データサイエンス基礎演習Ⅱ　（2023.04.16 版）

目的　・プログラミング言語「Python」を用い，実際の測定機器で測定されたデータにおいて，統計的な代表値（平均値や中央値等）を求めることができる。
・プログラミング言語「Python」を用い，ヒストグラムや散布図等によりデータの可視化することができる。
・プログラミング言語「Python」を用い，2 つの異なる種類のデータの間の関係性を示す指標の一つである相関係数を求めることができる。

使用機器：ノートパソコン（インターネットに接続済み），測定データ

注意事項：手引書の用語等，不明なところは予め書籍等を活用して調べておくこと。
前半にて設問を解き，後半で考察の内容を考え，レポートをまとめてください。
レポートは Google 社のクラスルームから電子ファイルにてご提出をお願いします。
なお，実行結果はジュピターノートブック形式のファイルにてご提出ください。

実験日のスケジュール：
（1）途中であっても午前で作業を打ち切り，午後から報告書作成へ進むこと。
（2）報告書を授業終了時間前までに完成させ，班員同士の評価（ピアレビュー）を行うこと。
（3）授業終了時に，班員全員がレポートを提出できるように協力してください。

実験概要

本テーマはプログラミング言語「Python」を用いて各種データ処理を行います。これを通し，データサイエンスを行う上で必要となる知識や技術を身につけていきます。今回は旭川高専に既設の施設にて収集された実際のデータを扱います。昨年度と違い，実際のデータには機器のトラブル等が原因による外れ値や測定漏れが含まれています。それらを踏まえ，より実際に近い状況で演習を行って行きます。

「Python」について

分からないことは参考文献やホームページ等を活用して，各自で積極的に調べてください。また，調べた結果をレポートにまとめておいてください。調べ方やまとめ方を身につけることも自分のための勉強になると思います。

なお，本テーマの Python のプログラムは全て Google 社「Colaboratory」にて動作を確認しています。

参考文献[1]のためのサポートページ（Python でデータ処理を進める上でのチュートリアル）
https://github.com/ghmagazine/python_stat_sample
ジュピターノートブック形式のファイルを開いて，下記について確認してください。

1. ライブラリ「pandas」についてのチュートリアル
データ処理の際に便利なものを集めたライブラリ
https://github.com/ghmagazine/python_stat_sample/blob/master/tutorial/pandas.ipynb
（1） ライブラリの読み込みの方法
（2） 「データフレーム」という形式で csv ファイルを読み込む方法
（3）「データフレーム」という形式の扱い方

2. ライブラリ「matplotlib」について
グラフを描画する等，データ可視化を扱う際に便利なものを集めたライブラリ

3. ライブラリ「numpy」について
行列（配列）の演算を扱う際に便利なものを集めたライブラリ
https://github.com/ghmagazine/python_stat_sample/blob/master/tutorial/numpy.ipynb
（1） データ型の一つであるリストを用いた処理方法

4. 簡単なチュートリアル
https://github.com/ghmagazine/python_stat_sample/blob/master/tutorial/python.ipynb
Python については知らないことが多いと思いますので，サンプルプログラム等を参考にして
必要なところは，真似してみてください。

準備手順

(1)Google クラスルームにあるデータファイルをダウンロードしてください。
(2)各自のアカウントにて Google 社「Colab」のページを開いてください。
(3)その「Colab」のページへデータファイルをアップロードしてください。
(4)演習を進めるために必要なファイル（ライブラリ）を読み込むために下記のプログラムを実行する。
なお，左側の数字は行番号，「＃」以降はコメントですので，入力する必要はありません。

```
1 # 以下のライブラリを使うので、あらかじめ読み込んでおいてください
2 import numpy as np
3 import numpy.random as random
4 import scipy as sp
5 import pandas as pd
6 from pandas import Series, DataFrame
7
8 # 可視化ライブラリ
9 import matplotlib.pyplot as plt
10 import matplotlib as mpl
11 import seaborn as sns
12 %matplotlib inline
13
14 # 小数第3位まで表示
15 %precision 3
```

資料を参考にすると下記のプログラムにてデータファイルを読み込むことができる。

```
1 df = pd.read_csv('/content/house_data_utf8.csv',index_col='時刻')
2 # dfの最初の5行を表示
3 df.head()
```

