통계와 시각화

- 통계(Statistic)과 통계량(Statistic)
- 시각화와 통계
- Pandas와 통계

작성자: sanggoo cho

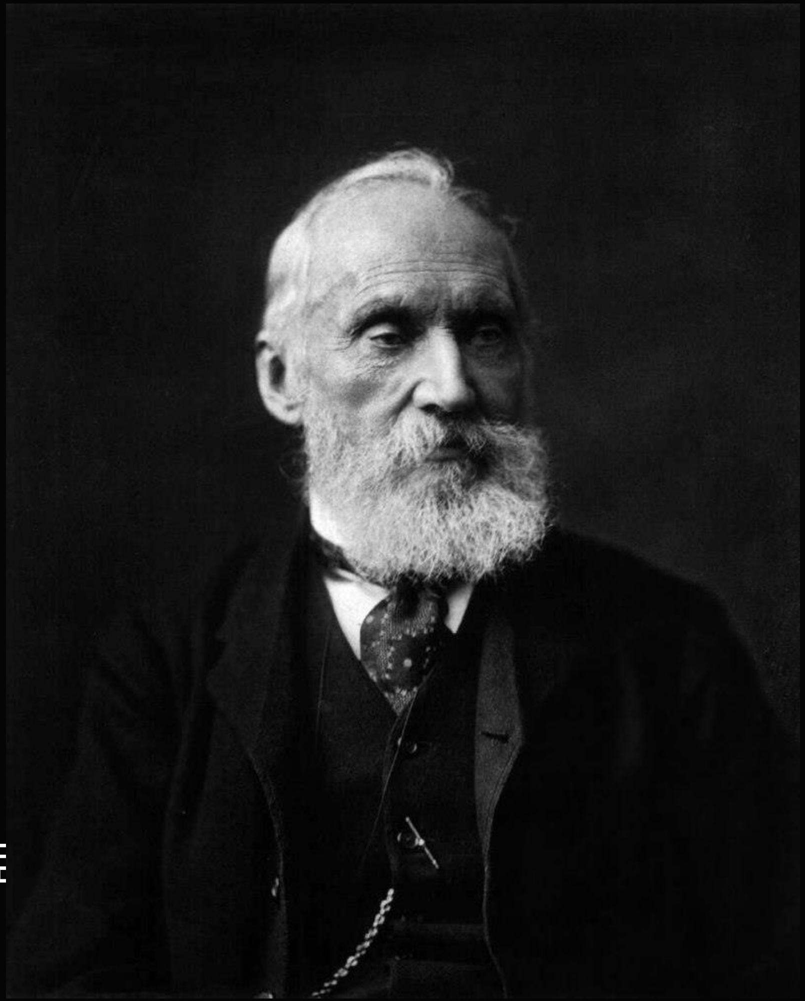
데이터 과학(Data Science)

“

저는 종종 이렇게 말합니다.
당신이 말하고 있는 것을 측정할 수 있고,
그것을 숫자로 표현할 수 있다면,
당신은 그 주제에 대해 뭔가를 알고 있는 것입니다.

하지만 그것을 측정할 수 없고, 숫자로 표현할 수 없다면,
당신의 지식은 불완전하고 미흡한 것입니다.
그것은 지식의 시작일 수 있지만, 아직 그 문제를 과학의 단계로
충분히 발전시키기에는 이르다고 할 수 있습니다.

"



Kelvin

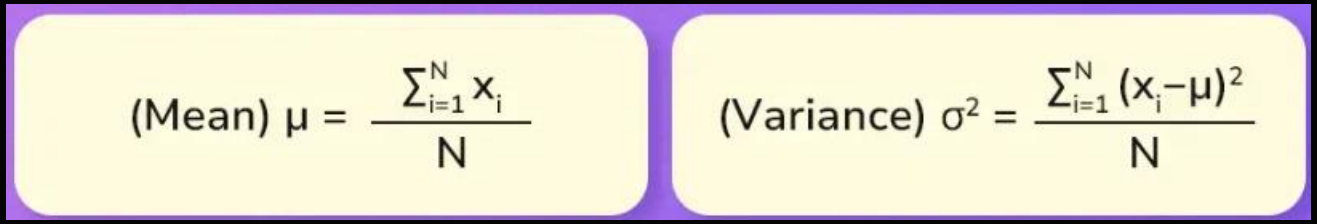
- 統計(합칠 통, 셀 계)
- Data Aggregation

통계(Statistics)

