

「Python プログラミング基礎その2」

九州大学 大学院システム情報科学研究院
情報知能工学部門
データサイエンス実践特別講座
備瀬竜馬, Diego Thomas, 正井克俊

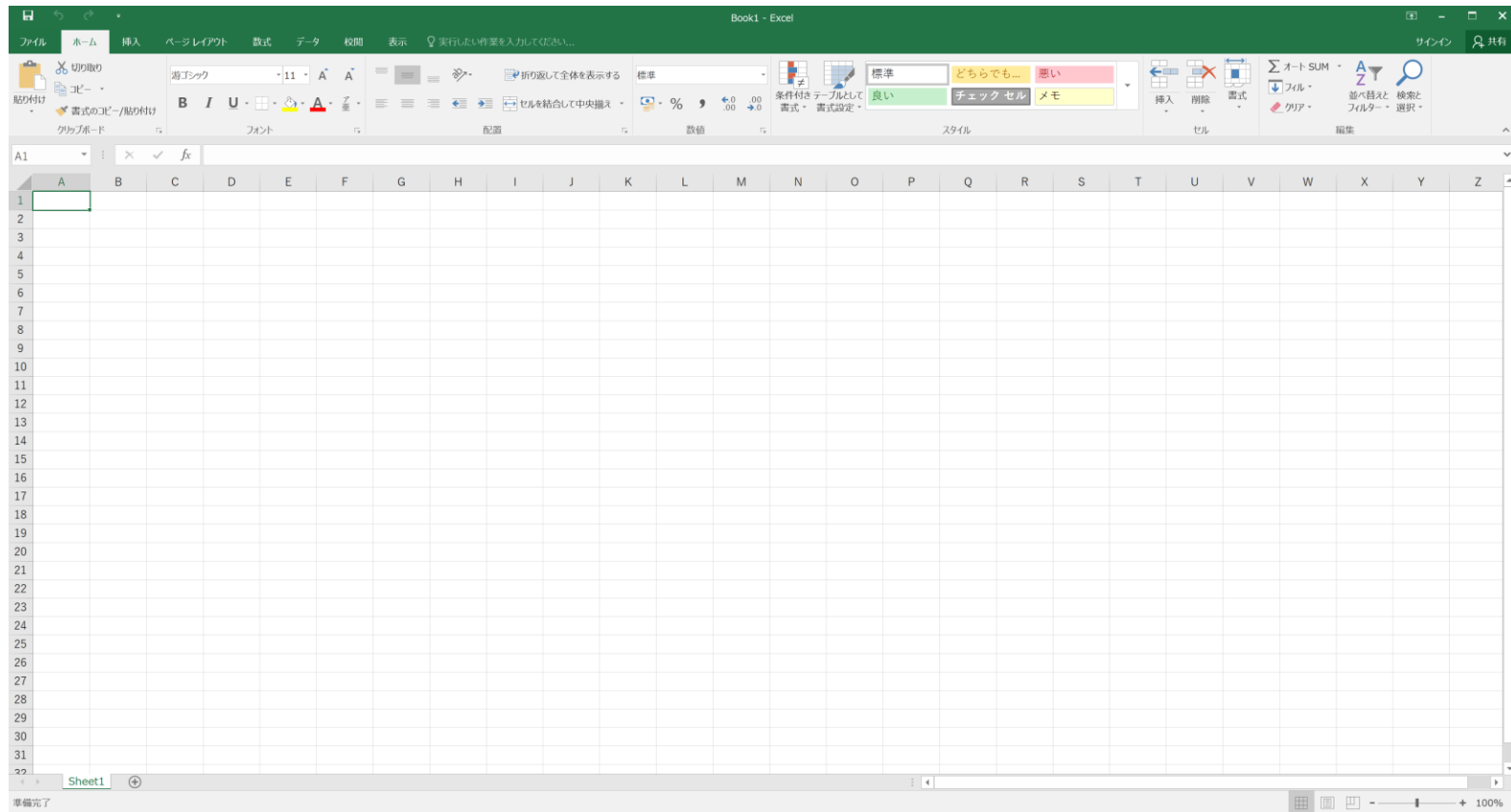
復習

CSV file format

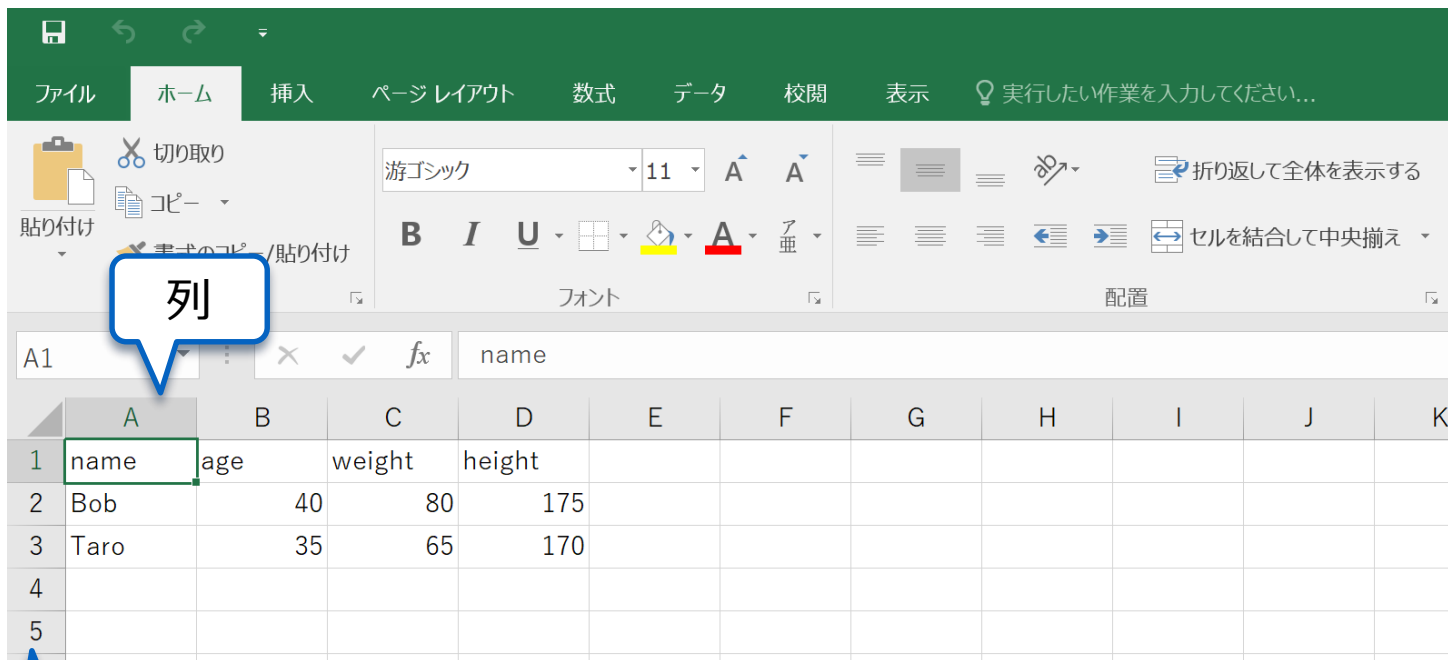
- CSV = comma separated values
- File format specifications:
 - 項目ごとにコンマ “,” で区切られたデータ
 - 1行が一まとめのデータで、改行で行を識別
- Example:
name, age, weight, height
Bob, 40, 80, 175
Taro, 35, 65, 170
▪ ▪ ▪

Excel で csv ファイル

● Excel



Excel で csv ファイル



The screenshot shows the Microsoft Excel interface with a CSV file open. The ribbon is set to 'ホーム' (Home). The worksheet contains the following data:

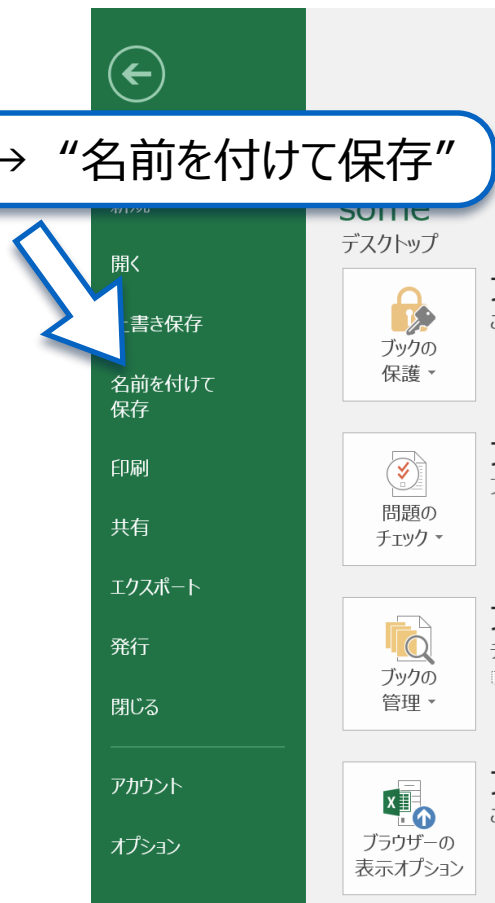
	A	B	C	D	E	F	G	H	I	J	K
1	name	age	weight	height							
2	Bob	40	80	175							
3	Taro	35	65	170							
4											
5											

行

Excel で csv ファイル

● CSV fileとして保存

ファイル → “名前を付けて保存”