演習 1「統計的な代表値な計算」

データファイルには 1 週間分の多くの種類の測定データがあります。例えば，次のプログラムにて測定開始から 20 分間の気温の統計的な代表値を計算できます。

```
1 kion0 = np.array(df['気温（℃）'])[:20]
2 kion0
3 avg = np.mean(kion0)
4 md = np.median(kion0)
5
6
7 print("平均",avg)
8 print("中央値", md)
9
```

```
平均 19.874999999999996
中央値 19.9
```

設問 1

(1)「気温」以外についても，測定開始から 20 分間の統計的な代表値を計算してください。
(相対湿度，絶対湿度，CO2 濃度，風向くらいまで計算してください。)
(2)「気温」について，測定開始から 120 分間の統計的な代表値を計算してください。

もし余裕があれば，「気温」以外のデータについても，測定開始から 120 分間の統計的な代表値を計算してください。

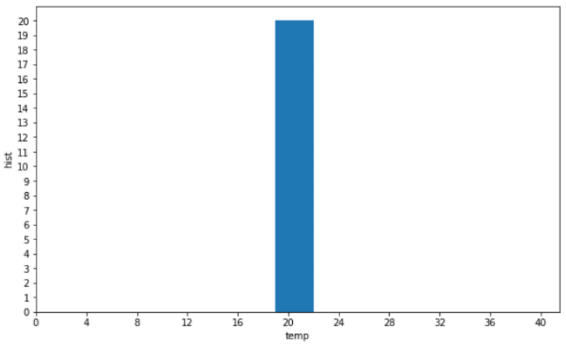
演習2「データの可視化」

例えば，下記にて測定開始から 20 分間の気温のヒストグラムを描画することができます。

```
kion1 = np.array(df['気温 (℃) '])[:20]

# キャンバスを作る
# figsizeで横・縦の大きさを指定
fig = plt.figure(figsize=(10, 6))
# キャンバス上にグラフを描画するための領域を作る
# 引数は領域を1×1個作り、1つめの領域に描画することを意味する
ax = fig.add_subplot(111)

# 階級数を10にしてヒストグラムを描画
freq, _, _ = ax.hist(kion1, bins=10, range=(10, 40))
# X軸にラベルをつける
ax.set_xlabel('temp')
# Y軸にラベルをつける
ax.set_ylabel('hist')
# X軸に0, 10, 20, ..., 100の目盛りをふる
ax.set_xticks(np.linspace(0, 40, 10+1))
# Y軸に0, 1, 2, ...の目盛りをふる
ax.set_yticks(np.arange(0, freq.max()+1))
# グラフの表示
plt.show()
```



設問 2

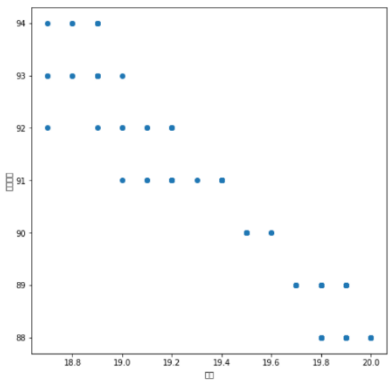
(1)「気温」以外のデータについても，測定開始から 20 分間のヒストグラムを描画してください。
(相対湿度，絶対湿度，CO2 濃度，風向くらいまで描画してください。)
(2)「気温」のデータについて，測定開始から 120 分間のヒストグラムを描画してください。

もし余裕があれば，「気温」以外のデータについても，測定開始から 120 分間のヒストグラムを描画してください。

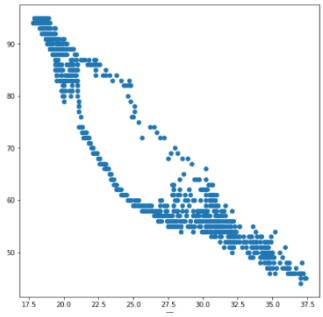
演習 3「データの可視化　2」

例えば，下記にて測定開始から 120 分間の気温(X 軸)と相対湿度(Y 軸)の散布図を描画することができます。

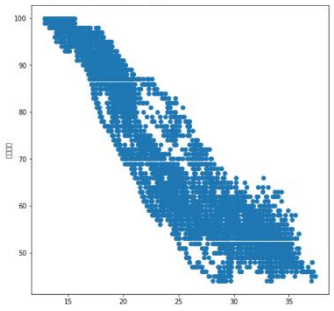
```
kion1 = np.array(df['気温 (℃) '])[:120]
shitsudo1 =np.array(df['相対湿度 (%) '])[:120]
fig = plt.figure(figsize=(8, 8))
ax = fig.add_subplot(111)
# 散布図
ax.scatter(kion1, shitsudo1)
ax.set_xlabel('気温')
ax.set_ylabel('相対湿度')
```



