# Getting familiar with R

Sep 25, 2020

Qingyao
Baudis group

# About me (Qingyao)

- PhD student in Baudis group
- Background:
  - biochemistry BSc
  - molecular medicine MSc
- Current area of work
  - Cancer genomics & Population genetics
  - Database restructuring
  - Curation, text mining, data standard, sharing

# R v.s. Python

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

# IDE

## Python Jupiter notebook

jupyter tutorial Last Checkpoint: 3 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

### PyCon 2018: Using pandas for Better (and Worse) Data Science

**GitHub:** https://github.com/justmarkham/pycon-2018-tutorial

In [1]:
```
import matplotlib.pyplot as plt
import pandas as pd
pd.__version__
```

Out[1]: '0.24.1'

### Dataset: Stanford Open Policing Project (video)

In [2]:
```
# ri stands for Rhode Island
ri = pd.read_csv('police.csv')
```

In [3]:
```
# what does each row represent?
ri.head()
```

Out[3]:

| | stop_date | stop_time | county_name | driver_gender | driver_age_raw | driver_age | driver_race | violation_raw | violation | search_ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2005-01-02 | 01:55 | NaN | M | 1985.0 | 20.0 | White | Speeding | Speeding | |
| 1 | 2005-01-18 | 08:15 | NaN | M | 1965.0 | 40.0 | White | Speeding | Speeding | |
| 2 | 2005-01-23 | 23:15 | NaN | M | 1972.0 | 33.0 | White | Speeding | Speeding | |
| 3 | 2005-02-20 | 17:15 | NaN | M | 1986.0 | 19.0 | White | Call for Service | Other | |

## R RStudio

RStudio

File Edit View Workspace Plots Tools Help

testscript.R

Source on Save Run Sou

```
11
12
13 for(i in 1:3)
14  {
15  print(i)
16  }
17
18
19
```

4:1 (Top Level) R Script

Workspace History

Load Save Import Dataset Clear All

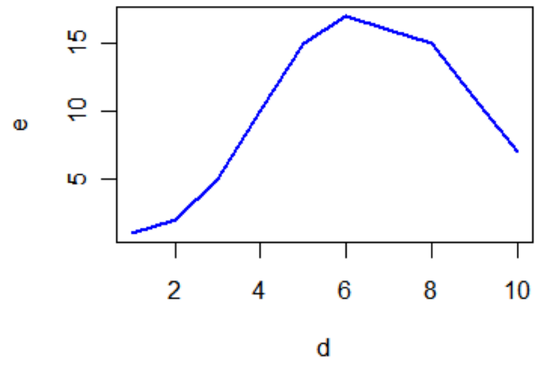
| Values | |
|---|---|
| a | 1 |
| b | numeric[3] |
| c | character[3] |
| d | numeric[10] |
| e | numeric[10] |
| i | 3L |

Console ~/

```
> a=1
> b=c(1,2,3)
> c=c("1","2","3")
> d=c(1,2,3,4,5,6,7,8,9,10)
> e=c(1,2,5,10,15,17,16,15,11,7)
> plot(d,e,type="l",col="blue",lwd=2)
>
>
> for(i in 1:3)
+  {
+  print(i)
+  }
[1] 1
[1] 2
[1] 3
>
```

