

Pythonデータ分析コトハジメ

PyLadies Caravan in Hokkaido-Muroran



■自己紹介■

まーや

@maaya8585



- PyLadies Caravan STAFF/PyLadies Tokyo Staff
- Azure 技術営業 / クラウド&アプリアーキテクト
- Disney好き
- ダイビング好き(ウミウシ)
- スノボ好き
- 分析よりデータ処理とwebバックエンドが好き
- 今日私のスマホは函館にいる。

※ 今日やること ※

■Pythonってどんな言語？

■jupyter notebookを使ってみよう

■データ分析のことはじめ

- +データの読み込み
- +データの形式を把握
- +要約統計量の算出
- +データの可視化

Pythonってどんな言語？



- コンピュータが処理を行う際には、コンピュータを動作させるための命令で構成される何らかのプログラムが動いている。このプログラムを記述するための言語がプログラミング言語である。
- 開発者はオランダ出身のグイド・ヴァン・ロッサム (Guido van Rossum) 氏。
 - ✓ Pythonという名前の由来は、イギリスのテレビ局BBCが製作したコメディ番組である『空飛ぶモンティ・パイソン』をグイド氏が好きで、この番組にちなんで付けられた。但し、ロゴはPython（蛇）が使われている。
 - ✓ 今はMicrosoftで働いてる。
- Pythonの開発や維持は、Pythonソフトウェア財団 (Python Software Foundation) が実施。



(出典 : [Guido's Personal Home Page](#))

Pythonってどんな言語？



➤ Pythonの特徴

- ✓ フリーのOSS（オープンソースソフトウェア）である
- ✓ Windows、Linux、MacOS等、様々な環境で動作するマルチプラットフォーム
- ✓ インターネット上に豊富なドキュメントが公開されている
- ✓ Webやデータ解析等、幅広い分野に適応可能な汎用言語
- ✓ プログラムの作成・実行・テストが容易なスクリプト言語
- ✓ 記述性と可読性が高く、またインタプリタ形式（対話的に1行ずつ実行）を採用
- ✓ ライブラリ（処理を簡単に行うための部品）が豊富
- ✓ そこまでプログラムの実行速度は速くない
- ✓ 実は習得は決して容易というわけではない



（出典：[Guido's Personal Home Page](#)）

Pythonってどんな言語？



- PythonはPSFによって日々バージョンアップをしている
 - ✓ Python 3.11.15 といったように、バージョン番号で管理されており、メジャーバージョン番号.マイナーバージョン番号.小さな変更やバグフィックスによって付けられる番号 という形式。
 - ✓ メジャーバージョンは1.0が1994年、2.0が2000年、3.0が2008年にリリース。
 - ✓ 言語として注目されたのは2系からであり、さらに3系では言語レベルで大きく変わった為、2系のプログラムは3系では、そのまま動作することはほぼない。
 - ✓ 3系でもマイナーバージョンが変わると、挙動が少し変わったり、ライブラリの整合性等で動作しない場合があるので、本格的な開発やWeb情報の参照等でも注意が必要

現在のpython最新バージョン
【 3.13.0 】

Pythonってどんな言語？



- Pythonが使われているサービス例
 - ✓ Amazon
 - ✓ Dropbox
 - ✓ Netflix
 - ✓ Instagram
 - ✓ YouTube
 - ✓ Spotify

Pythonの思想 : The Zen of Python

```
>>> import this
```

1. Beautiful is better than ugly.
2. Explicit is better than implicit.
3. Simple is better than complex.
4. Complex is better than complicated.
5. Flat is better than nested.
6. Sparse is better than dense.
7. Readability counts.
8. Special cases aren't special enough to break the rules.

醜いよりは美しい方がいい
暗黙の了解よりは明示した方がいい
複雑よりシンプルな方がいい
でも込み入るくらいなら複雑な方がいい
ネストは浅い方がいい
詰め込み過ぎよりはバラす方がいい
読みやすさの積み重ねは善
特殊だからってルールを破る理由にならない

Pythonの思想 : The Zen of Python

9. Although practicality beats purity.
10. Errors should never pass silently.
11. Unless explicitly silenced.
12. In the face of ambiguity, refuse the temptation to guess.
13. There should be one-- and preferably only one --obvious way to do it.
14. Although that way may not be obvious at first unless you're Dutch.