- ❖ 산술적 방법을 기초로 하여, 주로 다량의 데이터를 관찰하고 정리 및 분석하는 방법을 연구하는 수학의 한 분야

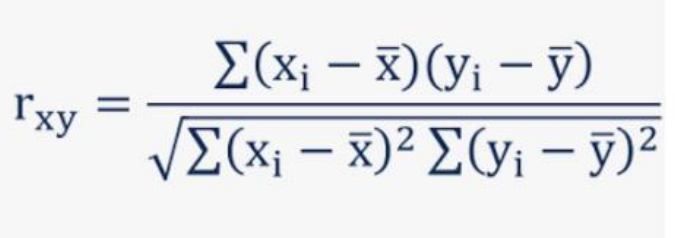
- ❖ 모집단(Population)을 대표하는 표본(Sample)의 평균, 분산 등의 통계량(Statistic)을 바탕으로 모집단을 기술(Description)하거나 추론(Inference)하는 것

통계와 시각화(Data Visualization)



$$\text{(Mean) } \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{(Variance) } \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$



$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

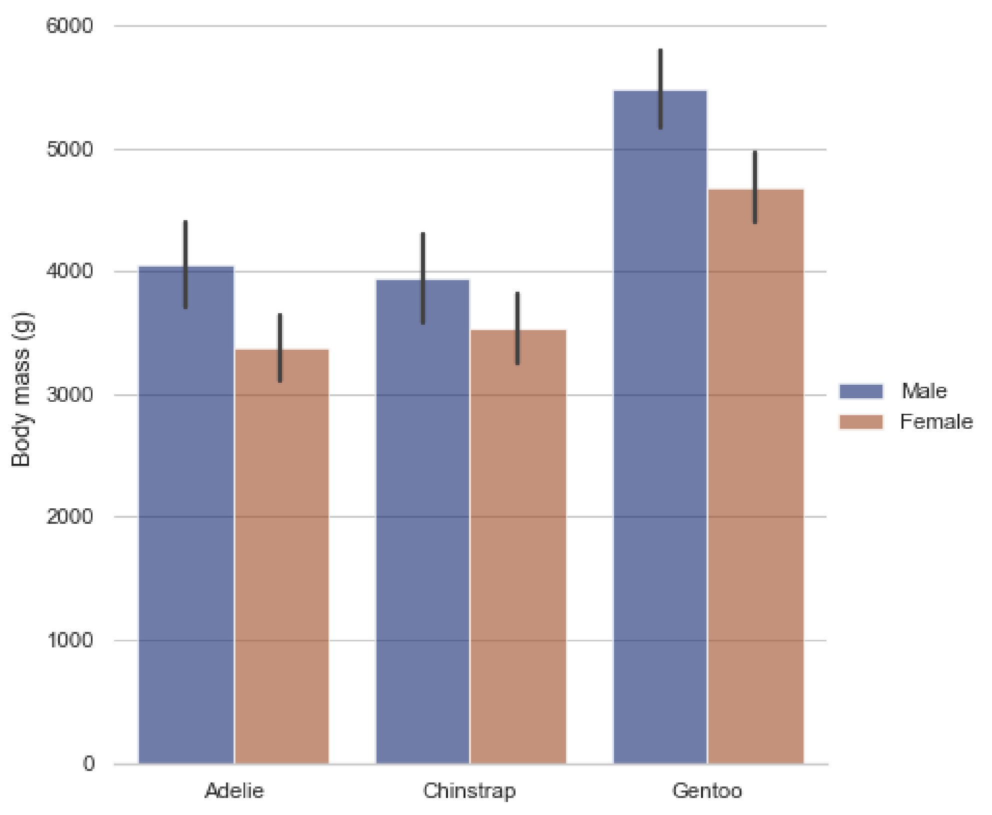
seaborn

tableau
from Salesforce