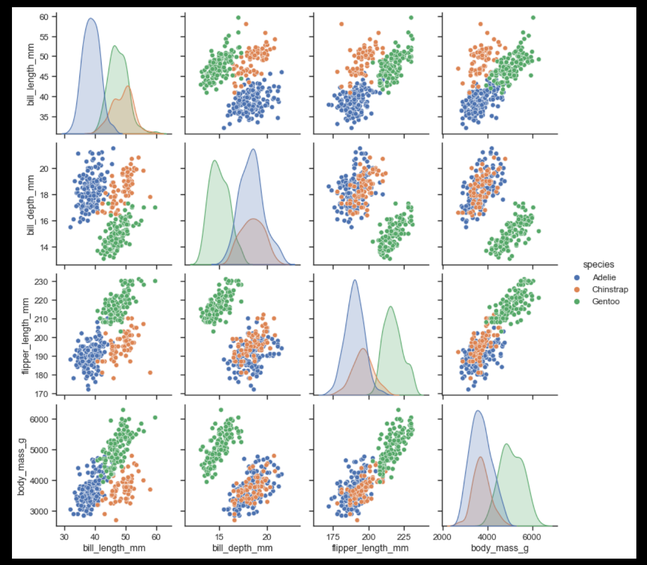
Files Plots Packages Help

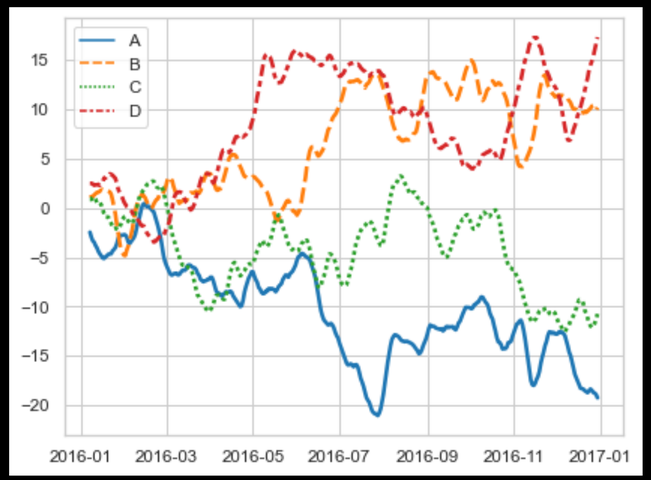
Zoom Export Clear All

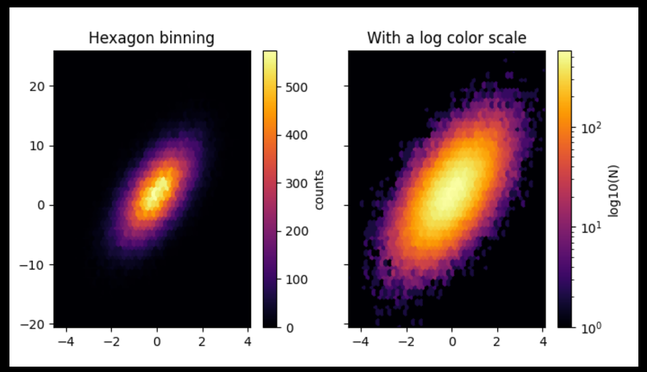


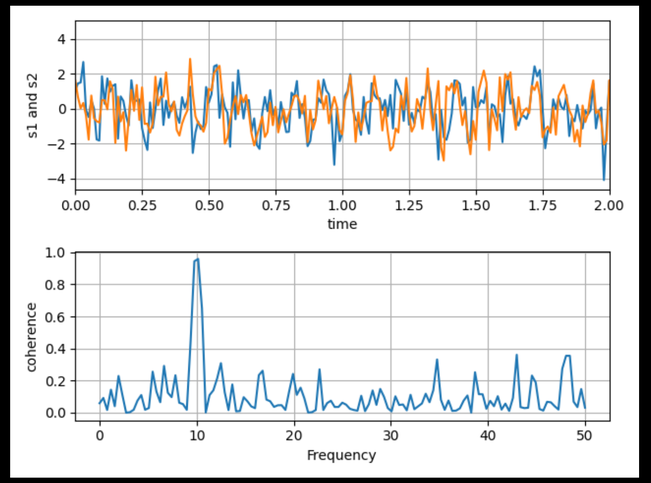
# Data visualisation

## Python seaborn, Matplotlib



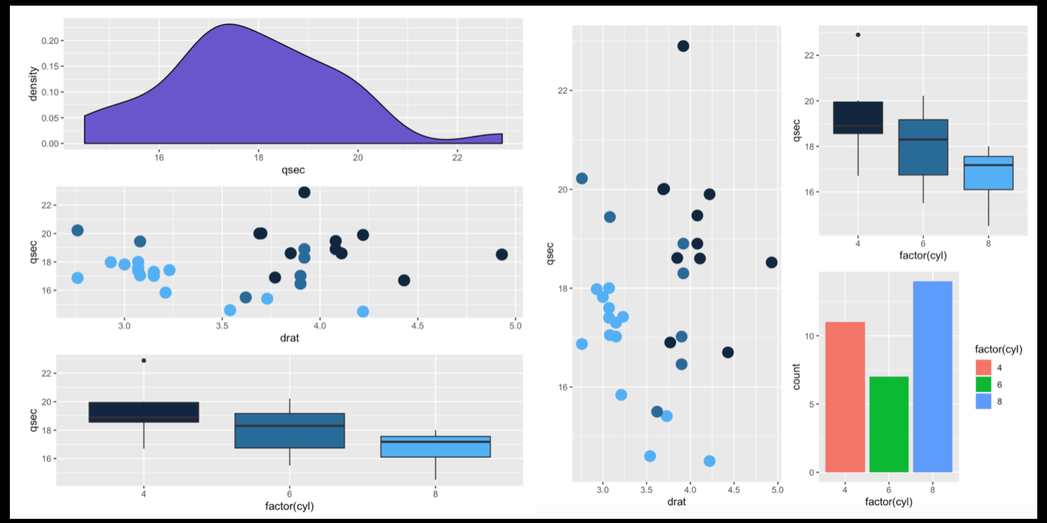




https://seaborn.pydata.org/examples/index.html
https://matplotlib.org/gallery/index.html
https://docs.bokeh.org/en/latest/index.html