とはいえ現実には臨機応変にせざるを得ない
エラーは絶対に隠してはいけない

エラーが無視できる理由が明示されない限り
曖昧なものに出会ったらその意味を勝手に
推測しては行けない

何かいいやり方があるはずだ。誰が見ても
明らかな、たったひとつのやり方が。

そのやり方は一目見ただけではわかりにくい
かもしれない。オランダ人にだけわかりやす
いなんてこともあるかもしれない。

Pythonの思想 : The Zen of Python

- 15. Now is better than never.
- 16. Although never is often better than *right* now.
- 17. If the implementation is hard to explain, it's a bad idea.
- 18. If the implementation is easy to explain, it may be a good idea.
- 19. Namespaces are one honking great idea -- let's do more of those!

ずっとやらないより今やるべき
でも今「すぐ」じゃない方が良い時の方が
往々にしてある
コードの意味を説明できないのであれば、
それは悪い実装である
コードの意味を簡単に説明できるのであれば、
それは良い実装である
名前空間の概念は素晴らしいので、
積極的に使っていこう

環境を整えよう！

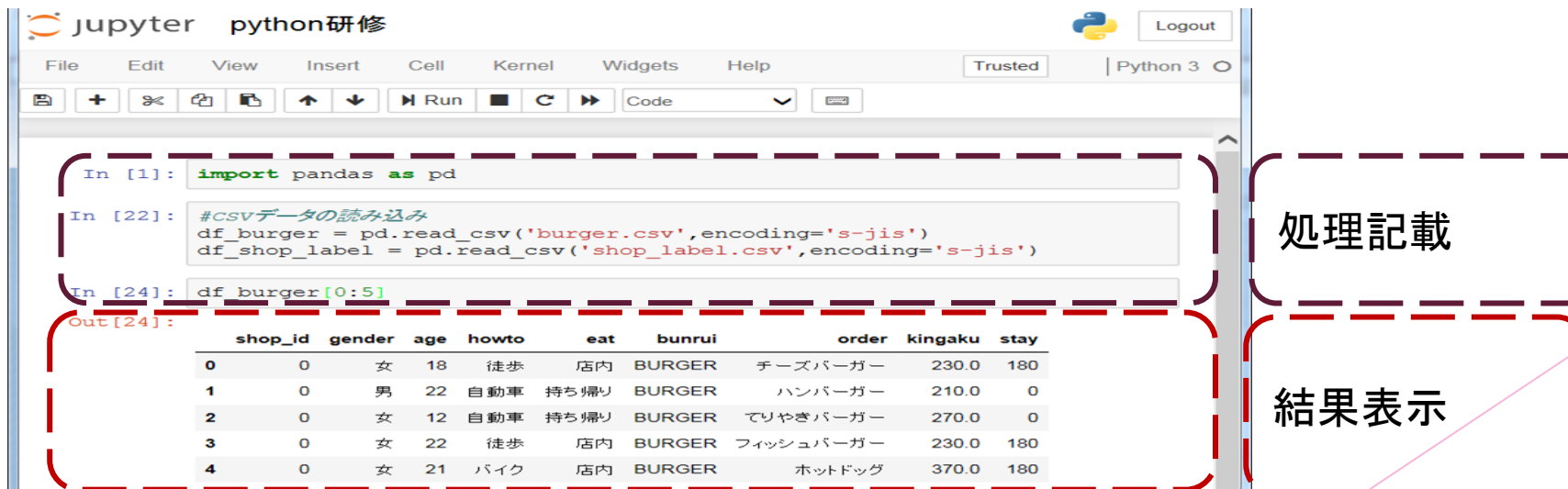
今回はJupyter Notebookを使用します

Jupyter Notebookとは、

ブラウザ形式のテキストエディタ。

ノートブックと呼ばれる形式でプログラムを作成でき、
実行結果を確認しながら作業を進めるためのツールです。

<実行画面>



The screenshot shows a Jupyter Notebook interface with the title 'python研修'. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running, and saving. The code area contains three input cells:

```
In [1]: import pandas as pd
```

```
In [22]: #CSVデータの読み込み
df_burger = pd.read_csv('burger.csv',encoding='s-jis')
df_shop_label = pd.read_csv('shop_label.csv',encoding='s-jis')
```

```
In [24]: df_burger[0:5]
```