matplotlib

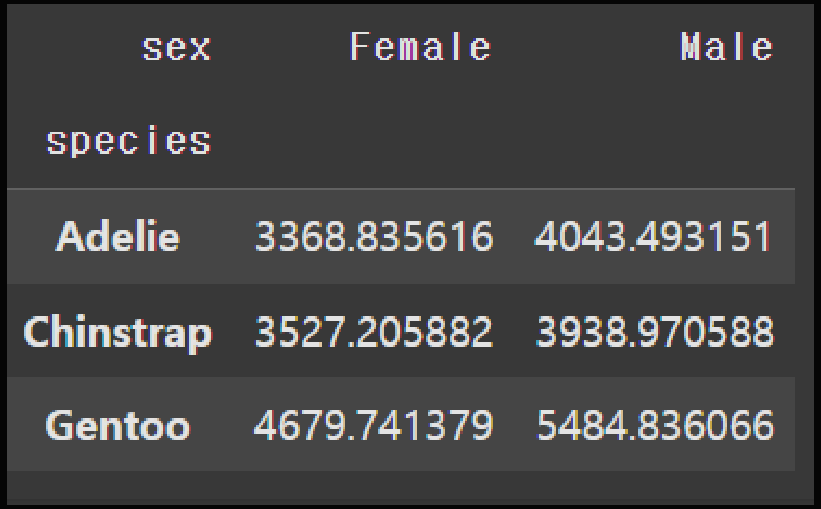
Power BI

Pandas groupby()



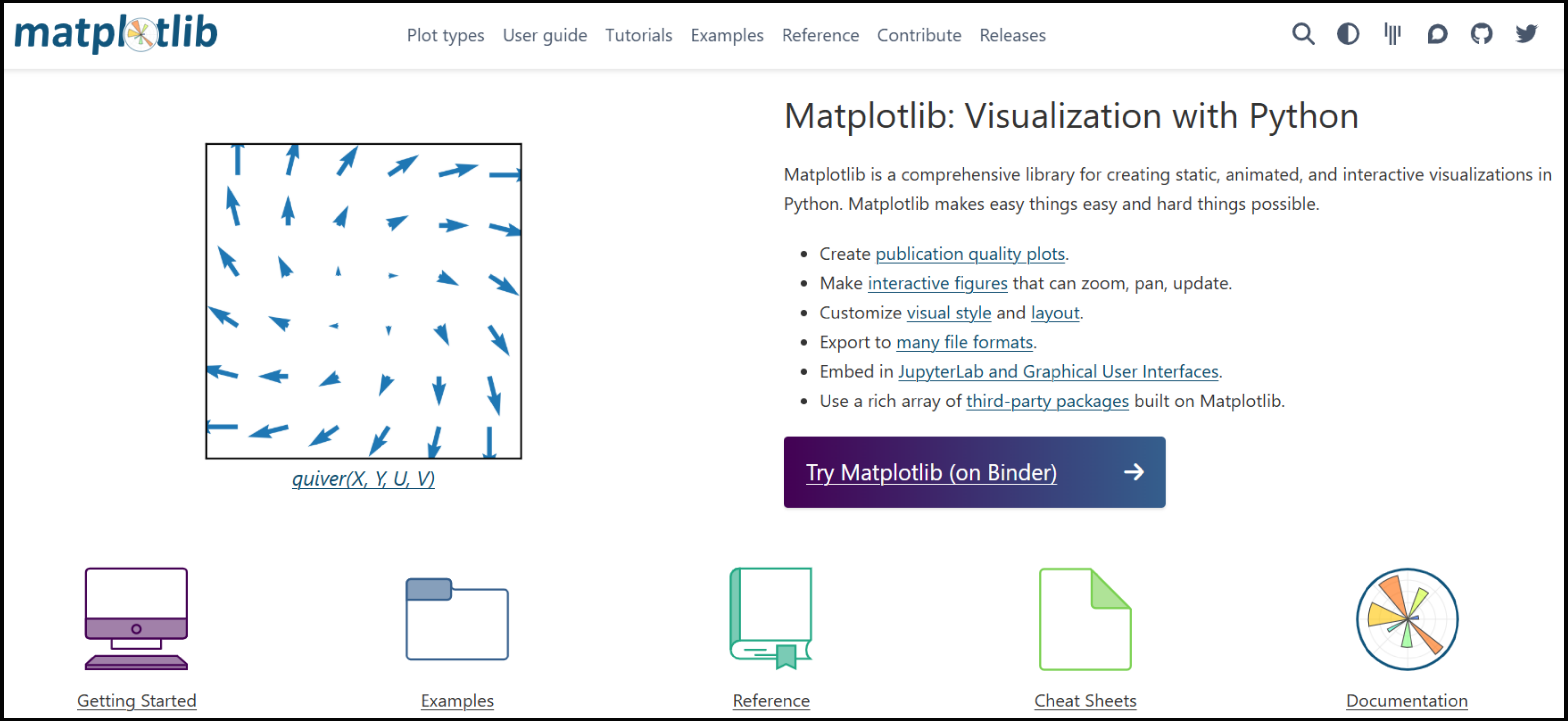
https://seaborn.pydata.org/examples/grouped_barplot.html

```
import seaborn as sns
penguins = sns.load_dataset("penguins")
penguins.groupby(["species", "sex"])["body_mass_g"].mean()
# ----------------------------------------------------------------#
import pandas as pd
pd.pivot_table(penguins,
            values='body_mass_g',
            index='species',
            columns='sex')
```