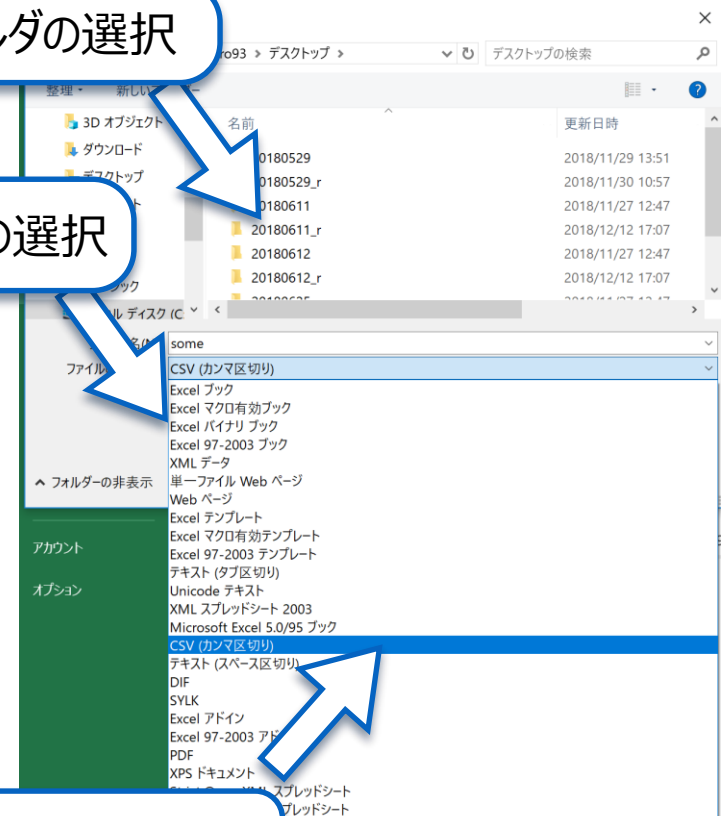
フォルダの選択

ファイルフォーマットの選択



csv (カンマ区切り)

適当に名前をつけて終了



ExcelでCSVファイルを作ってみよう！

- Excelを開いて、次の内容を自分で書いてみよう！
- Bob, Taroに加えて、1行分、適当に自分で、データを書いて、CSVファイルとしてsaveしてみよう！

name, age, weight, height

Bob, 40, 80, 175

Taro, 35, 65, 170

(ここに好きなデータを記載)

Fileのアップロード

introduction.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール

+ コード + テキスト

> [12] 1
2
3
4
4
2
3
4

▶ # 外部データの読み込み

```
# pandas
import pandas as pd
data = pd.read_csv(data_dir+"some.csv")
# 好きな行を見る
print(data.iloc[0])
print(data.iloc[0:2])
# 好きな列を見る
print(data.iloc[:,0])
print(data.iloc[:,0:3])

print(data.columns)

# 項目を指定して見る
print(data.name)
print(data.height)

# 項目を指定して好きな行を見る
print(data.iloc[0].weight)
print(data.weight[1])

for i in range(data.shape[0]):
    print(data.iloc[i])

# 行を追加
s = pd.Series(['Jiro',30,75,180],index=data.index)
data = data.append(s,ignore_index=True)

# 新規列名を指定して追加
data['BMI'] = [20, 30, 25]
print(data)

# 行、列を指定して編集
data.iloc[0,4] = 40
print(data)
```



introduction.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ

+ コード + テキスト

目次 コードスニペット ファイル X

⬆ アップロード ⌂ 更新 🗑️ ドライブをマウント

📁 ..

📁 sample_data

> [12] 1
2
3
4
4
2
3
4

▶ # 外部データの読み込み

```
# pandas
import pandas as pd
data = pd.read_csv(data_dir+"some.csv")
# 好きな行を見る
print(data.iloc[0])
print(data.iloc[0:2])
# 好きな列を見る
print(data.iloc[:,0])
print(data.iloc[:,0:3])

print(data.columns)

# 項目を指定して見る
print(data.name)
print(data.height)

# 項目を指定して好きな行を見る
print(data.iloc[0].weight)
print(data.weight[1])

for i in range(data.shape[0]):
    print(data.iloc[i])

# 行を追加
s = pd.Series(['Jiro',30,75,180],index=data.index)
data = data.append(s,ignore_index=True)

# 新規列名を指定して追加
data['BMI'] = [20, 30, 25]
print(data)

# 行、列を指定して編集
data.iloc[0,4] = 40
print(data)
```

右クリック

Fileのアップロード



The screenshot shows the JupyterLab interface. The 'File' menu is open, and the 'Upload' option is highlighted. A blue arrow points from the 'Upload' option to the 'some.csv' file in the file explorer. A speech bubble with the word 'クリック' (Click) points to the 'Upload' option.

introduction.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ

+ コード + テキスト

目次 コード ファイル

アップロード

sample_data

- README
- anscom
- californi
- californi
- mnist_test.csv
- mnist_train_small.csv