The output area shows the result of the last execution, labeled 'Out[24]:', which is a DataFrame table with 10 columns and 5 rows of data. The table is enclosed in a red dashed box. To the right of the code area, there are two labels in dashed boxes: '処理記載' (Processing Record) and '結果表示' (Result Display).

	shop_id	gender	age	howto	eat	bunrui	order	kingaku	stay
0	0	女	18	徒歩	店内	BURGER	チーズバーガー	230.0	180
1	0	男	22	自動車	持ち帰り	BURGER	ハンバーガー	210.0	0
2	0	女	12	自動車	持ち帰り	BURGER	てりやきバーガー	270.0	0
3	0	女	22	徒歩	店内	BURGER	フィッシュバーガー	230.0	180
4	0	女	21	バイク	店内	BURGER	ホットドッグ	370.0	180

Pythonの稼働確認！

【Windows】

1. コマンドプロンプトを起動する。

見つからない場合は、スタートメニュー横の検索バーに cmd と入力すると候補としてコマンドプロンプトが表示されます



コマンドプロンプトやターミナルは、OS搭載のシステムツールで、コマンドと呼ばれる命令文を実行することができます

2. コマンドプロンプトに `python -V` または `python --version` と入力し、enterキーを押下。結果に「Python 3.xx.xx」と表示されることを確認する。（コマンドプロンプトを閉じる）

```
Microsoft Windows [Version 10.0.22621.1992]
(c) Microsoft Corporation. All rights reserved.

C:\Users\kanan>python -V
Python 3.11.5
```

【Mac】

1. ターミナルを起動する。



2. ターミナルに `python3 -V` または `python3 --version` と入力し、enterキーを押下。結果に「Python 3.xx.xx」と表示されることを確認する。（ターミナルを閉じる）

上手く結果が表示されない場合は、アンインストールして再度インストールを行ってください。

必要なPythonライブラリのインストール

- ▶ 以下の手順で、jupyter notebookをインストールしてください。
(要インターネット接続)