sex	Female	Male
species		
Adelie	3368.835616	4043.493151
Chinstrap	3527.205882	3938.970588
Gentoo	4679.741379	5484.836066

Matplotlib



Matplotlib: Visualization with Python

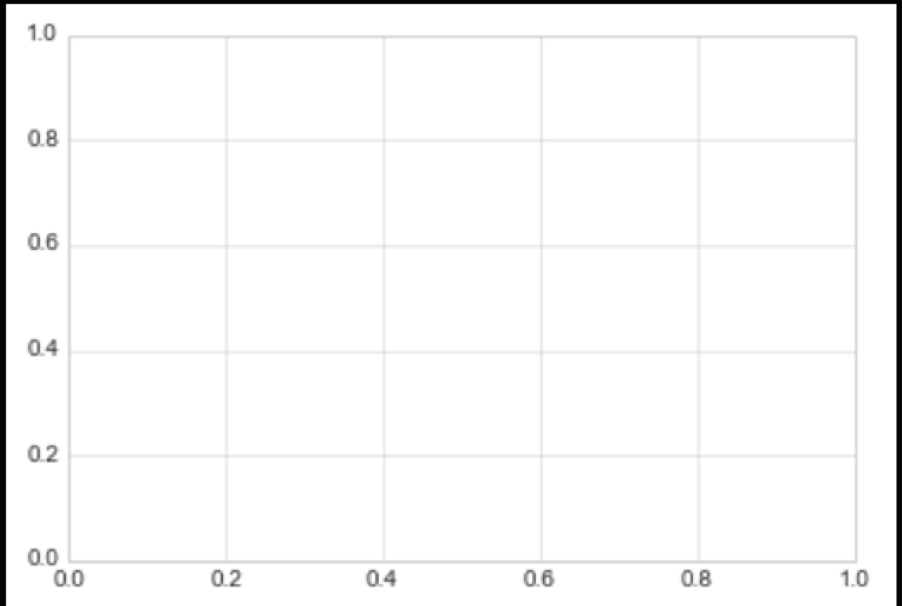
Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

quiver(X, Y, U, V)

