Pythonでggplotってみた話

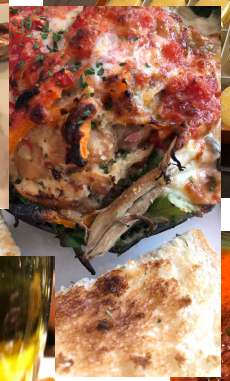
15th August 2020, Pythonもくもく自習室3周年記念LT大会

Yuta Kanzawa @yutakanzawa

Data Science Senior Analyst at Janssen Pharmaceutical K.K., Tokyo
A Family Company of Johnson & Johnson

3周年おめでとうございます:)))










I am...

- **神沢雄大 Yuta Kanzawa** (twitter: @yutakanzawa)
- Data scientist at Janssen Japan, Tokyo
  - A pharmaceutical company of **J&J**
- Opera & wine lover
  - Wagner
  - Bourgogne
- 7 languages
  - Human: Japanese, English, German
  - Computer: R, Python, SAS, SQL

（宣伝）今年のPyCon JPのスポンサーをします！

Johnson & Johnson
FAMILY OF COMPANIES IN JAPAN

ジョンソン・エンド・ジョンソン日本法人グループ

日本最大級のPythonイベント「PyCon JP 2020」 協賛のお知らせ

2020年8月14日

ジョンソン・エンド・ジョンソン日本法人グループ[*1]（本社：東京都千代田区、以下ジョンソン・エンド・ジョンソン）は、2020年8月28日（金）、29日（土）に開催される日本最大級のPython（パイソン）のイベント「PyCon（パイコン） JP 2020 」に協賛します。

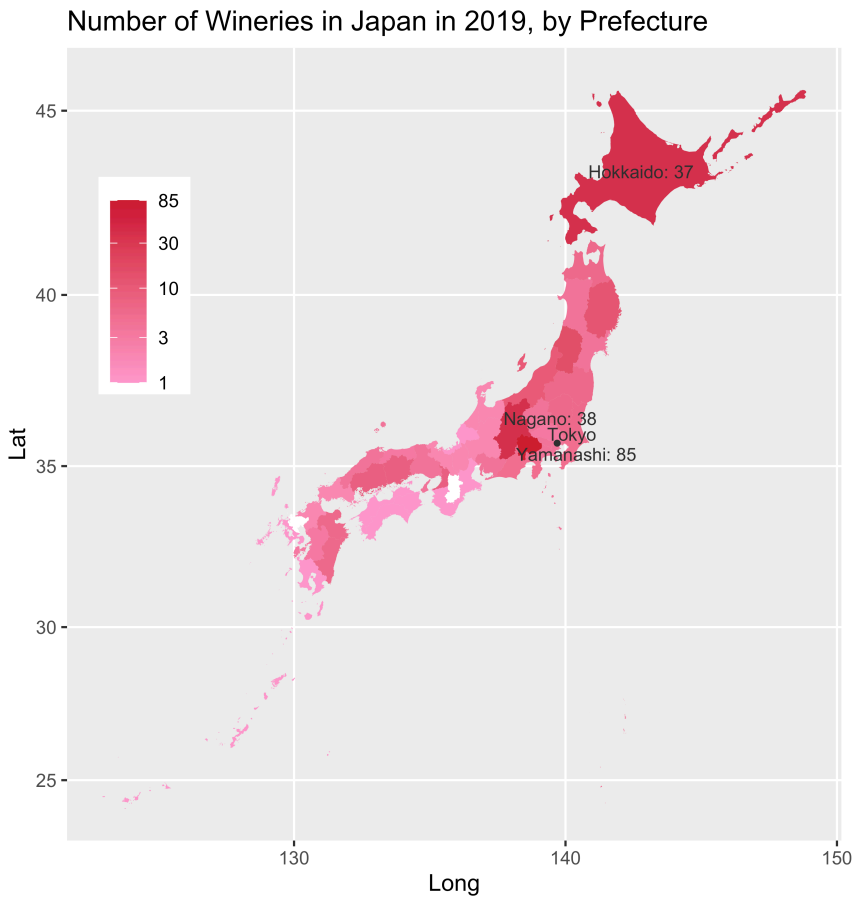
「Python（パイソン）」は、データサイエンスの分野で広く用いられるプログラミング言語の一つで、ジョンソン・エンド・ジョンソンのデータサイエンスチームも採用しています。