ちなみに 1 日分および 1 週間分の気温（X 軸）と相対湿度（Y 軸）の散布図は下図のようになります。



図：1 日分の散布図



図：1 週間分の散布図

設問 3　「気温」と「相対湿度」以外の組み合わせのデータについても，測定開始から 120 分間の散布図を描画してください。
（絶対湿度，CO2 濃度，風向くらいまで描画してください。）

演習 4「相関係数の計算」

例えば，下記にて測定開始から 20 分間の気温と相対湿度の相関係数を計算することができます。

```
1 np.cov(kion0, shitsudo0, ddof=0)[0, 1] / (np.std(kion0) * np.std(shitsudo0))
2
```

```
-0.363
```

設問 4　「気温」と「相対湿度」以外の組み合わせのデータについても，測定開始から 20 分間の相関係数を計算してください。
（絶対湿度，CO2 濃度，風向くらいまで描画してください。）

演習 5「データのクリーニング」

例えば，下記にて測定開始から 120 分間の気温と相対湿度の相関係数を計算しようとしても計算ができません。（「NaN」は「Not a Number」の略で非数と訳されます。）

```
1 kion1 = np.array(df['気温（℃）'])[:120]
2 shitsudo1 =np.array(df['相対湿度（%）'])[:120]
3 np.cov(kion1, shitsudo1, ddof=0)[0, 1] /(np.std(kion1) * np.std(shitsudo1))
```

```
nan
```

これは，下記にある赤枠や青枠のようにデータが欠損している箇所（装置の誤動作等）があるからです。

	A	B	C	D	E	F	G	H	I	J	K	
1	時刻	気温（℃）	相対湿度（%）	絶対湿度（g/m3	露点温度（℃）	飽差（g/m3）	CO2濃度（ppm）	日射量（kW/m2	体積含水率（%）	土中温度（℃）	電気伝導度：EC	重量
2	2021/8/29 0:00	19.9	88	15.2	17.9	2.1	145	0	53.7	34.1	0.8	
3	2021/8/29 0:01	19.9	88	15.2	17.9	2	139	0	53.7	34.2	0.8	
4	2021/8/29 0:02	19.9	88	15.2	17.9	2	137	0	53.7	34.2	0.8	
5	2021/8/29 0:03	19.9	88	15.2	17.9	2	149	0	53.6	34.1	0.8	
6	2021/8/29 0:04	19.9	88	15.2	17.8	2	142	0	53.6	34	0.8	
7	2021/8/29 0:05	19.9	89	15.3	18	1.9	149	0	53.6	33.9	0.8	
8	2021/8/29 0:06	19.9	89	15.3	18	1.9	146	0	53.6	33.8	0.8	
9	2021/8/29 0:07	19.9	89	15.2	17.9	2	144	0	53.6	33.9	0.8	
10	2021/8/29 0:08	19.9	88	15.2	17.9	2	157	0	53.6	33.6	0.8	
11	2021/8/29 0:09	19.9	89	15.3	18	1.9	148	0	53.6	33.6	0.8	
12	2021/8/29 0:10	19.9	89	15.3	18	1.9	156	0	53.6	33.6	0.8	
13	2021/8/29 0:11	19.9	89	15.3	17.9	1.9	141	0	53.6	33.6	0.8	
14	2021/8/29 0:12	19.9	89	15.3	18	1.9	155	0	53.6	33.9	0.8	
15	2021/8/29 0:13	19.9	88	15.2	17.9	2	151	0	53.6	33.9	0.8	
16	2021/8/29 0:14	19.9	88	15.2	17.9	2	145	0	53.6	33.9	0.8	
17	2021/8/29 0:15	19.9	88	15.2	17.9	2	142	0	53.6	33.8	0.8	
18	2021/8/29 0:16	19.9	89	15.2	17.9	2	154	0	53.6	33.8	0.8	
19	2021/8/29 0:17	19.8	89	15.2	17.9	1.9	149	0	53.6	33.8	0.8	
20	2021/8/29 0:18	19.7	89	15.1	17.8	1.9	145	0	53.6	33.8	0.8	
21	2021/8/29 0:19	19.7	89	15.3	17.9	1.8	144	0	53.6	33.7	0.8	
22	2021/8/29 0:20	19.7	89	15.2	17.9	1.9	140	0	53.6	33.7	0.8	
23	2021/8/29 0:21	19.8	89	15.2	17.9	2	150	0	53.6	33.7	0.8	
24	2021/8/29 0:22	19.9	89	15.2	17.9	2	148	0	53.5	33.7	0.8	
25	2021/8/29 0:23	19.9	89	15.2	18	2	133	0	53.5	33.7	0.8	
26	2021/8/29 0:24	19.9	88	15.2	17.9	2.1	142	0	53.5	33.8	0.8	
27	2021/8/29 0:25	20	88	15.2	17.9	2.1	154	0	53.5	33.8	0.8	
28	2021/8/29 0:26	20	88	15.2	17.9	2.1	153	0	53.5	33.8	0.8	
29	2021/8/29 0:27	20	88	15.2	17.9	2.1	150	0	53.5	33.8	0.8	
30	2021/8/29 0:28	20	88	15.2	17.9	2.1	150	0	53.5	33.8	0.8	
31	2021/8/29 0:29	20	88	15.2	17.9	2.1	155	0	53.5	33.7	0.8	
32	2021/8/29 0:30	20	88	15.2	17.9	2.1	149	0	53.5	33.7	0.8	
33	2021/8/29 0:31	19.9	88	15.2	17.9	2.1	152	0	53.5	33.6	0.8	
34	2021/8/29 0:32	19.9	88	15.2	17.9	2	145	0	53.5	33.6	0.8	
35	2021/8/29 0:33	19.8	89	15.2	17.9	1.9	146	0	53.5	33.6	0.8	
36	2021/8/29 0:34	19.9	89	15.4	18.1	1.8	143	0	53.5	33.6	0.8	
37	2021/8/29 0:35	19.8	89	15.2	17.9	2	151	0	53.5	33.6	0.8	
38	2021/8/29 0:36	19.9	88	15.2	17.9	2	147	0	53.5	33.6		
39	2021/8/29 0:37		88	15.2	17.9			0		33.6		
40	2021/8/29 0:38			15.1		2	139	0		33.6		
41	2021/8/29 0:39											
42	2021/8/29 0:40											
43	2021/8/29 0:41		88		17.8	2	143		53.4	33.6	0.8	
44	2021/8/29 0:42	19.8	88	15.1	17.8	2	149	0	53.4	33.6	0.8	
45	2021/8/29 0:43	19.8	88	15.1	17.8	2	140	0	53.4		0.8	
46	2021/8/29 0:44	19.8	88	15.2	17.9	1.9	140	0	53.4	33.5	0.8	