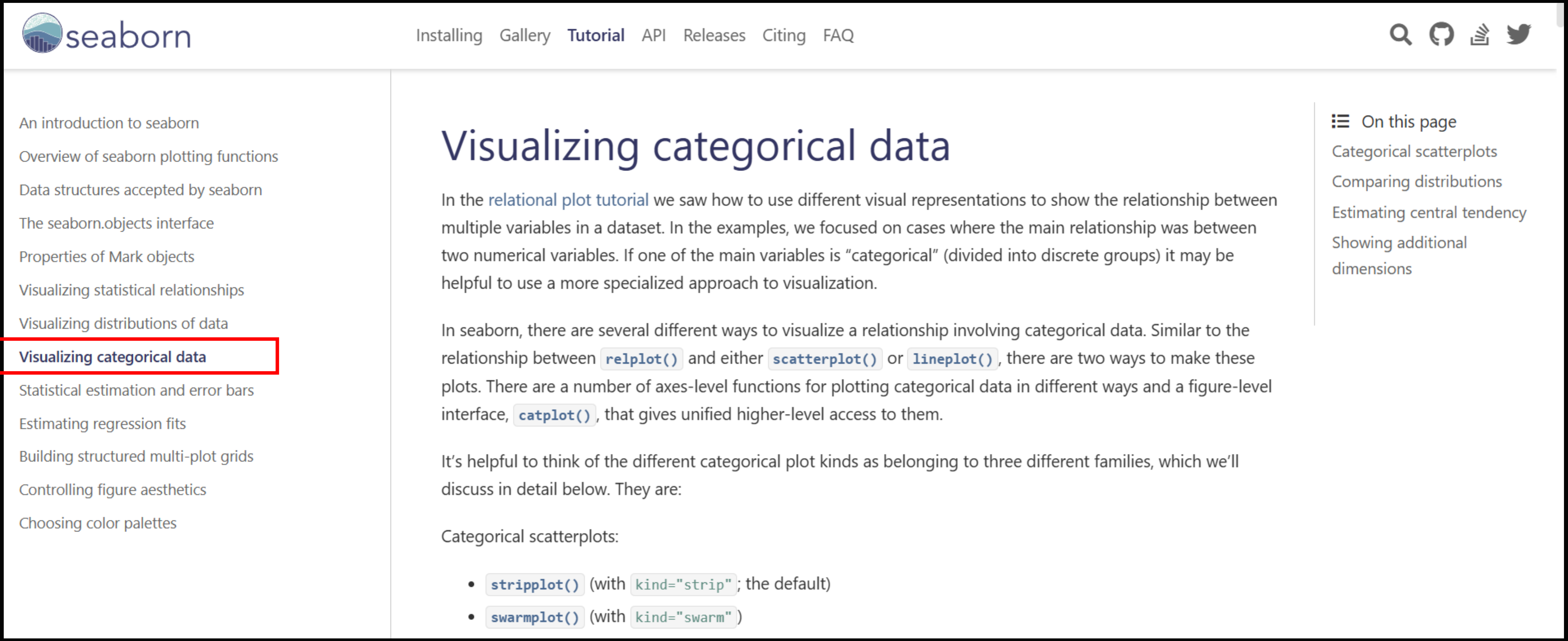
Matplotlib

```
import matplotlib.pyplot as plt
plt.style.use('seaborn-whitegrid')
fig = plt.figure()
ax = plt.axes()
```



https://jakevdp.github.io/PythonDataScienceHandbook/04.01-simple-line-plots.html

Seaborn Tutorial



An introduction to seaborn
Overview of seaborn plotting functions
Data structures accepted by seaborn
The seaborn.objects interface
Properties of Mark objects
Visualizing statistical relationships
Visualizing distributions of data
Visualizing categorical data
Statistical estimation and error bars
Estimating regression fits
Building structured multi-plot grids
Controlling figure aesthetics
Choosing color palettes

Visualizing categorical data

In the relational plot tutorial we saw how to use different visual representations to show the relationship between multiple variables in a dataset. In the examples, we focused on cases where the main relationship was between two numerical variables. If one of the main variables is "categorical" (divided into discrete groups) it may be helpful to use a more specialized approach to visualization.

In seaborn, there are several different ways to visualize a relationship involving categorical data. Similar to the relationship between `relplot()` and either `scatterplot()` or `lineplot()`, there are two ways to make these plots. There are a number of axes-level functions for plotting categorical data in different ways and a figure-level interface, `catplot()`, that gives unified higher-level access to them.