Jupyter notebookはコードの逐次実行が可能なブラウザベースのPython開発環境です。

- 1) コマンドプロンプト (Windows) または、ターミナル (Mac) を起動する。
- 2) 以下コマンドを入力し、実行する。

【Windows】

```
pip install jupyter
```

【Mac】

```
pip3 install jupyter
```

- 3) ほかに必要なライブラリも一緒にインストールしておく。

【Windows】

```
pip install pandas  
pip install matplotlib
```

【Mac】

```
pip3 install pandas  
pip3 install matplotlib
```

Jupyter Notebookの起動

▶ 以下の手順で、jupyter notebookを稼働確認をしてください。

1) コマンドプロンプト（Windows）または、ターミナル（Mac）を起動する。（前頁のインストールで起動済みの人はそのままでOK）

2) ダウンロードしたファイルを置いた任意のフォルダまで移動する。

【Windows・Mac共通】

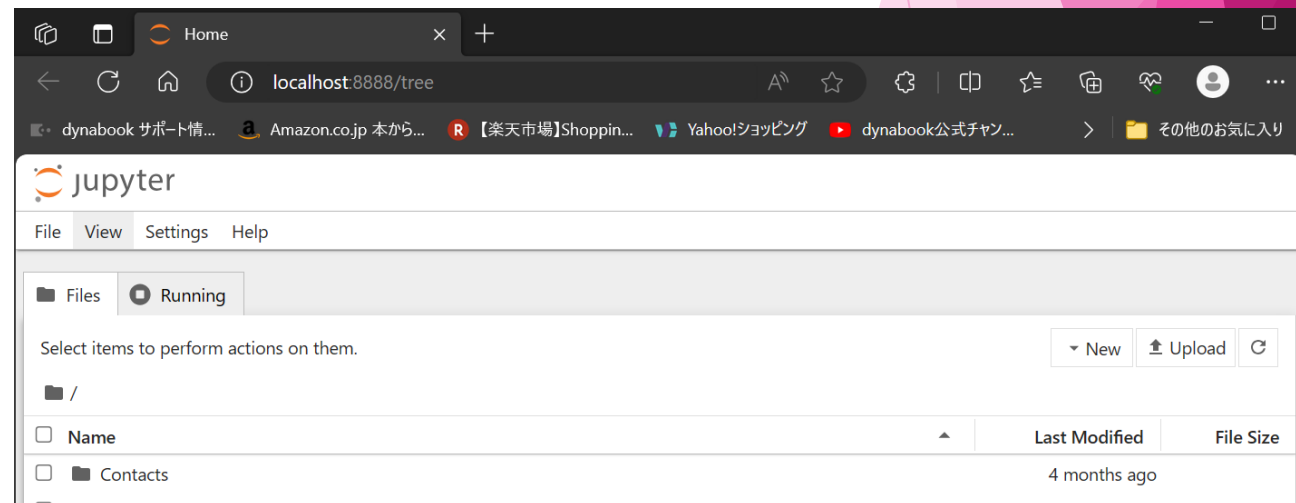
```
cd 任意のフォルダパス
```

3) 以下コマンドを入力し、実行する。

【Windows・Mac共通】

```
Jupyter notebook
```

4) ブラウザが起動し、Jupyter Notebookの画面が表示される



Jupyter Notebookを使って
Pythonデビューしてみよう。

Jupyter Notebookを使って
データ分析をやってみよう！

データ分析で大事なこと

データ分析っていうと、
多変量解析とか機械学習とかってすぐやりたくなる

でもとても大事なことはデータを理解すること。

そのデータがどんな姿をしているのかを

数値化と**グラフ化**で泥臭く地道に

解き明かす過程がデータ分析の大半を占めたりする

(参考) describe : 要約統計量

count	: 件数
mean	: 平均
std	: 標準偏差
min	: 最小値
25%	: 25%点 (第 1 四分位数)
50%	: 50%点 (第 2 四分位数、中央値)
75%	: 75%点 (第 3 四分位数)
max	: 最大値

(参考) 平均と中央値 (50%点)

- **平均**は少数の外れ値 (異常値) に大きな影響を受ける統計量！！
- 平均は全体の中心を表す統計量としてよく利用されるけど、外れ値を含んだデータでは、外れ値に影響を受けやすいので注意が必要。

ユーザー	スマホゲーム月課金額
A	200円
B	50円
C	300円
D	250円
E	600円
F	5,000円

平均 : 1,067円
中央値 : 275円

(参考) パーセンタイル値

パーセンタイル値とは、データを昇順に並べた時の位置を表します。

※百分位で位置を表す場合にパーセンタイル値となる。

データを昇順に並べ等分した時の位置を分位数 (quantile) という。

よく使われるのは4等分する四分位数 (quartile) である。

【四分位数】

25%値 : 全データの25%が入る値 第1四分位点 (Q1)