ヘルスケア業界においても、テクノロジーの積極的な活用とデータに基づく意思決定の重要性は日々増加しており、ジョンソン・エンド・ジョンソンは、日本をはじめ、世界各国でデータサイエンスを活用した製品開発やビジネス活動を進めています。

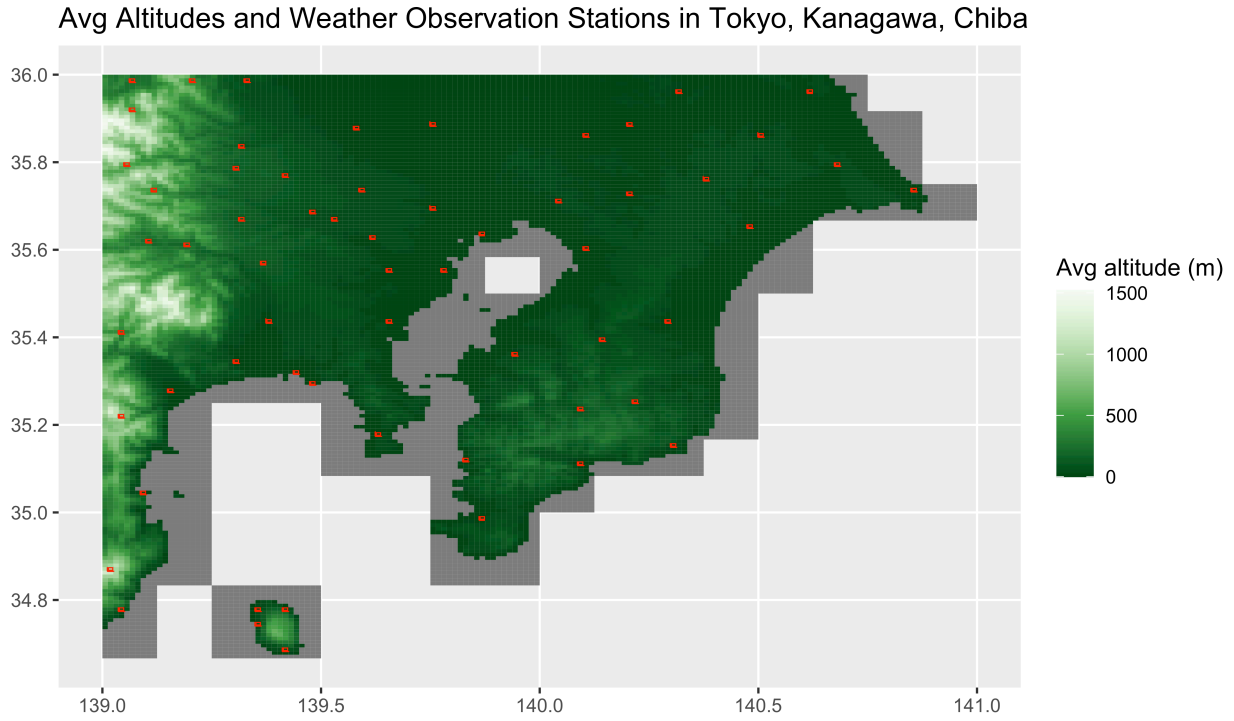
イベント当日は、一般消費者から患者さん、医療従事者まで幅広い顧客ニーズに応えるジョンソン・エンド・ジョンソン日本法人グループのデータサイエンス担当者が、それぞれの事業分野での具体的事例や、業務紹介を行います。ぜひご参加ください。