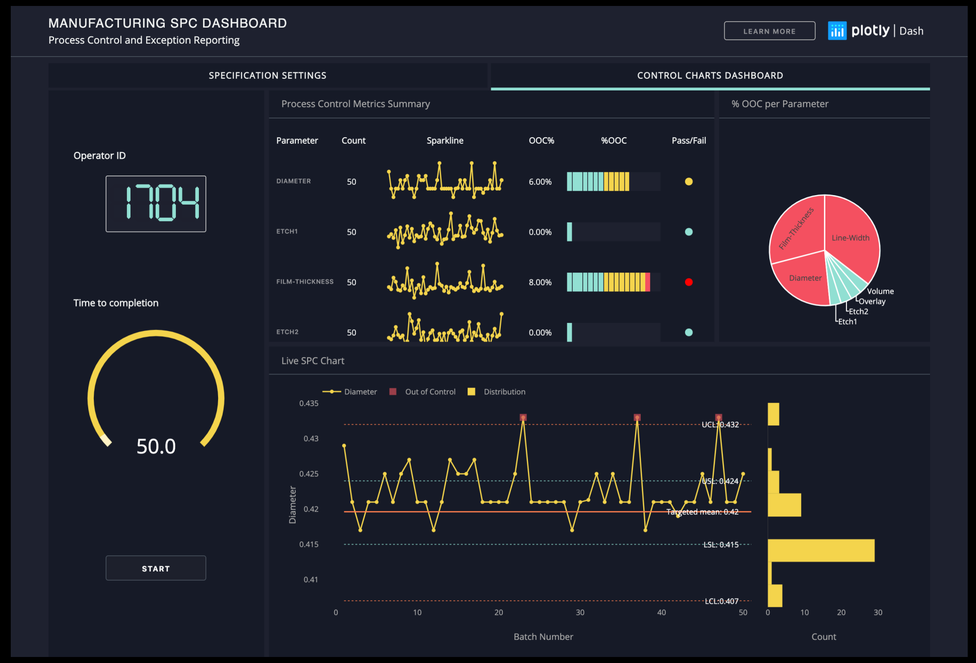
## R ggplot2




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

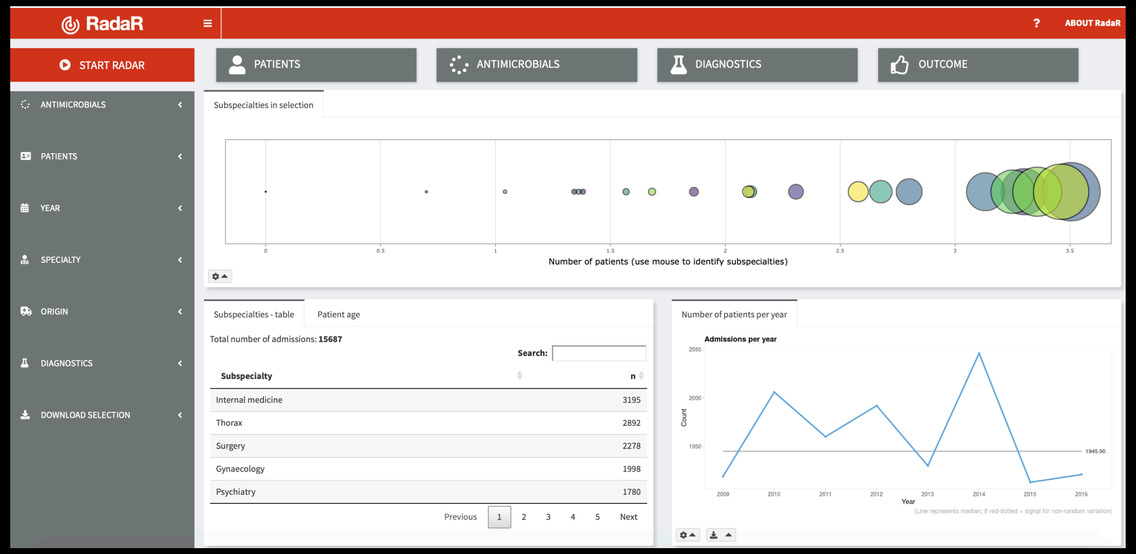
# Web app

## Python Dash




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

## R Rshiny




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

# 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/)

# Before next week

GATK workshop

- Register at app.terra.bio
- Send your email address to MS Teams page

# 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 c(1,2,3)
- 2-D: matrix, data frame
- 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[:,1], df$Value
- apply, sapply, lapply

**df**

| | Name | Value |
|---|---|---|
| 1 | John | 20 |
| 2 | Alice | 45 |

# Plotting

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