クリック

アップロード

パスをコピー

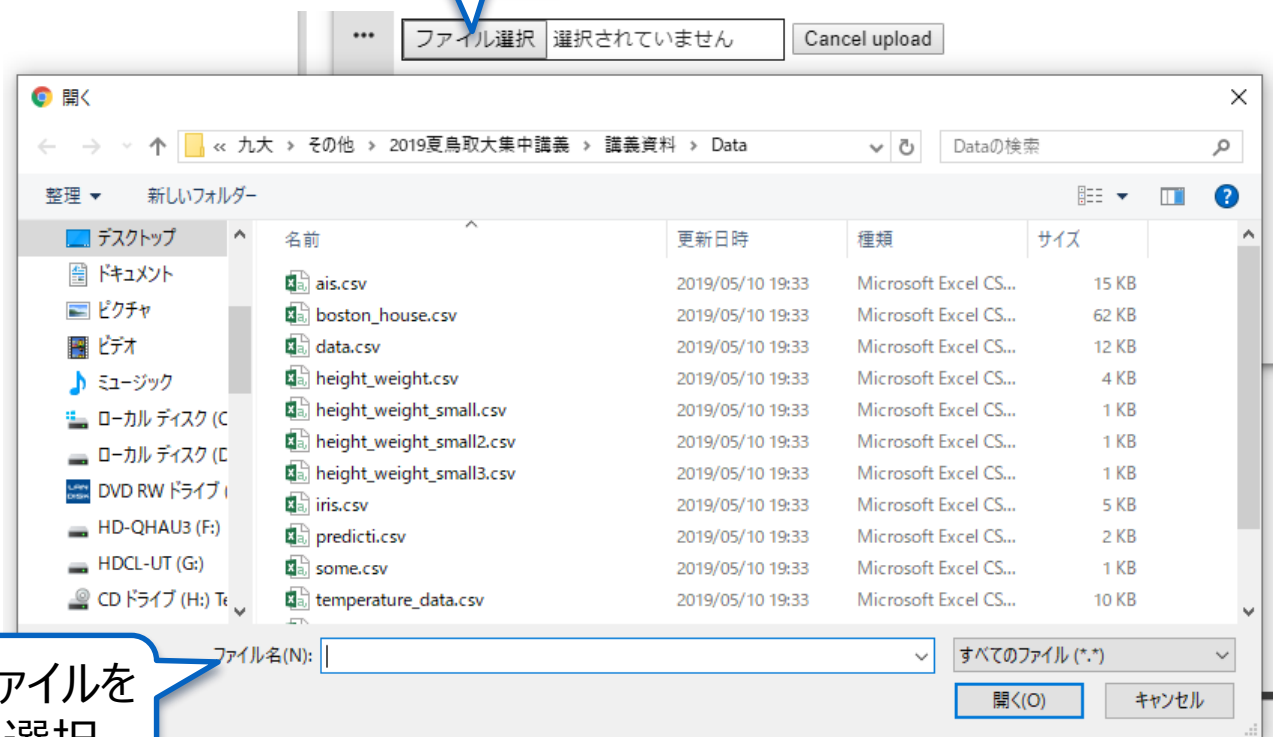
更新

```
[12] 1
      2
      3
      4
      5
      6
      7
      8
      9
     10
     11
     12
     13
     14
     15
     16
     17
     18
     19
     20
     21
     22
     23
     24
     25
     26
     27
     28
     29
     30
     31
     32
     33
     34
     35
     36
     37
     38
     39
     40
     41
     42
     43
     44
     45
     46
     47
     48
     49
     50
     51
     52
     53
     54
     55
     56
     57
     58
     59
     60
     61
     62
     63
     64
     65
     66
     67
     68
     69
     70
     71
     72
     73
     74
     75
     76
     77
     78
     79
     80
     81
     82
     83
     84
     85
     86
     87
     88
     89
     90
     91
     92
     93
     94
     95
     96
     97
     98
     99
    100
    101
    102
    103
    104
    105
    106
    107
    108
    109
    110
    111
    112
    113
    114
    115
    116
    117
    118
    119
    120
    121
    122
    123
    124
    125
    126
    127
    128
    129
    130
    131
    132
    133
    134
    135
    136
    137
    138
    139
    140
    141
    142
    143
    144
    145
    146
    147
    148
    149
    150
    151
    152
    153
    154
    155
    156
    157
    158
    159
    160
    161
    162
    163
    164
    165
    166
    167
    168
    169
    170
    171
    172
    173
    174
    175
    176
    177
    178
    179
    180
    181
    182
    183
    184
    185
    186
    187
    188
    189
    190
    191
    192
    193
    194
    195
    196
    197
    198
    199
    200
    201
    202
    203
    204
    205
    206
    207
    208
    209
    210
    211
    212
    213
    214
    215
    216
    217
    218
    219
    220
    221
    222
    223
    224
    225
    226
    227
    228
    229
    230
    231
    232
    233
    234
    235
    236
    237
    238
    239
    240
    241
    242
    243
    244
    245
    246
    247
    248
    249
    250
    251
    252
    253
    254
    255
    256
    257
    258
    259
    260
    261
    262
    263
    264
    265
    266
    267
    268
    269
    270
    271
    272
    273
    274
    275
    276
    277
    278
    279
    280
    281
    282
    283
    284
    285
    286
    287
    288
    289
    290
    291
    292
    293
    294
    295
    296
    297
    298
    299
    300
    301
    302
    303
    304
    305
    306
    307
    308
    309
    310
    311
    312
    313
    314
    315
    316
    317
    318
    319
    320
    321
    322
    323
    324
    325
    326
    327
    328
    329
    330
    331
    332
    333
    334
    335
    336
    337
    338
    339
    340
    341
    342
    343
    344
    345
    346
    347
    348
    349
    350
    351
    352
    353
    354
    355
    356
    357
    358
    359
    360
    361
    362
    363
    364
    365
    366
    367
    368
    369
    370
    371
    372
    373
    374
    375
    376
    377
    378
    379
    380
    381
    382
    383
    384
    385
    386
    387
    388
    389
    390
    391
    392
    393
    394
    395
    396
    397
    398
    399
    400
    401
    402
    403
    404
    405
    406
    407
    408
    409
    410
    411
    412
    413
    414
    415
    416
    417
    418
    419
    420
    421
    422
    423
    424
    425
    426
    427
    428
    429
    430
    431
    432
    433
    434
    435
    436
    437
    438
    439
    440
    441
    442
    443
    444
    445
    446
    447
    448
    449
    450
    451
    452
    453
    454
    455
    456
    457
    458
    459
    460
    461
    462
    463
    464
    465
    466
    467
    468
    469
    470
    471
    472
    473
    474
    475
    476
    477
    478
    479
    480
    481
    482
    483
    484
    485
    486
    487
    488
    489
    490
    491
    492
    493
    494
    495
    496
    497
    498
    499
    500
    501
    502
    503
    504
    505
    506
    507
    508
    509
    510
    511
    512
    513
    514
    515
    516
    517
    518
    519
    520
    521
    522
    523
    524
    525
    526
    527
    528
    529
    530
    531
    532
    533
    534
    535
    536
    537
    538
    539
    540
    541
    542
    543
    544
    545
    546
    547
    548
    549
    550
    551
    552
    553
    554
    555
    556
    557
    558
    559
    560
    561
    562
    563
    564
    565
    566
    567
    568
    569
    570
    571
    572
    573
    574
    575
    576
    577
    578
    579
    580
    581
    582
    583
    584
    585
    586
    587
    588
    589
    590
    591
    592
    593
    594
    595
    596
    597
    598
    599
    600
    601
    602
    603
    604
    605
    606
    607
    608
    609
    610
    611
    612
    613
    614
    615
    616
    617
    618
    619
    620
    621
    622
    623
    624
    625
    626
    627
    628
    629
    630
    631
    632
    633
    634
    635
    636
    637
    638
    639
    640
    641
    642
    643
    644
    645
    646
    647
    648
    649
    650
    651
    652
    653
    654
    655
    656
    657
    658
    659
    660
    661
    662
    663
    664
    665
    666
    667
    668
    669
    670
    671
    672
    673
    674
    675
    676
    677
    678
    679
    680
    681
    682
    683
    684
    685
    686
    687
    688
    689
    690
    691
    692
    693
    694
    695
    696
    697
    698
    699
    700
    701
    702
    703
    704
    705
    706
    707
    708
    709
    710
    711
    712
    713
    714
    715
    716
    717
    718
    719
    720
    721
    722
    723
    724
    725
    726
    727
    728
    729
    730
    731
    732
    733
    734
    735
    736
    737
    738
    739
    740
    741
    742
    743
    744
    745
    746
    747
    748
    749
    750
    751
    752
    753
    754
    755
    756
    757
    758
    759
    760
    761
    762
    763
    764
    765
    766
    767
    768
    769
    770
    771
    772
    773
    774
    775
    776
    777
    778
    779
    780
    781
    782
    783
    784
    785
    786
    787
    788
    789
    790
    791
    792
    793
    794
    795
    796
    797
    798
    799
    800
    801
    802
    803
    804
    805
    806
    807
    808
    809
    810
    811
    812
    813
    814
    815
    816
    817
    818
    819
    820
    821
    822
    823
    824
    825
    826
    827
    828
    829
    830
    831
    832
    833
    834
    835
    836
    837
    838
    839
    840
    841
    842
    843
    844
    845
    846
    847
    848
    849
    850
    851
    852
    853
    854
    855
    856
    857
    858
    859
    860
    861
    862
    863
    864
    865
    866
    867
    868
    869
    870
    871
    872
    873
    874
    875
    876
    877
    878
    879
    880
    881
    882
    883
    884
    885
    886
    887
    888
    889
    890
    891
    892
    893
    894
    895
    896
    897
    898
    899
    900
    901
    902
    903
    904
    905
    906
    907
    908
    909
    910
    911
    912
    913
    914
    915
    916
    917
    918
    919
    920
    921
    922
    923
    924
    925
    926
    927
    928
    929
    930
    931
    932
    933
    934
    935
    936
    937
    938
    939
    940
    941
    942
    943
    944
    945
    946
    947
    948
    949
    950
    951
    952
    953
    954
    955
    956
    957
    958
    959
    960
    961
    962
    963
    964
    965
    966
    967
    968
    969
    970
    971
    972
    973
    974
    975
    976
    977
    978
    979
    980
    981
    982
    983
    984
    985
    986
    987
    988
    989
    990
    991
    992
    993
    994
    995
    996
    997
    998
    999
    1000
    1001
    1002
    1003
    1004
    1005
    1006
    1007
    1008
    1009
    1010
    1011
    1012
    1013
    1014
    1015
    1016
    1017
    1018
    1019
    1020
    1021
    1022
    1023
    1024
    1025
    1026
    1027
    1028
    1029
    1030
    1031
    1032
    1033
    1034
    1035
    1036
    1037
    1038
    1039
    1040
    1041
    1042
    1043
    1044
    1045
    1046
    1047
    1048
    1049
    1050
    1051
    1052
    1053
    1054
    1055
    1056
    1057
    1058
    1059
    1060
    1061
    1062
    1063
    1064
    1065
    1066
    1067
    1068
    1069
    1070
    1071
    1072
    1073
    1074
    1075
    1076
    1077
    1078
    1079
    1080
    1081
    1082
    1083
    1084
    1085
    1086
    1087
    1088
    1089
    1090
    1091
    1092
    1093
    1094
    1095
    1096
    1097
    1098
    1099
    1100
    1101
    1102
    1103
    1104
    1105
    1106
    1107
    1108
    1109
    1110
    1111
    1112
    1113
    1114
    1115
    1116
    1117
    1118
    1119
    1120
    1121
    1122
    1123
    1124
    1125
    1126
    1127
    1128
    1129
    1130
    1131
    1132
    1133
    1134
    1135
    1136
    1137
    1138
    1139
    1140
    1141
    1142
    1143
    1144
    1145
    1146
    1147
    1148
    1149
    1150
    1151
    1152
    1153
    1154
    1155
    1156
    1157
    1158
    1159
    1160
    1161
    1162
    1163
    1164
    1165
    1166
    1167
    1168
    1169
    1170
    1171
    1172
    1173
    1174
    1175
    1176
    1177
    1178
    1179
    1180
    1181
    1182
    1183
    1184
    1185
    1186
    1187
    1188
    1189
    1190
    1191
    1192
    1193
    1194
    1195
    1196
    1197
    1198
    1199
    1200
    1201
    1202
    1203
    1204
    1205
    1206
    1207
    1208
    1209
    1210
    1211
    1212
    1213
    1214
    1215
    1216
    1217
    1218
    1219
    1220
    1221
    1222
    1223
    1224
    1225
    1226
    1227
    1228
    1229
    1230
    1231
    1232
    1233
    1234
    1235
    1236
    1237
    1238
    1239
    1240
    1241
    1242
    1243
    1244
    1245
    1246
    1247
    1248
    1249
    1250
    1251
    1252
    1253
    1254
    1255
    1256
    1257
    1258
    1259
    1260
    1261
    1262
    1263
    1264
    1265
    1266
    1267
    1268
    1269
    1270
    1271
    1272
    1273
    1274
    1275
    1276
    1277
    1278
    1279
    1280
    1281
    1282
    1283
    1284
    1285
    1286
    1287
    1288
    1289
    1290
    1291
    1292
    1293
    1294
    1295
    1296
    1297
    1298
    1299
    1300
    1301
    1302
    1303
    1304
    1305
    1306
    1307
    1308
    1309
    1310
    1311
    1312
    1313
    1314
    1315
    1316
    1317
    1318
    1319
    1320
    1321
    1322
    1323
    1324
    1325
    1326
    1327
    1328
    1329
    1330
    1331
    1332
    1333
    1334
    1335
    1336
    1337
    1338
    1339
    1340
    1341
    1342
    1343
    1344
    1345
    1346
    1347
    1348
    1349
    1350
    1351
    1352
    1353
    1354
    1355
    1356
    1357
    1358
    1359
    1360
    1361
    1362
    1363
    1364
    1365
    1366
    1367
    1368
    1369
    1370
    1371
    1372
    1373
    1374
    1375
    1376
    1377
    1378
    1379
    1380
    1381
    1382
    1383
    1384
    1385
    1386
    1387
    1388
    1389
    1390
    1391
    1392
    1393
    1394
    1395
    1396
    1397
    1398
    1399
    1400
    1401
    1402
    1403
    1404
    1405
    1406
    1407
    1408
    1409
    1410
    1411
    1412
    1413
    1414
    1415
    1416
    1417
    1418
    1419
    1420
    1421
    1422
    1423
    1424
    1425
    1426
    1427
    1428
    1429
    1430
    1431
    1432
    1433
    1434
    1435
    1436
    1437
    1438
    1439
    1440
    1441
    1442
    1443
    1444
    1445
    1446
    1447
    1448
    1449
    1450
    1451
    1452
    1453
    1454
    1455
    1456
    1457
    1458
    1459
    1460
    1461
    1462
    1463
    1464
    1465
    1466
    1467
    1468
    1469
    1470
    1471
    1472
    1473
    1474
    1475
    1476
    1477
    1478
    1479
    1480
    1481
    1482
    1483
    1484
    1485
    1486
    1487
    1488
    1489
    1490
    1491
    1492
    1493
    1494
    1495
    1496
    1497
    1498
    1499
    1500
    1501
    1502
    1503
    1504
    1505
    1506
    1507
    1508
    1509
    1510
    1511
    1512
    1513
    1514
    1515
    1516
    1517
    1518
    1519
    1520
    1521
    1522
    1523
    1524
    1525
    1526
    1527
    1528
    1529
    1530
    1531
    1532
    1533
    1534
    1535
    1536
    1537
    1538
    1539
    1540
    1541
    1542
    1543
    1544
    1545
    1546
    1547
    1548
    1549
    1550
    1551
    1552
    1553
    1554
    1555
    1556
    1557
    1558
    1559
    1560
    1561
    1562
    1563
    1564
    1565
    1566
    1567
    1568
    1569
    1570
    1571
    1572
    1573
    1574
    1575
    1576
    1577
    1578
    1579
    1580
    1581
    1582
    1583
    1584
    1585
    1586
    1587
    1588
    1589
    1590
    1591
    1592
    1593
    1594
    1595
    1596
    1597
    1598
    1599
    1600
    1601
    1602
    1603
    1604
    1605
    1606
    1607
    1608
    1609
    1610
    1611
    1612
    1613
    1614
    1615
    1616
    1617
    1618
    1619
    1620
    1621
    1622
    1623
    1624
    1625
    1626
    1627
    1628
    1629
    1630
    1631
    1632
    1633
    1634
    1635
    1636
    1637
    1638
    1639
    1640
    1641
    1642
    1643
    1644
    1645
    1646
    1647
    1648
    1649
    1650
    1651
    1652
    1653
    1654
    1655
    1656
    1657
    1658
    1659
    1660
    1661
    1662
    1663
    1664
    1665
    1666
    1667
    1668
    1669
    1670
    1671
    1672
    1673
    1674
    1675
    1676
    1677
    1678
    1679
    1680
    1681
    1682
    1683
    1684
    1685
    1686
    1687
    1688
    1689
    1690
    1691
    1692
    1693
    1694
    1695
    1696
    1697
    1698
    1699
    1700
    1701
    1702
    1703
    1704
    1705
    1706
    1707
    1708
    1709
    1710
    1711
    1712
    1713
    1714
    1715
    1716
    1717
    1718
    1719
    1720
    1721
    1722
    1723
    1724
    1725
    1726
    1727
    1728
    1729
    1730
    1731
    1732
    1733
    1734
    1735
    1736
    1737
    1738
    1739
    1740
    1741
    1742
    1743
    1744
    1745
    1746
    1747
    1748
    1749
    1750
    1751
    1752
    1753
    1754
    1755
    1756
    1757
    1758
    1759
    1760
    1761
    1762
    1763
    1764
    1765
    1766
    1767
    1768
    1769
    1770
    1771
    1772
    1773
    1774
    1775
    1776
    1777
    1778
    1779
    1780
    1781
    1782
    1783
    1784
    1785
    1786
    1787
    1788
    1789
    1790
    1791
    1792
    1793
    1794
    1795
    1796
    1797
    1798
    1799
    1800
    1801
    1802
    1803
    1804
    1805
    1806
    1807
    1808
    1809
    1810
    1811
    1812
    1813
    1814
    1815
    1816
    1817
    1818
    1819
    1820
    1821
    1822
    1823
    1824
    1825
    1826
    1827
    1828
    1829
    1830
    1831
    1832
    1833
    1834
    1835
    1836
    1837
    1838
    1839
    1840
    1841
    1842
    1843
    1844
    1845
    1846
    1847
    1848
    1849
    1850
    1851
    1852
    1853
    1854
    1855
    1856
    1857
    1858
    1859
    1860
    1861
    1862
    1863
    1864
    1865
    1866
    1867
    1868
    1869
    1870
    1871
    1872
    1873
    1874
    1875
    1876
    1877
    1878
    1879
    1880
    1
```