* https://www.jnj.co.jp/media-center/press-releases/20200814

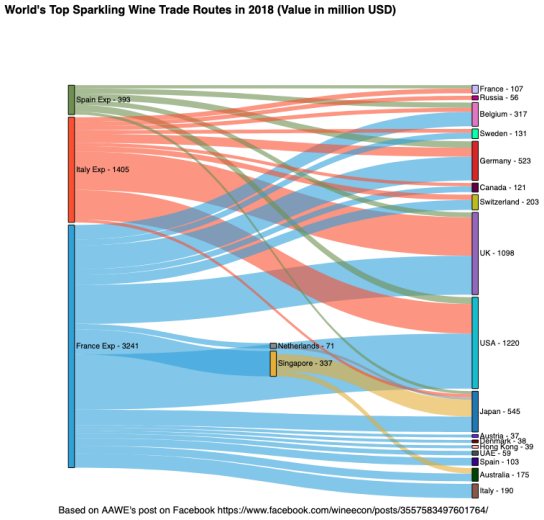
ポートフォリオ（Rでは最近は地図が多い）

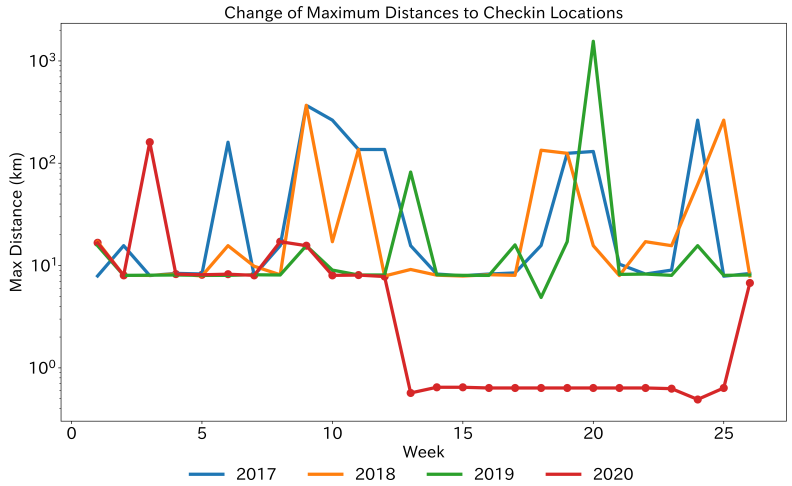


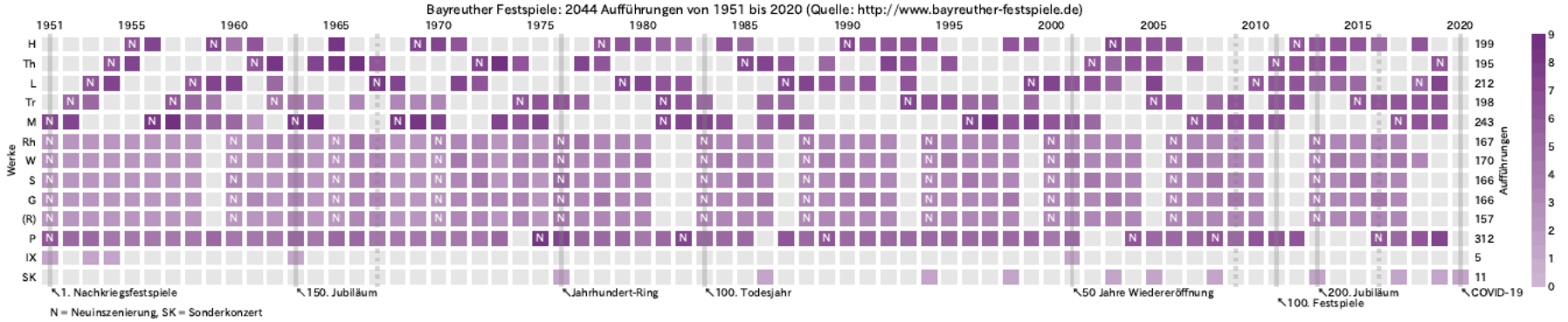
Source: https://www.nta.go.jp/taxes/sake/shiori-gaikyo/seizogaikyo/kajitsu/pdf/h30/30wine01.pdf



ポートフォリオ（Pythonも）







アジェンダ

- 今日話すこと
  - ggplot2
  - plotnine

- 今日話さないこと
  - Rそのもの

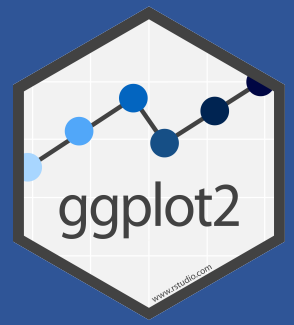
TL;DR

- ggplot2でのグラフ描画
  - グラフの要素をそれぞれ指定していく。→レイヤー
  - ggplot()：全レイヤーに関わる要素を指定
  - aes()*：見映えの要素となる変数
  - geom関数：グラフの種類