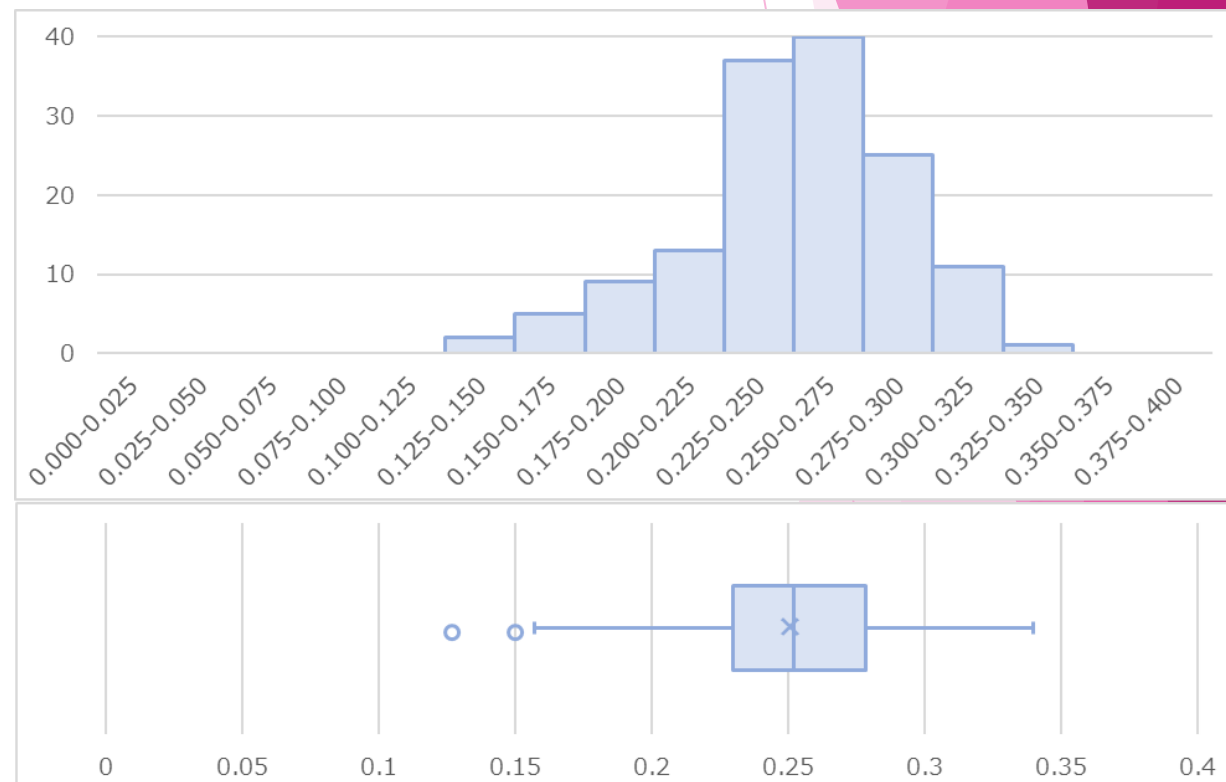