Fileのアップロード: コマンド

from google.colab import files
files.upload()

クリック



PythonによるCSVファイルの読み込み

- CSVファイル読み込み用の道具をPython上で呼び出す

```
import pandas as pd
```

CSVファイルの読み込み

● CSVファイルを読み出すおまじない

注) 他のフォルダにファイルがある場合
`pd.read_csv("path名¥ファイル名")`
Windowsでは、¥¥とバックスラッシュ2つ
でないと動かない場合もある。

CSVファイルの中身の
データが格納

Name	age	weight	height
Bob	40	80	175
Taro	35	65	170

読み込むCSVファイルのパス

```
data = pd.read_csv("some.csv")
```

読み込むCSVファイルの名前

以下を事前に行うことを忘れないように！

- (1) **`import pandas as pd`**
- (2) "some.csv" をコードと同じフォルダに置く

データの見方1

●好きな行を見る

```
In [399]: print(data.iloc[0])  
name      Bob  
age       40  
weight    80  
height    175  
Name: 0, dtype: object
```

```
In [400]: print(data.iloc[0:2])  
   name  age  weight  height  
0   Bob   40     80     175  
1  Taro   35     65     170
```

↑ 0,1行目を見る

●好きな列を見る

```
In [401]: print(data.iloc[:,0])  
0      Bob  
1     Taro  
Name: name, dtype: object
```

```
In [402]: print(data.iloc[:,0:3])  
   name  age  weight  
0   Bob   40     80  
1  Taro   35     65
```

↑ 0,1,2列目を見る

データの見方2

- 項目(index)を指定してみる

↓ 項目を指定

```
In [406]: print(data.name)
0      Bob
1      Taro
Name: name, dtype: object
```

```
In [407]: print(data.height)
0      175
1      170
Name: height, dtype: int64
```

- 項目を指定して好きな行を見る

どちらもO.K.

```
In [125]: print(data.iloc[0].weight)
80
```

```
In [126]: print(data.weight[1])
65
```

行を指定して項目を指定

項目を指定して行を指定

データの見方2

- 項目(index)を指定してみる

↓ 項目を指定

```
In [406]: print(data.name)
0      Bob
1     Taro
Name: name, dtype: object
```

```
In [407]: print(data.height)
0     175
1     170
Name: height, dtype: int64
```

- 項目を指定して好きな行を見る

どちらでもO.K.

```
In [125]: print(data.iloc[0].weight)
80
```

```
In [126]: print(data.weight[1])
65
```

行を指定して項目を指定

項目を指定して行を指定

一個一個見てたらキリがない！

データの可視化（プロット）

- プロットには色々な方法があります
 - 折れ線グラフ
 - ヒストグラム
 - 散布図
 - etc.

はじめに

- 今回はサンプルとして次の2つのデータを使ってみます
- 那覇, 福岡, 札幌の気温データ
→ temperature_data.csv
<http://www.data.jma.go.jp/gmd/risk/obsdl/index.php>
- 体重身長データ (オーストラリア国立スポーツ研究所の男女202人のアスリートのデータ)
→ height_weight.csv
<http://www.statsci.org/data/oz/ais.html>

おまじないと準備

- import pandas as pd
- データの読み込み

```
In [66]: temper_data=pd.read_csv("../DS_enshu1/temperature_data.csv")
```

```
In [67]: HW_data=pd.read_csv("../DS_enshu1/height_weight.csv")
```



ファイルの場所は各自置いた場所を
指定しよう

読み込んだデータの中身

In [68]: temper_data

```
Out[68]:
```

	date	Naha	Fukuoka	Sapporo
0	2016/1/1	18.1	7.3	-1.1
1	2016/1/2	20.2	11.6	1.6
2	2016/1/3	21.2	11.6	0.3
3	2016/1/4	19.8	11.0	-1.7
4	2016/1/5	22.5	9.9	-3.9
5	2016/1/6	19.5	9.5	-2.3
6	2016/1/7	19.1	9.1	-2.7
7	2016/1/8	16.6	7.5	-2.5
8	2016/1/9	17.1	7.9	-3.6
9	2016/1/10	18.3	7.9	-4.2
10	2016/1/11	19.4	8.3	-6.7
11	2016/1/12	18.2	7.5	-6.4
12	2016/1/13	15.9	5.8	-4.2
13	2016/1/14	15.1	6.7	-3.7
14	2016/1/15	16.6	5.8	-6.3
15	2016/1/16	17.5	7.5	-5.0
16	2016/1/17	19.5	7.1	-5.1
17	2016/1/18	17.6	6.7	-5.9
18	2016/1/19	15.3	2.3	-1.4
19	2016/1/20	15.0	2.7	-0.5
20	2016/1/21	17.3	2.8	-2.1
21	2016/1/22	16.8	5.8	-2.7
22	2016/1/23	14.7	3.5	-3.7
23	2016/1/24	9.2	-2.0	-5.7
24	2016/1/25	10.4	1.2	-6.7
25	2016/1/26	12.6	4.4	0.0
26	2016/1/27	15.7	6.6	-1.7
27	2016/1/28	20.2	8.3	-3.9
28	2016/1/29	22.2	11.0	-5.0
29	2016/1/30	18.6	11.8	-5.8
...
336	2016/12/2	21.5	11.6	1.1
337	2016/12/3	22.5	10.6	3.6
338	2016/12/4	24.1	11.3	7.1
339	2016/12/5	23.3	12.4	6.3
340	2016/12/6	21.3	11.3	-2.1
341	2016/12/7	19.8	8.8	-5.0
342	2016/12/8	20.7	10.6	-2.5
343	2016/12/9	21.6	12.4	-0.3
344	2016/12/10	20.9	10.6	-2.6
345	2016/12/11	20.3	9.4	-4.6
346	2016/12/12	21.9	11.0	-4.1
347	2016/12/13	22.8	12.3	0.2
348	2016/12/14	20.0	10.8	-2.0
349	2016/12/15	17.4	7.9	-4.5
350	2016/12/16	16.6	6.1	-5.3
351	2016/12/17	17.8	7.4	-3.6
352	2016/12/18	20.7	8.9	1.0
353	2016/12/19	22.2	12.4	1.0
354	2016/12/20	23.4	14.0	-0.1
355	2016/12/21	23.6	16.1	2.7
356	2016/12/22	22.7	15.8	0.3
357	2016/12/23	18.9	9.7	-1.4
358	2016/12/24	18.5	9.1	-4.4
359	2016/12/25	20.2	10.8	-2.3
360	2016/12/26	22.3	13.3	1.3
361	2016/12/27	18.5	10.7	-2.5
362	2016/12/28	16.9	6.8	-6.4
363	2016/12/29	17.9	6.8	-3.6
364	2016/12/30	17.9	6.6	-2.8
365	2016/12/31	19.3	7.3	-2.6