- plotnine
  - Pythonによるggplot2の実装

* 日本語では「エステティック」と表されることもある。

ggplot2入門

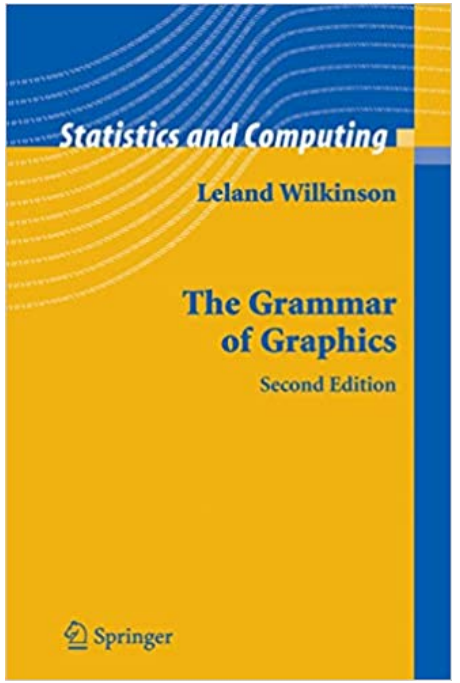


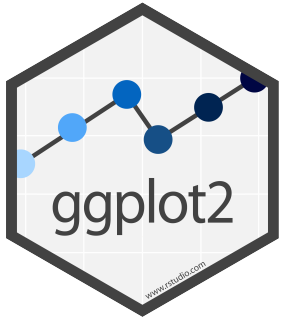
ggplot2 101

預言の書：‘The Grammar of Graphics’*1

- 告解：ちゃんと読んだことはありません…

- ggplot2の哲学的土台
  - グラフとは何か？
  - グラフ作成の基本的ルール
  - →Hadley Wickhamがコードで実装。
    - ‘A layered grammar of graphics’*2





*1 https://www.amazon.co.jp/Grammar-Graphics-Statistics-Computing/dp/0387245448/
*2 https://vita.had.co.nz/papers/layered-grammar.html

グラフの内部構造としてのレイヤー

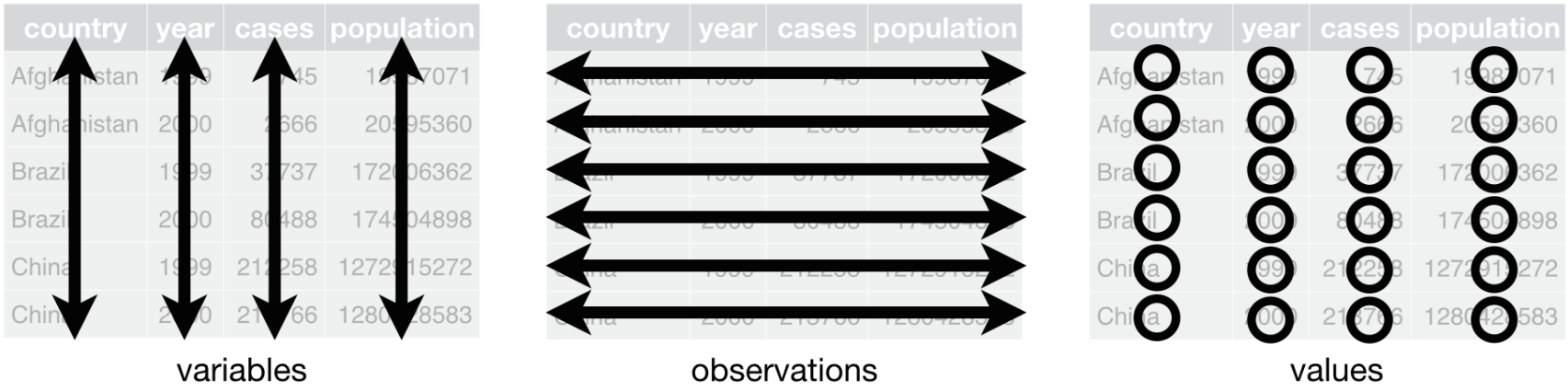
- ‘A layered grammar of graphics’
- 参考
  - ‘Making the complex simple in data viz’*
    - T. Vasilikioti, PyCon DE & PyData Berlin 2019