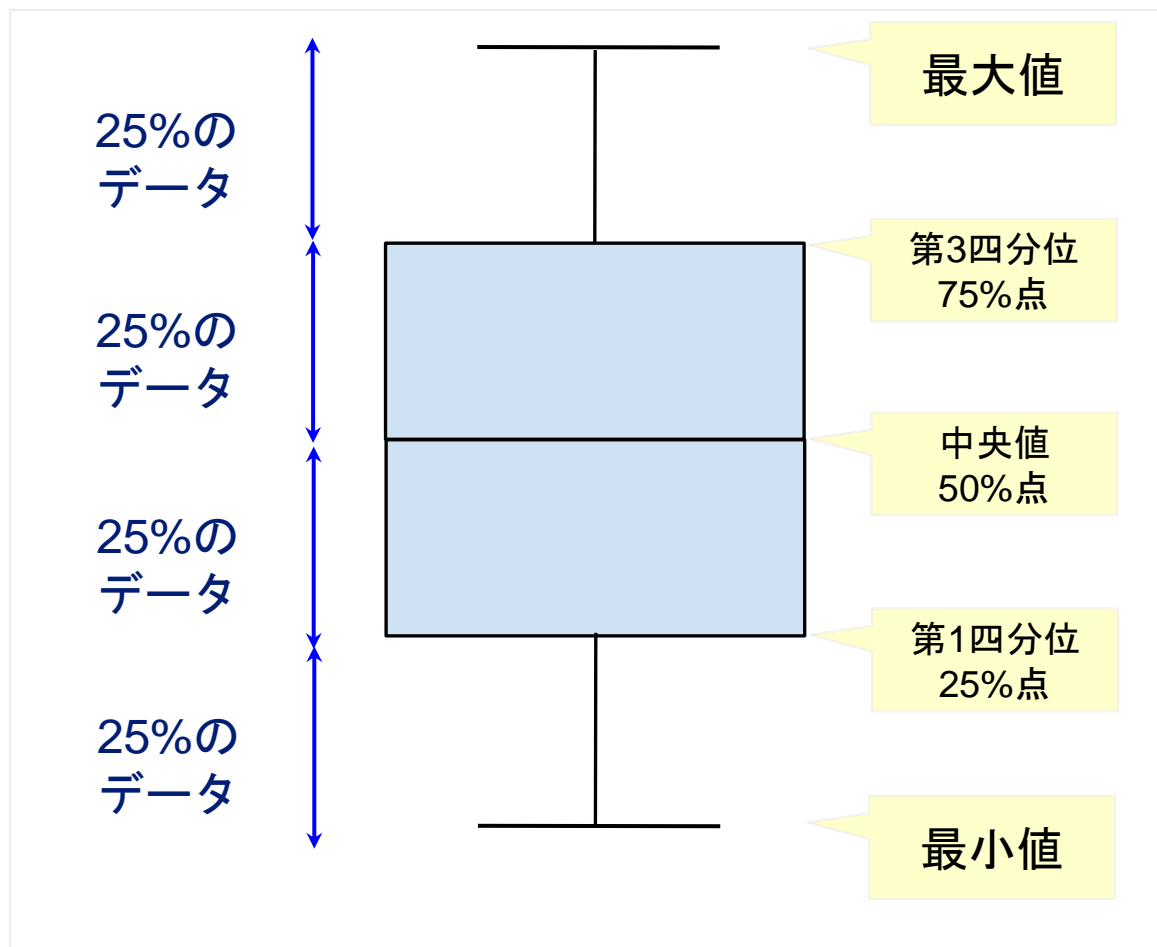
50%値 : 全データの50%が入る値 第2四分位点 (中央値) (Q2)