（データクリーニング作業）

欠損値を取り除くため，表計算ソフト（マイクロソフト社 EXCEL を利用する場合はマルチメディア室の PC を利用）を利用して，欠損している箇所の行を削除してください。

データクリーニングの後に下記の設問 5 を実行してください。ただし，ファイルを保存する際，文字コードは「UTF-8」としてください。

設問 5

（1）「気温」と「相対湿度」の組み合わせのデータについても，測定開始から 120 分間の相関係数を計算してください。

（2）「気温」と「相対湿度」以外の組み合わせのデータについても，測定開始から 120 分間の相関係数を計算してください。

（もし余力があれば，）

（3）「気温」と「相対湿度」の組み合わせのデータについても，測定開始から 1 日分の相関係数を計算してください。

（4）「気温」と「相対湿度」以外の組み合わせのデータについても，測定開始から 1 日分の相関係数を計算してください。

ちなみに，手元の計算では 1 日分の 「気温」と「相対湿度」の組み合わせの相関係数は「-0.951」くらいかと思います。

さらに，余裕があれば，上記の相関係数について，1 週間分のものを計算してみてください。

考察

a.「相関係数」の意味を調べ，まとめなさい。

b. 設問 3 と設問 5 を鑑みて，分かることをまとめなさい。

c. 今回はデータクリーニング（データクレンジング）として，欠損値を目視で発見し，その行を削除しました。この欠損値の処理以外に，データクリーニングとして予め処理しておくものについて調べ，まとめなさい。

ピアレビュー（他者の評価）について：

（1）別紙のチェックシートに従って，自分のレポートを他の班員 1 名以上，できれば全員に評価してもらうこと。
（2）もちろん，自分だけレビューをしてもらうのではなく，他の班員も評価すること。

【参考文献】

[1]「Python で理解する統計解析の基礎」 谷合 廣紀 著， 技術評論社
[2]「東京大学のデータサイエンティスト育成講座」 塚本邦尊ら 著，マイナビ出版社
[3]演習用データ 旭川高専 中村先生が作成・収集されたデータを利用させて頂いています。