- ggplot2の原理
  - 表現の層（レイヤー）を重ねてグラフを描く。
  - 層ごとに異なる役割

* https://www.youtube.com/watch?v=pwzsGHjTDa4

入力データの形

- ‘tidy’なデータセット
  - ここでは説明を省略。
  - matplotlibに最適な形とは異なることがある。



* https://r4ds.had.co.nz/tidy-data.html#fig:tidy-structure

ggplot2の基本的用例（構文）*

- (1) 最初に、`ggplot()`を呼ぶ。`+`でつないでいく。
  - 元になるデータセットを指定し、「見映えの要素」となる変数を`aes()`に指定。
- それに加え、(2) グラフの種類 `geom_...()`
  - 例：散布図、棒グラフ、折れ線グラフ
- 必要に応じて以下も。
  - (3) scale関数：カラーパレット、凡例、軸
  - (4) ファセット：グループごとに描き分け
  - (5) 座標系：座標反転
- 画像として保存：`ggsave()`

例：x軸がログスケールの散布図

```
ggplot(data, aes(…)) +
  geom_point() +
  scale_colour_brewer(…) +
  scale_x_log10()
```

* https://ggplot2.tidyverse.org/

aes()（エステティックマッピング）

- 見映えに関わる「変数」を指定
  - x軸、y軸の値
  - グループごとの塗り分け*1：colour, fill
  - 点のサイズ（バブルチャート）：size
  - 色の濃淡度合い：alpha
- ポイント
  - ggplot()に指定するのがよいが、geom関数でもよい。
    - 宇宙本*2p.131参照（データやマッピングの継承）
  - 引数の値が定数のときは、aes()に入れない。

例：世界の都市の緯度と気温

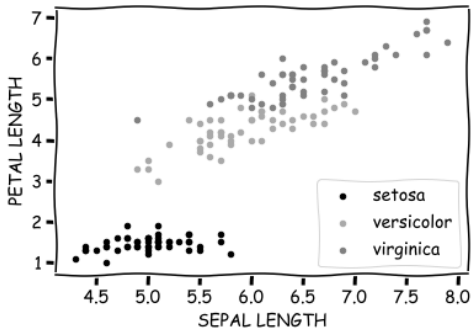
```
aes(x = latitude,
    y = temperature,
    colour = region,
    size = population)
```

*1 塗る色を指定する訳ではない。→ scale関数
*2『RユーザのためのRStudio入門 tidyverseによるモダンな分析フローの世界』（松村、湯谷、紀ノ定、前田、2018年）

例：アヤメのデータセット

- ボス：大至急、萼と花弁の長さの関係を種ごとに示せ！

- 私：（そうだ、散布図を描こう！）
  - x軸：萼の長さ
  - y軸：花弁の長さ
  - 種に応じて各点を色付けする。

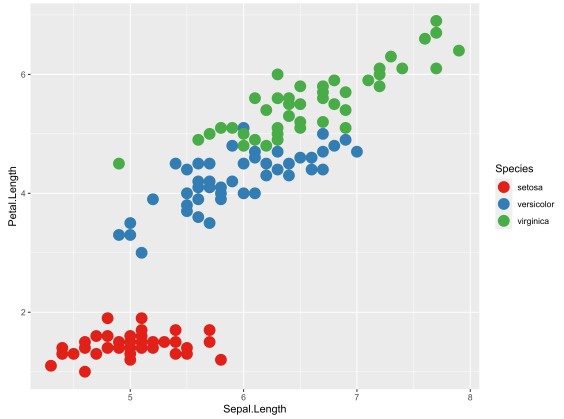


* https://xkcd.com/2207/
* Matplotlib’s xkcd style: https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.xkcd.html

コーディング例

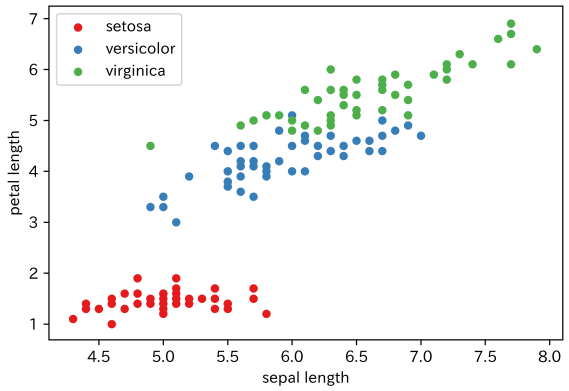
ggplot2

```
library(ggplot2)
ggplot(iris, aes(x = Sepal.Length,
                 y = Petal.Length,
                 colour = Species)) +
  geom_point(size = 5) +
  scale_colour_brewer(palette = "Set1")
```



matplotlib*

```
import matplotlib.pyplot as plt
fig, ax = plt.subplots()
cmap = plt.get_cmap('Set1')
for i, (key, df) in enumerate(iris.groupby('species')):
    df.plot.scatter(x='sepal length', y='petal length',
                    ax=ax, color=cmap(i), label=key, s=20)
ax.legend()
plt.show()
```