[366 rows x 4 columns]

In [70]: HW_data

```
Out[70]:
```

	Unnamed: 0	Sex	Ht	Wt
0	1	1	195.9	78.9
1	2	1	189.7	74.4
2	3	1	177.8	69.1
3	4	1	185.0	74.9
4	5	1	184.6	64.6
5	6	1	174.0	63.7
6	7	1	186.2	75.2
7	8	1	173.8	62.3
8	9	1	171.4	66.5
9	10	1	179.9	62.9
10	11	1	193.4	96.3
11	12	1	188.7	75.5
12	13	1	169.1	63.0
13	14	1	177.9	80.5
14	15	1	177.5	71.3
15	16	1	179.6	70.5
16	17	1	181.3	73.2
17	18	1	179.7	68.7
18	19	1	185.2	80.5
19	20	1	177.3	72.9
20	21	1	179.3	74.5
21	22	1	175.3	75.4
22	23	1	174.0	69.5
23	24	1	183.3	66.4
24	25	1	184.7	79.7
25	26	1	180.2	73.6
26	27	1	180.2	78.7
27	28	1	176.0	75.0
28	29	1	156.0	49.8
29	30	1	179.7	67.2
...
172	173	0	178.5	71.0
173	174	0	171.3	69.1
174	175	0	178.0	62.9
175	176	0	189.1	94.8
176	177	0	195.4	94.6
177	178	0	179.1	108.2
178	179	0	180.1	97.9
179	180	0	179.6	75.2
180	181	0	174.7	74.8
181	182	0	192.7	94.2
182	183	0	179.3	76.1
183	184	0	197.5	94.7
184	185	0	182.7	86.2
185	186	0	190.5	79.6
186	187	0	191.0	85.3
187	188	0	179.6	74.4
188	189	0	192.6	93.5
189	190	0	194.1	87.6
190	191	0	193.0	85.4
191	192	0	193.9	101.0
192	193	0	187.7	74.9
193	194	0	185.3	87.3
194	195	0	191.5	90.0
195	196	0	184.6	94.7
196	197	0	179.9	76.3
197	198	0	183.9	93.2
198	199	0	183.5	80.0
199	200	0	183.1	73.8
200	201	0	178.4	71.1
201	202	0	190.8	76.7

[202 rows x 4 columns]

- 「データフレーム」という形式で保存
- 行列（縦が行，横が列）
- 気温データは
 - 縦が時間
 - 横が地方
- 身長体重データは
 - 縦が人（サンプル）
 - 横が項目（性別，身長，体重）

データの見方

●好きな行をみる

```
In [79]: temper_data.iloc[0]
```

```
Out[79]:
```

```
date      2016/1/1
Naha      18.1
Fukuoka   7.3
Sapporo   -1.1
Name: 0, dtype: object
```

```
In [82]: temper_data.iloc[:3]
```

```
Out[82]:
```

	date	Naha	Fukuoka	Sapporo
0	2016/1/1	18.1	7.3	-1.1
1	2016/1/2	20.2	11.6	1.6
2	2016/1/3	21.2	11.6	0.3

●好きな列をみる

↑ 0,1,2行目をみる

```
In [80]: temper_data.iloc[:,0]
```

```
In [84]: temper_data.iloc[:,1:3]
```

↑ 1,2列目をみる

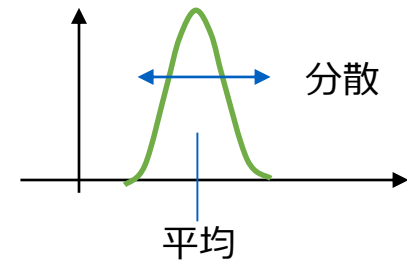
簡単な統計値を出す

● 平均値

```
In [94]: temper_data.mean()  
Out[94]:  
Naha      24.131421  
Fukuoka   18.091803  
Sapporo    9.295902  
dtype: float64
```

● 分散

```
In [95]: temper_data.var()  
Out[95]:  
Naha      24.559640  
Fukuoka   62.146179  
Sapporo   95.555079  
dtype: float64
```



● 標準偏差

```
In [96]: temper_data.std()  
Out[96]:  
Naha      4.955768  
Fukuoka   7.883285  
Sapporo    9.775228  
dtype: float64
```

まとめてみることもできます

- 統計量のまとめを表示

```
temper_data.describe()
```

	Naha	Fukuoka	Sapporo
count	366.000000	366.000000	366.000000
mean	24.131421	18.091803	9.295902
std	4.955768	7.883285	9.775228
min	9.200000	-2.000000	-6.800000
25%	20.225000	11.025000	0.300000
50%	25.050000	18.500000	8.650000
75%	28.600000	24.650000	18.050000
max	30.800000	31.400000	27.100000

データの25%は～以下

データの50%は～以下

データの75%は～以下（上位25%は～以上）

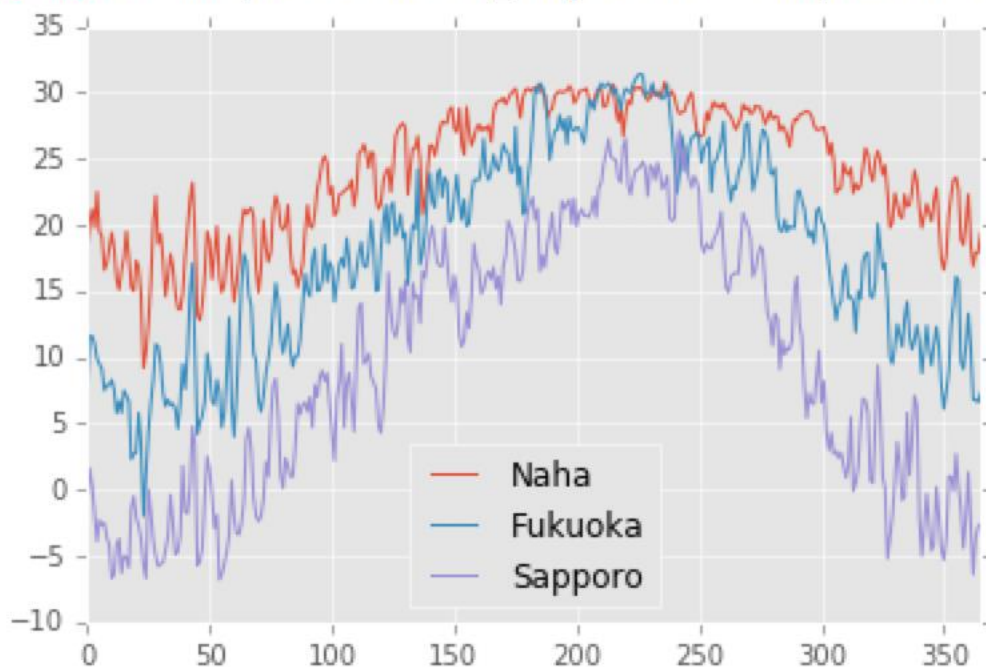
データの可視化（プロット）

折れ線グラフ

●折れ線グラフ

```
In [97]: temper_data.plot()
```

```
Out[97]: <matplotlib.axes._subplots.AxesSubplot at 0x120dc2050>
```