75%値 : 全データの75%が入る値 第3四分位点 (Q3)

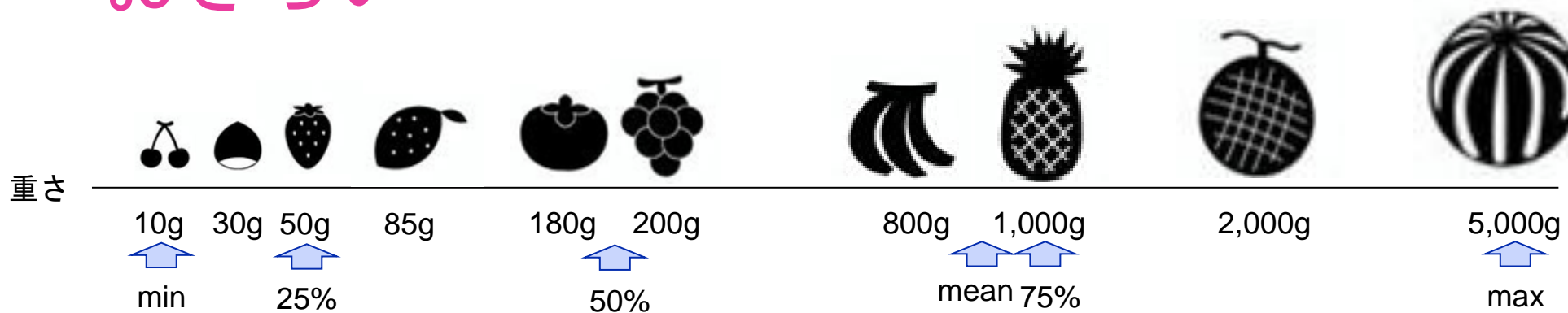


中央値 : 190.0g / 平均値 : 935.5g

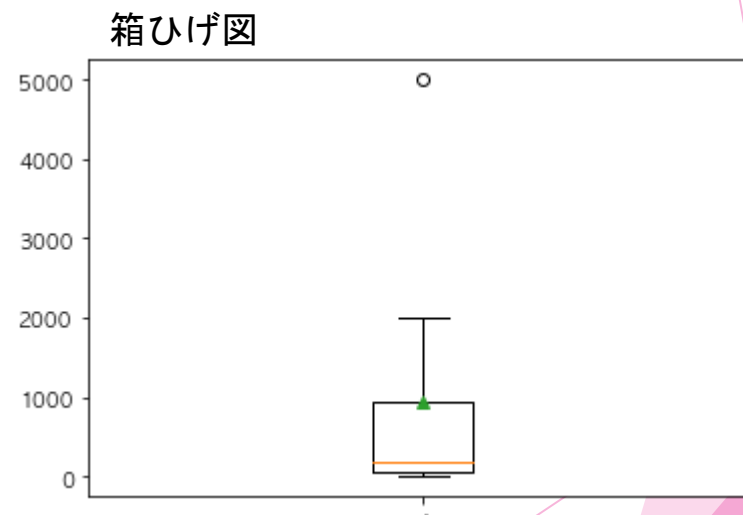
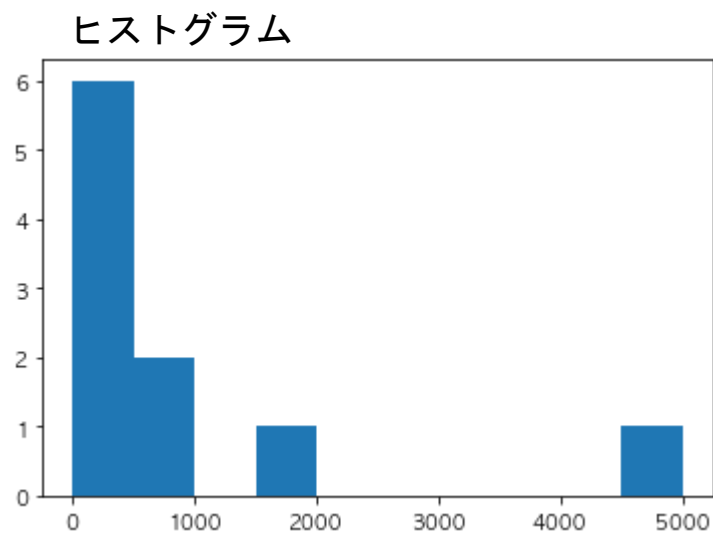
(参考) 箱ひげ図



おさらい



代表値	
count	10
mean	935.5
std	1,482.2
min	10
25%	50
50%	190
75%	1,000
max	5,000



本日の“LET'S TRY”

sample_data.csv

項目名	数値/カテゴリ	内容
No	数値	通し番号
地方	カテゴリ	8地方名（漢字）
chihou	カテゴリ	8地方名（アルファベット）
都道府県	カテゴリ	都道府県（漢字）
area_km2	数値	面積（単位：km ² ）
population_k	数値	人口（単位：千人）
female_k	数値	女性人口（単位：千人）
ramen_shop	数値	ラーメン店舗数（単位：軒）
sake_l	数値	総アルコール消費量（単位：ℓ）
sake_l_person	数値	20歳以上1人あたりアルコール消費量（単位：ℓ）
mcdnald_shop	数値	マクドナルド店舗数（単位：軒）
yakitori_shop	数値	焼き鳥屋店舗数（単位：軒）
name_sato	数値	苗字が「佐藤」さんの人数（単位：人）
name_kato	数値	苗字が「加藤」さんの人数（単位：人）
online_game	数値	オンラインゲーム利用率
source_ml	数値	2人以上の世帯の年間ソース消費量（単位：ml）
moyashi_g	数値	2人以上の世帯の年間もやし消費量（単位：g）
mikan_g	数値	2人以上の世帯の年間みかん消費量（単位：g）
orange_g	数値	2人以上の世帯の年間オレンジ消費量（単位：g）

challenge

問題 1 :

name_satoは、佐藤さんの人数です。その都道府県にどのくらい佐藤さんの割合が多いのかを確認してみよう。

※注意：都道府県の人口（population_k）は千人単位です。

問題 2 :

佐藤さん比率が多ければ多いほど、ラーメン店の数は多くなるのか検証してみよう！

問題 3 :

自分の出身（もしくは縁のある）都道府県について、データから得られる知見を探してみましょう。

はじめようPython Life★