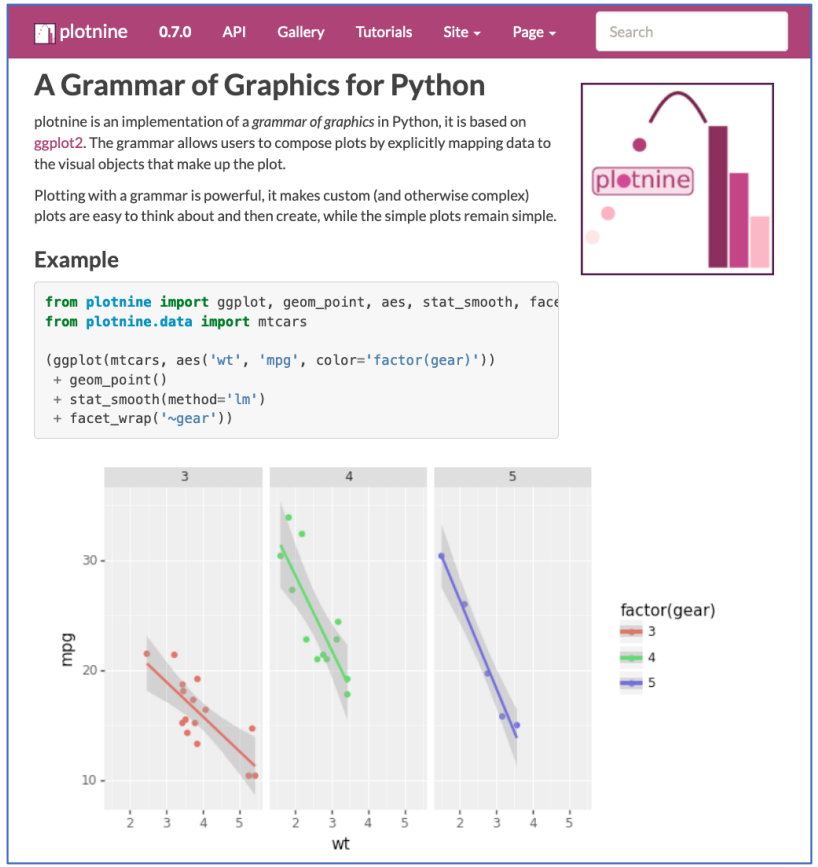


* scikit-learnのirisデータセットを事前に加工した。

plotnine

plotnine

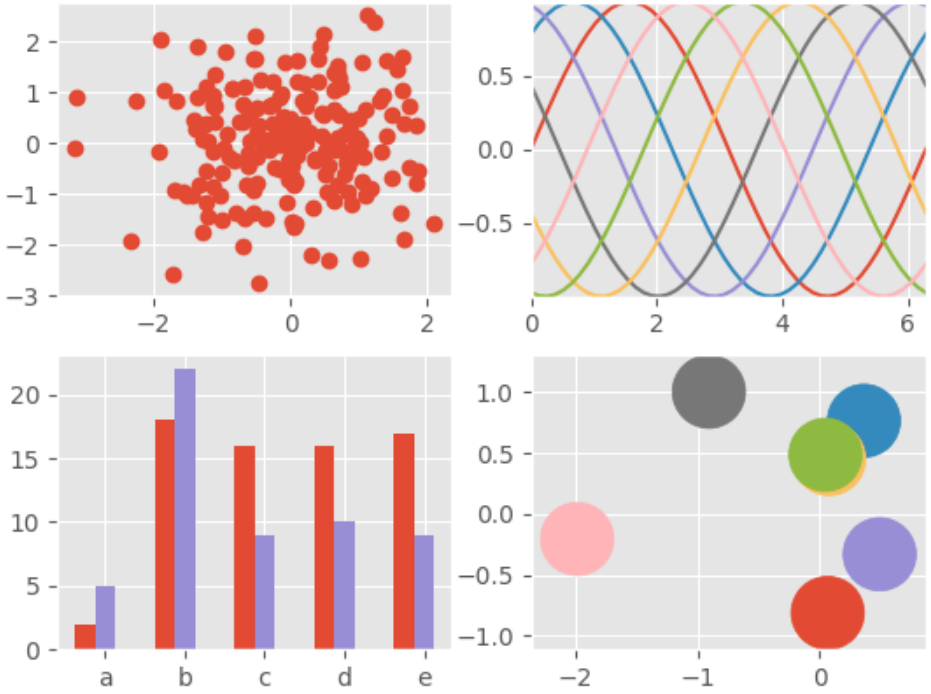
ちょっと前に知ったこと：



* https://twitter.com/medi_data0826/status/1289338732349255682
* https://plotnine.readthedocs.io/en/stable/

ちなみに、Pythonでggplotと言えば...

- matplotlibのggplot style sheet
  - `plt.style.use('ggplot')`
  - あくまでも出力がggplot風になるだけ。
    - 特徴的なライトグレーの背景
  - 書くコードはmatplotlib



* https://matplotlib.org/gallery/style_sheets/ggplot.html

plotnineとは？

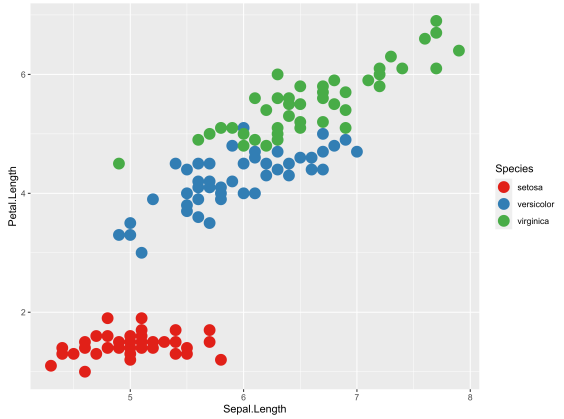
- （公式サイトの説明）
  plotnine is an implementation of a *grammar of graphics* in Python, it is based on ggplot2. The grammar allows users to compose plots by explicitly mapping data to the visual objects that make up the plot.