グラフの横軸が行の番号 (index) でわかりにくい

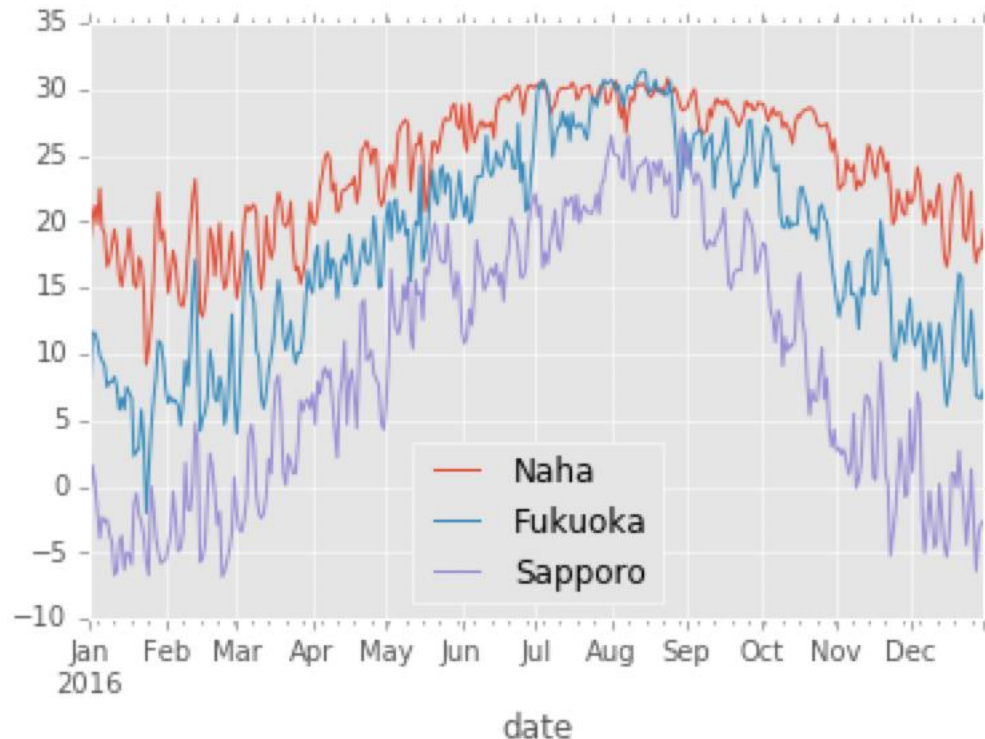
折れ線グラフ

●折れ線グラフ

```
In [98]: temper_data.index = pd.to_datetime(temper_data.date)
```

```
In [99]: temper_data.plot()
```

```
Out[99]: <matplotlib.axes._subplots.AxesSubplot at 0x120d95910>
```



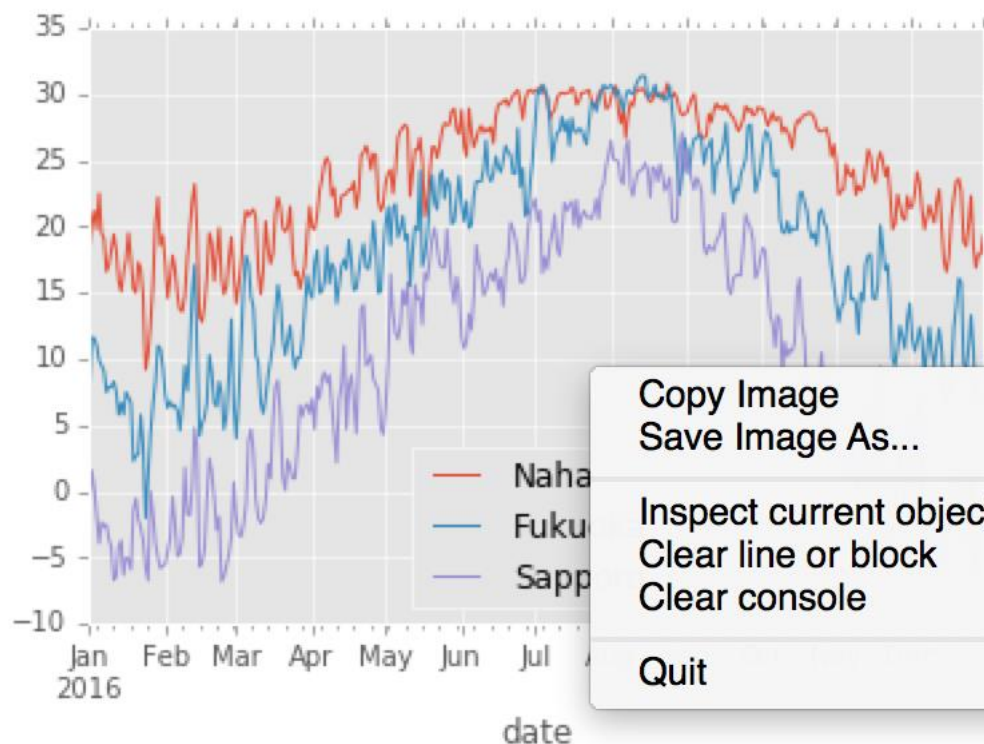
indexを date に変更

折れ線グラフ

●折れ線グラフ

```
In [99]: temper_data.plot()
```

```
Out[99]: <matplotlib.axes._subplots.AxesSubplot at 0x120d95910>
```



右クリックすれば
保存もできます

Copy Image
Save Image As...

Inspect current object

⌘I

Clear line or block

⇧⌘

Clear console

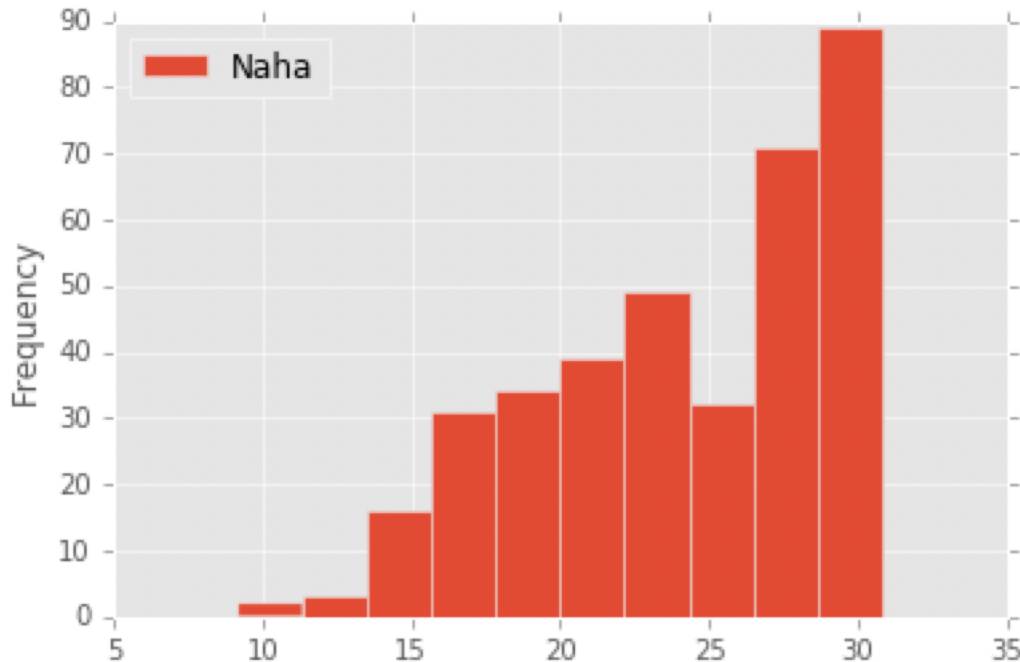
⌘L

Quit

ヒストグラム

- ヒストグラム
- 那覇の気温の分布を見してみる

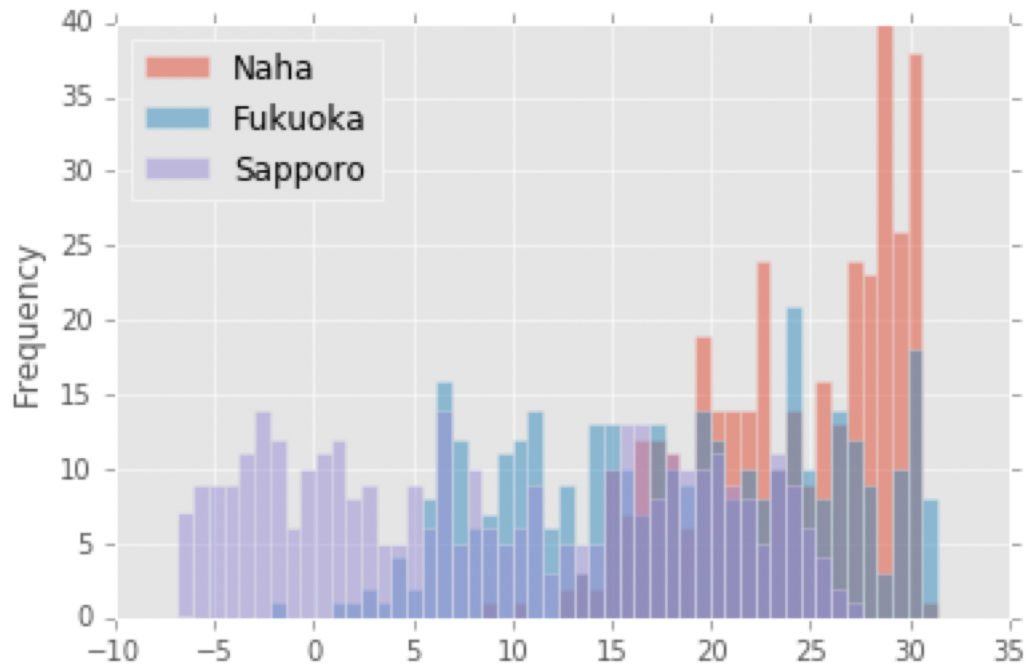
```
In [109]: temper_data.plot(y='Naha',kind='hist')  
Out[109]: <matplotlib.axes._subplots.AxesSubplot at 0x1281584d0>
```



ヒストグラム（発展）

- 那覇，福岡，札幌の分布を比較したい

```
In [115]: temper_data.plot.hist(bins=50, alpha=0.5)  
Out[115]: <matplotlib.axes._subplots.AxesSubplot at 0x128cdec0>
```



