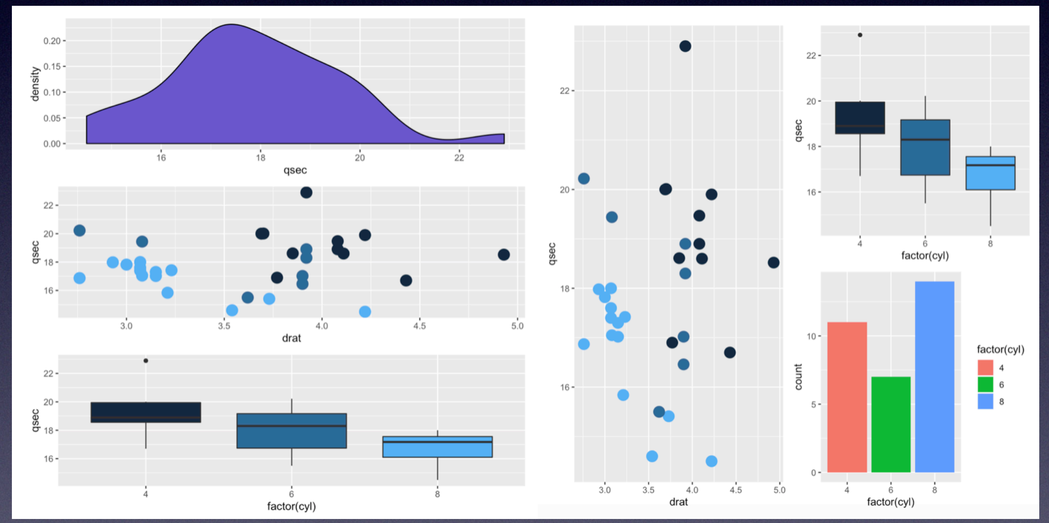


https://seaborn.pydata.org/examples/index.html

https://matplotlib.org/gallery/index.html

https://docs.bokeh.org/en/latest/index.html

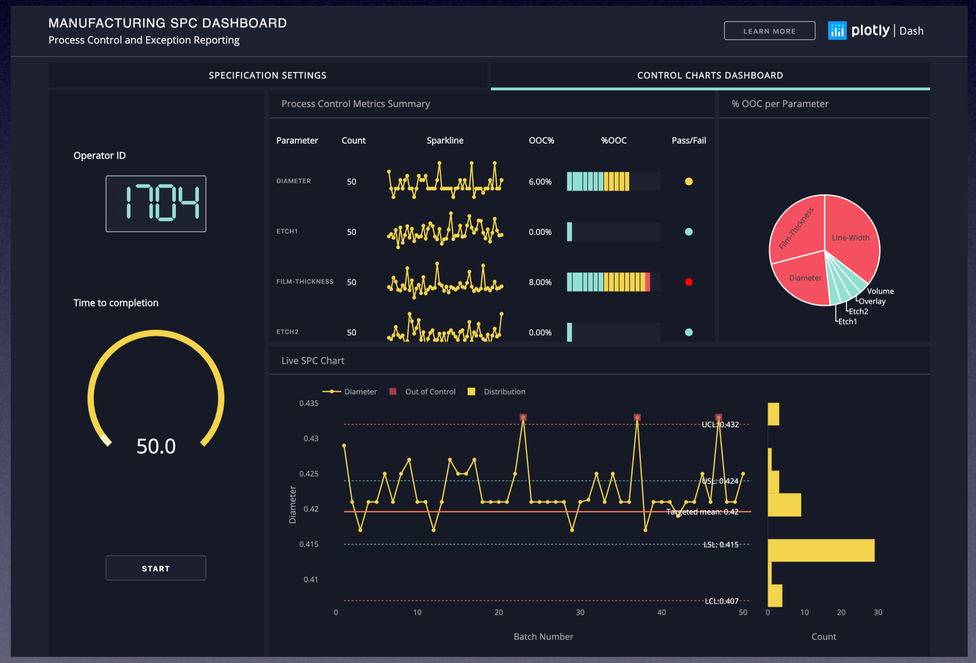
R ggplot2



https://www.r-graph-gallery.com/ggplot2-package.html

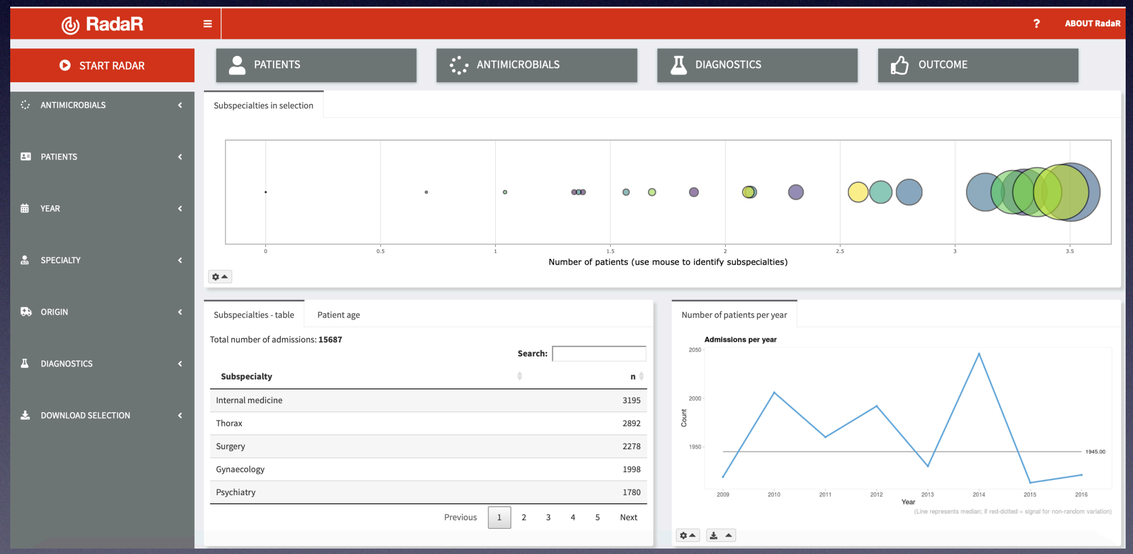
Web app

Python Dash



https://dash-gallery.plotly.host/dash-manufacture-spc-dashboard/

R Rshiny



https://shiny.rstudio.com/gallery/hospital-data-antimicrobial.html

Reading

A (very) short introduction to R (https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf)

edX (HarvardX) Data Science: R basics (https://courses.edx.org/courses/course-v1:HarvardX+PH125.1x+2T2020/course/)

HarvardX Class Note (https://rafalab.github.io/dsbook/getting-started.html)

Cheatsheets, e.g. ggplot2, tidyverse (https://rstudio.com/resources/cheatsheets/)

Basics

- Mathematical operators: + - * / ^
- Logical operator: == <= >= %in%
- Assignment operator: <- =
- Programming: *if* conditional, *for* loop
- Calculation: mean max median sum log
- Comment: #

Classes

- 0-D: character, numeric, logical
- 1-D: vector

```
v = c(1,2,3)
```

- 2-D: matrix, data frame

```
M = matrix(data = c(1,2,3,4), nrow = 2)
df = data.frame(M)
```

- flexible: list

Function

```
> SumTwoNumber <- function(a,b){
    return (a+b)
  }
> result <- SumTwoNumber(1,2)
> result
3
```

Import data

- read.table(path_to_file)
- read_delim(path_to_file) #readr library

Use libraries

- install.packages(package_name)
- library(package_name)

Data frame

- nrow, ncol, dim
- colnames, rownames
- df[:,2], df$Value
- apply, sapply, lapply

df

	Name	Value
1	John	20
2	Alice	45

Plotting

```
v1 = rnorm(100); v2 = rnorm(100)+1
```

- hist(v1) plot(x=v1, y=v2) #basic
- ggplot(data_frame_name) + geom_*xxx*() #ggplot2 library
  - bar, histogram, boxplot, violin, line

Cheatsheet Here! →

https://github.com/rstudio/cheatsheets/raw/master/data-visualization.pdf