- matplotlibベース

使ってみた

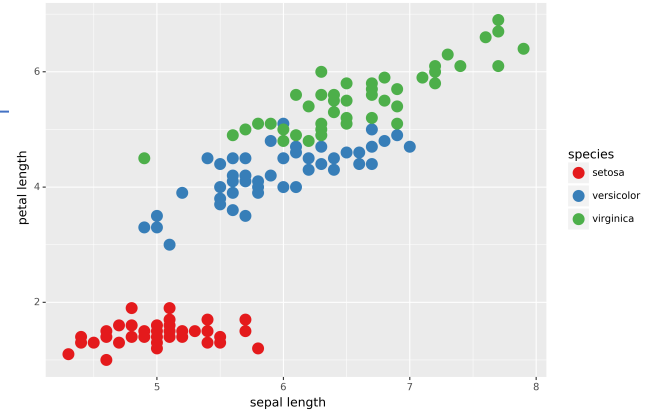
ggplot2

```
library(ggplot2)
ggplot(iris, aes(x = Sepal.Length,
                 y = Petal.Length,
                 colour = Species)) +
  geom_point(size = 5) +
  scale_colour_brewer(palette = "Set1")
```



plotnine*

```
from plotnine import ggplot, geom_point, scale_colour_brewer, aes
iris_plot = (
    ggplot(iris, aes(x="sepal length", y="petal length",
                     colour="species")) +
        geom_point(size=5) +
        scale_colour_brewer(type="qual", palette="Set1")
)
iris_plot
```