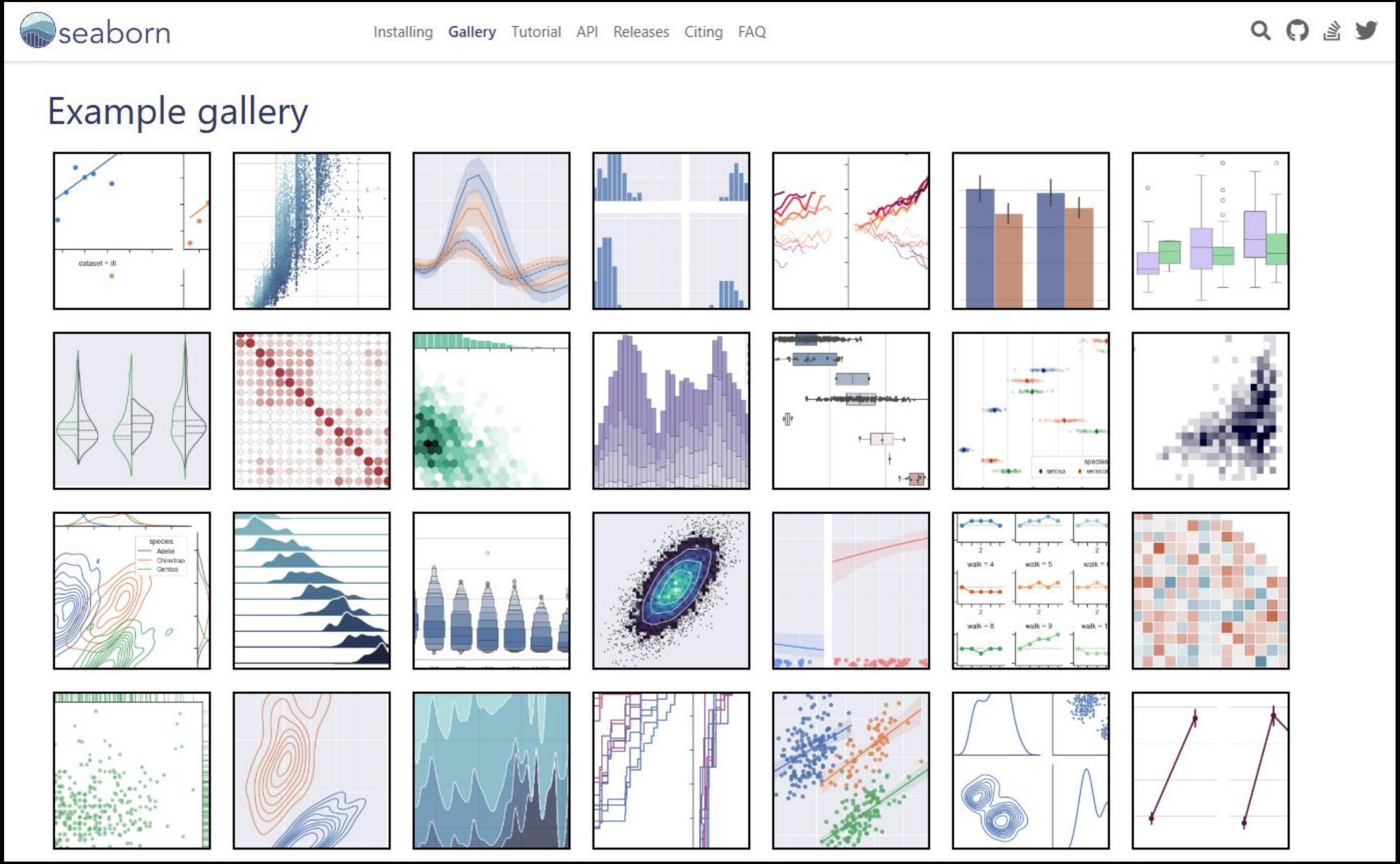
It's helpful to think of the different categorical plot kinds as belonging to three different families, which we'll discuss in detail below. They are:

Categorical scatterplots:

- `stripplot()` (with `kind="strip"`; the default)
- `swarmplot()` (with `kind="swarm"`)

On this page
Categorical scatterplots
Comparing distributions
Estimating central tendency
Showing additional dimensions

Practice_ to pandas



Example gallery