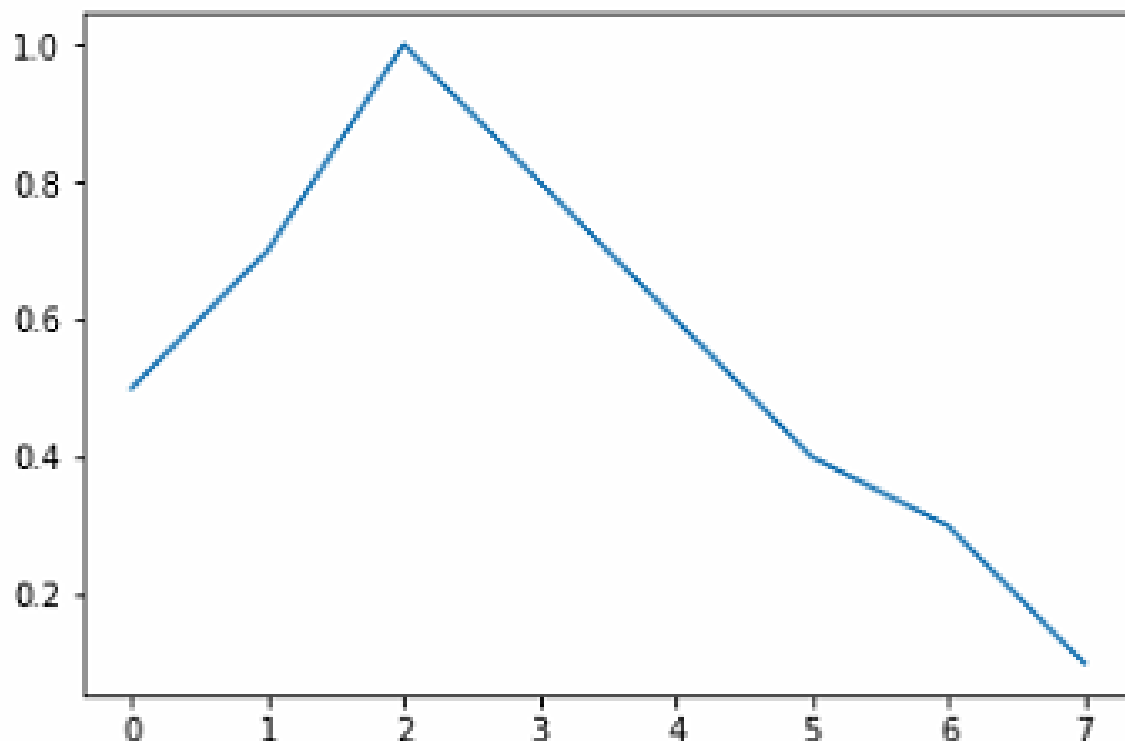
`data.plot(kind='hist')`
`data.plot.hist()`
は同じ処理

binsは棒の数， alpha は透明度

Matplotlibを使った可視化

● import matplotlib.pyplot as plt

```
In [12]: a = [0.5,0.7,1,0.8,0.6,0.4,0.3,0.1]
...: plt.plot(a)
Out[12]: [<matplotlib.lines.Line2D at 0x1d8b0bc9ba8>]
```



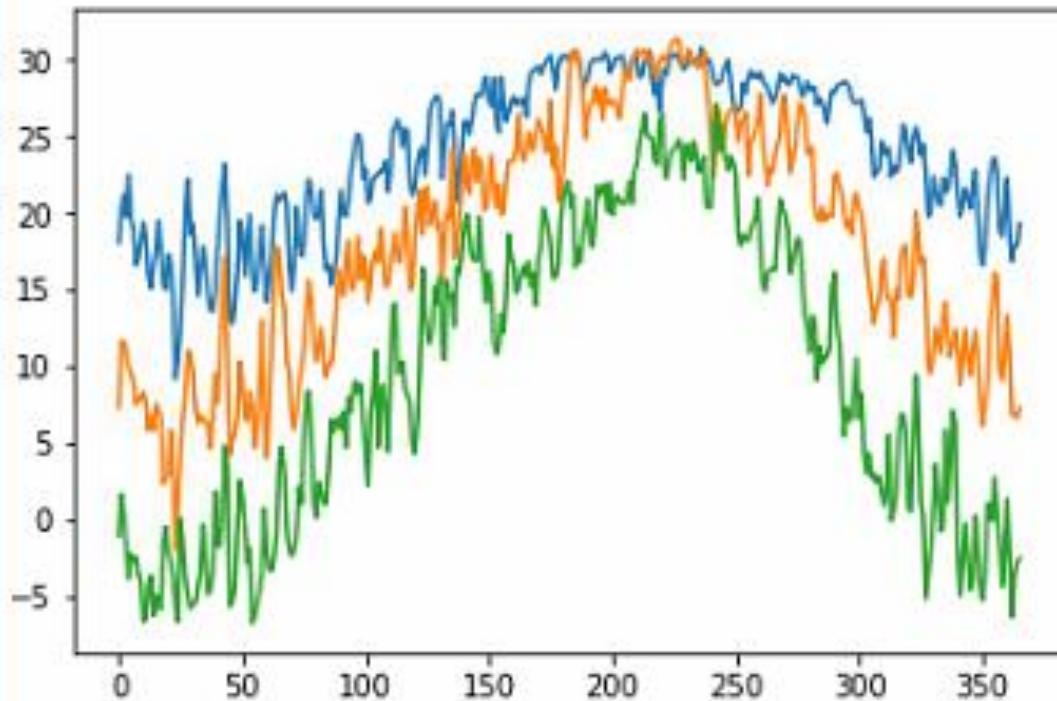
リストもしくは
numpyのarray型
を入力

pandas以外でも
使えます。
csvデータは基本的に
pandas の plotでも十分
ですが、より高度な
可視化（自由な可視化）
をしたいときには便利
（後ほど紹介）

Numpy array への変換

```
In [34]: Naha = temper_data.values[:,1]
...: Fukuoka = temper_data.values[:,2]
...: Sapporo = temper_data.values[:,3]
...:
...: plt.plot(Naha)
...: plt.plot(Fukuoka)
...: plt.plot(Sapporo)
Out[34]: [<matplotlib.lines.Line2D at 0x1d8b0b14d30>]
```

Pandas の values を使うと
numpy の array型として
抽出できる



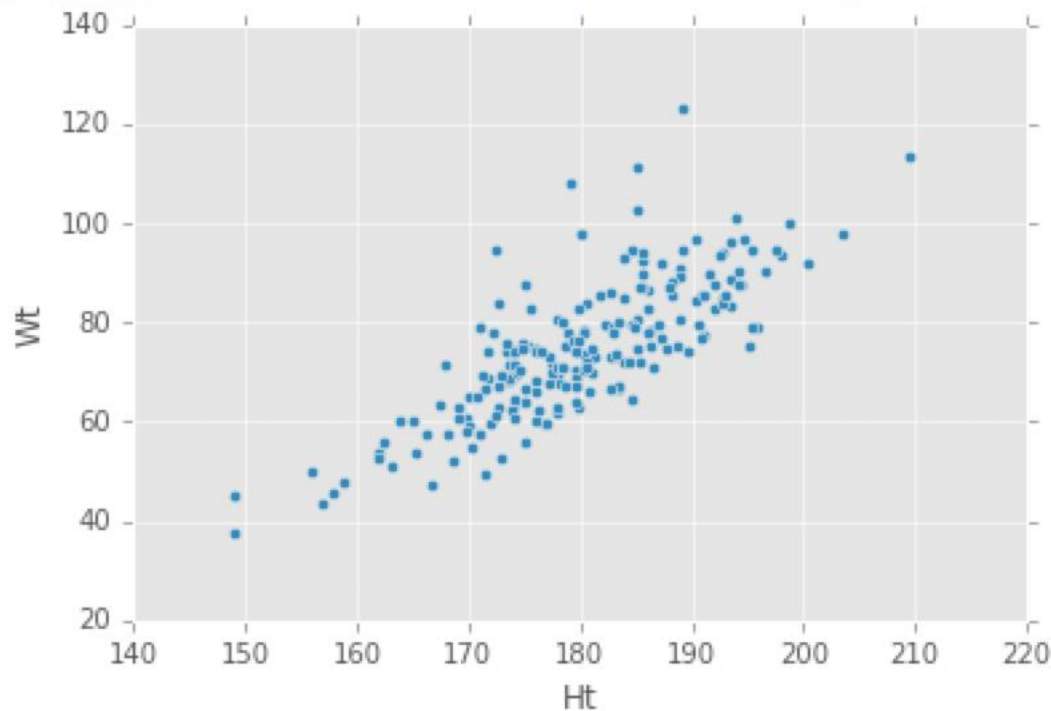
練習 1 : データの可視化

- 他の地方の分布も見よう
 - 福岡、札幌、那覇それぞれのヒストグラムを作成
- bins と alpha を変えて 3 地点の分布を比較してみよう
- 身長体重データでの統計値も見よう