* scikit-learnのirisデータセットを事前に加工した。

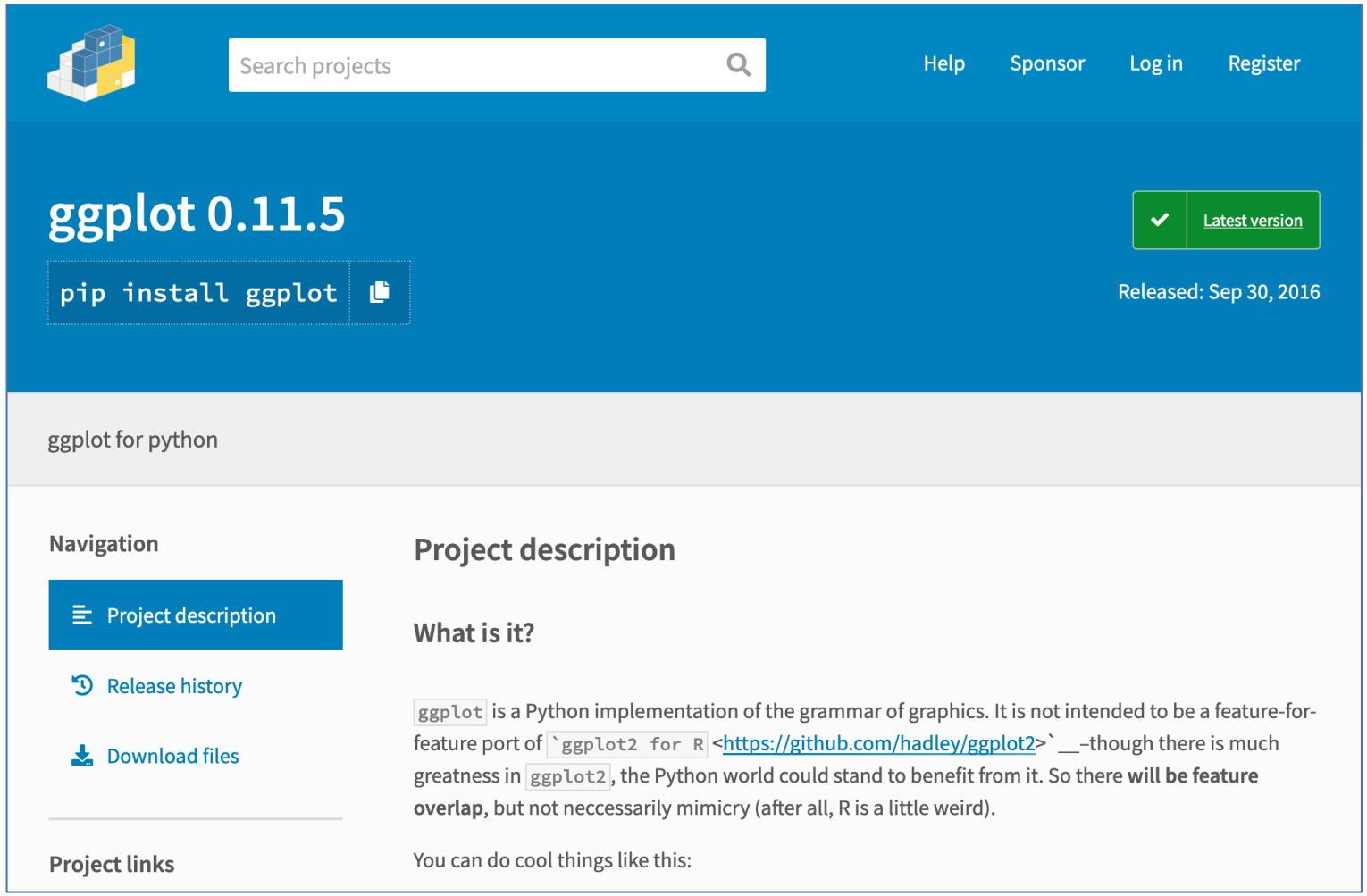
感想

- 良い点
  - ggplot2とほぼ同じ文法で書ける。
  - 出力もそっくり。
  - matplotlib, plotly以外の選択肢ができる。

- これからに期待する点
  - カラーパレット
  - 引数の型
  - pandasとの連携（難しいかな...）

おまけ

- ggplot
  - 最後は4年前…



まとめ

Long story short

Long story short

- ggplot2でのグラフ描画
  - グラフの要素をそれぞれ指定していく。→レイヤー
  - ggplot()：全レイヤーに関わる要素を指定
  - aes()*：見映えの要素となる変数
  - geom関数：グラフの種類

- plotnine
  - Pythonによるggplot2の実装
  - 便利なので（Rな人もPyな人も）試してみてほしい。

* 日本語では「エステティック」と表されることもある。

Enjoy!