散布図

- 身長と体重はどんな関係がある？

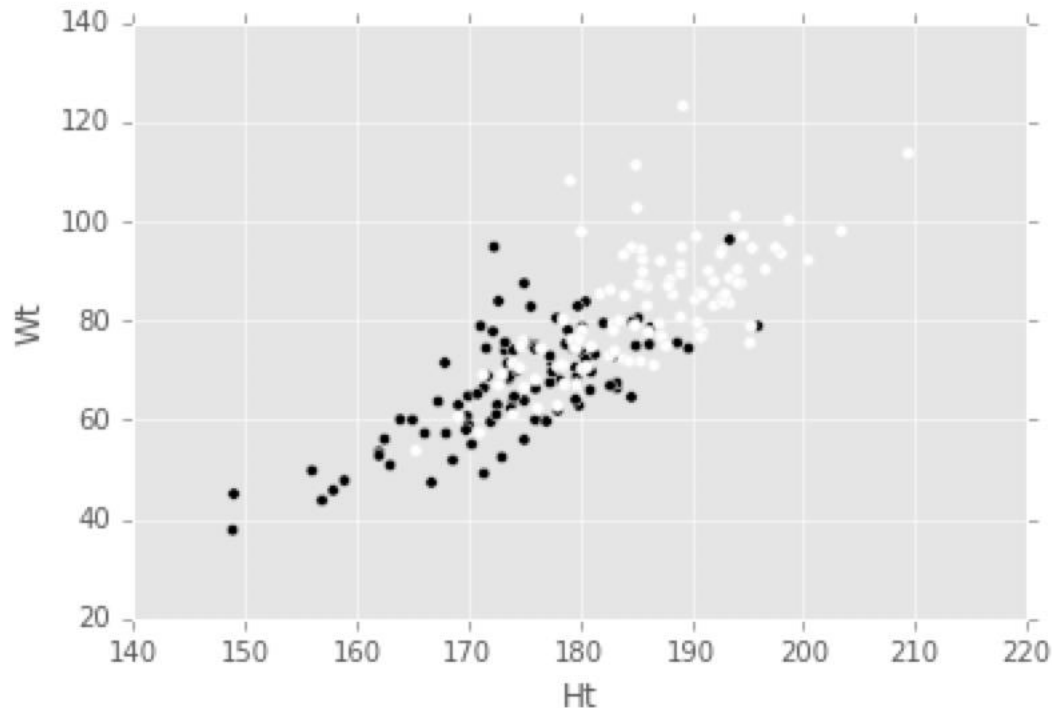
```
In [111]: HW_data.plot(x='Ht', y='Wt', kind='scatter')  
Out[111]: <matplotlib.axes._subplots.AxesSubplot at 0x12855f310>
```



散布図

- 身長と体重はどんな関係があるか？
- 性別でどんな違いがあるか？

```
In [122]: HW_data.plot(kind='scatter',x='Ht',y='Wt',c=HW_data.Sex)  
Out[122]: <matplotlib.axes._subplots.AxesSubplot at 0x12be06d10>
```



項目同士の関係性を見るのに便利なプロット

- 項目がたくさんあるとき、まとめて関係をプロット

```
import pandas as pd
wine_data = pd.read_csv("./winequality-red.csv", sep=";")
import seaborn as sns
sns.pairplot(wine_data)
```

- 実際に実行してみましょう
- 図は右クリックで保存できます



欠損値について

- csv の中に空白（欠損）があると, NaN（Not a Number）というものが入る
- サンプルファイルを読み込んでみましょう

```
In [92]: temper_data2=pd.read_csv("../DS_enshu1/temperature_data_NaN.csv")
```

```
In [93]: temper_data2
```

欠損値について

- csv の中に空白（欠損）があると, NaN（Not a Number）というものが入る
- 欠損値のあるデータ（行）を取り除く

```
In [130]: temper_data2.dropna(0)
```

```
Out [130]:
```

	date	Naha	Fukuoka	Sapporo
0	2016/1/1	18.1	7.3	-1.1
1	2016/1/2	20.2	11.6	1.6
2	2016/1/3	21.2	11.6	0.3
3	2016/1/4	19.8	11.0	-1.7
4	2016/1/5	22.5	9.9	-3.9
5	2016/1/6	19.5	9.5	-2.3
6	2016/1/7	19.1	9.1	-2.7
7	2016/1/8	16.6	7.5	-2.5
8	2016/1/9	17.1	7.9	-3.6
9	2016/1/10	18.3	7.9	-4.2
10	2016/1/11	19.4	8.3	-6.7
11	2016/1/12	18.2	7.5	-6.4
12	2016/1/13	15.9	5.8	-4.2
13	2016/1/14	15.1	6.7	-3.7
15	2016/1/16	17.5	7.5	-5.0

ベクトル表現と ベクトル演算の関数化

ベクトル

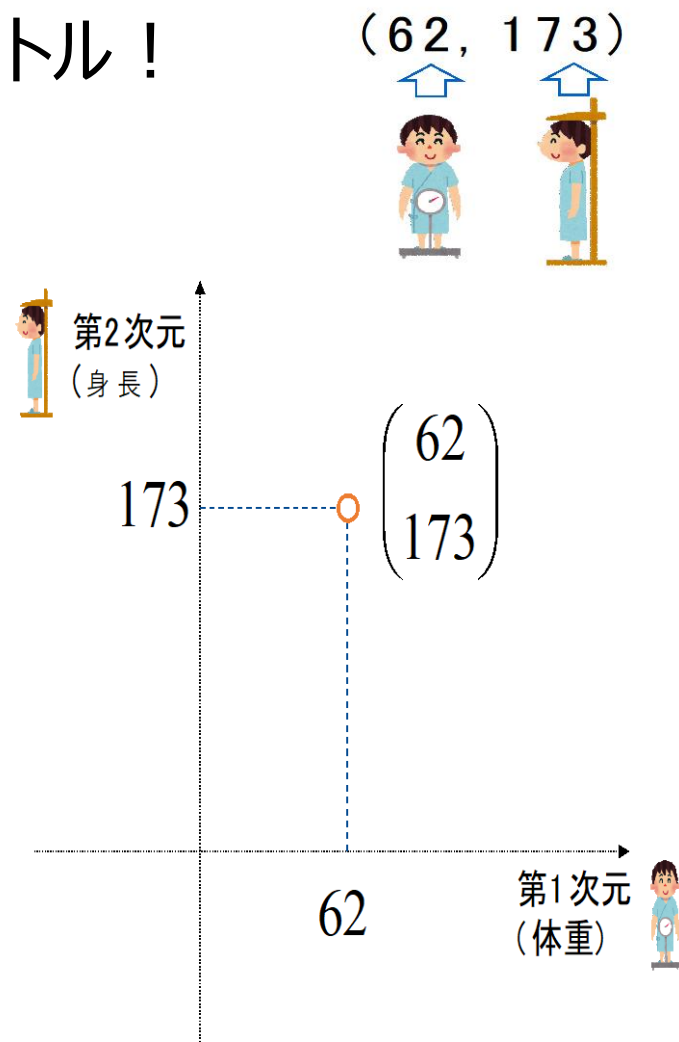
- データの多くは数字の組，つまりベクトル！

- 例： $x = [62, 173]$

- ベクトルで表現すると，

- データ間の「距離」

- 類似度
などが計算できました



ベクトルと python

- np.array を使うと色々な演算が楽
- まず「import numpy as np」を実行
- 例：

```
a=np.array([0,1,2])
```

```
print(a)
```

```
[0 1 2]
```

ベクトルの演算:和

- ベクトルの和（要素同士の和を取る）

$$\begin{array}{c} \mathbf{x} \\ \downarrow \\ \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) \end{array} + \begin{array}{c} \mathbf{y} \\ \downarrow \\ \left(\begin{array}{c} y_1 \\ y_2 \end{array} \right) \end{array} = \left(\begin{array}{c} x_1 + y_1 \\ x_2 + y_2 \end{array} \right)$$

```
a=np.array([1,2,3,4,5])  
b=np.array([2,2,3,3,4])  
print(a+b)
```


ベクトルの演算: 差

- ベクトルの差 (要素同士の差を取る)

$$\begin{array}{c} \mathbf{x} \\ \downarrow \\ \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) \end{array} - \begin{array}{c} \mathbf{y} \\ \downarrow \\ \left(\begin{array}{c} y_1 \\ y_2 \end{array} \right) \end{array} = \left(\begin{array}{c} x_1 - y_1 \\ x_2 - y_2 \end{array} \right)$$

```
a=np.array([1,2,3,4,5])  
b=np.array([2,2,3,3,4])  
print(a-b)
```

ベクトルと数値の積

- ベクトルと数値の積

$$s \times \begin{matrix} \mathbf{x} \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \end{matrix} \equiv \begin{pmatrix} s \times x_1 \\ s \times x_2 \end{pmatrix}$$

```
a=np.array([1,2,3,4,5])
```

```
print(2*a)
```

```
[ 2  4  6  8 10]
```

練習 2 : 2つのベクトル間の和と差を 「プログラムで」計算させよう

$$x = \begin{pmatrix} 3 \\ 5 \end{pmatrix}, y = \begin{pmatrix} 6 \\ 1 \end{pmatrix} \text{ のとき } x + y, x - y \text{ は ?}$$

$$x = \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix}, y = \begin{pmatrix} 6 \\ 1 \\ 2 \end{pmatrix} \text{ のとき } x + y, x - y \text{ は ?}$$

データ「セット」

- 複数のベクトルを行列で表現

```
a=np.array([1,2,3,4,5])  
b=np.array([2,2,3,3,4])  
c=np.array([5,4,2,3,3])
```

```
d=np.array([a,b,c])
```

```
In [474]: d
```

```
Out[474]:
```

```
array([[1, 2, 3, 4, 5],  
       [2, 2, 3, 3, 4],  
       [5, 4, 2, 3, 3]])
```

ユークリッド距離 (2)

- 2次元の場合

x と y の距離の二乗 \equiv

要素の差の二乗
+
要素の差の二乗

- 3次元の場合

x と y の距離の二乗 \equiv

要素の差の二乗
+
要素の差の二乗
+
要素の差の二乗

ユークリッド距離 (3)

- D 次元の場合

x と y の距離の二乗 \equiv

$$\begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \quad \begin{pmatrix} y_1 \\ \vdots \\ y_D \end{pmatrix}$$

要素の差の二乗
+
:
+
要素の差の二乗

というわけで、何次元ベクトルでも距離は計算可能

もちろん1次元ベクトル(数値)間の距離も計算可能

$$(x_1 - y_1)^2$$

演習 1

- 那覇, 福岡, 札幌の気温データを読み込んで
→ `temperature_data.csv`
- 各都市の365日分の気温を一つのベクトル (`numpy array`型) として定義し、ベクトル間(那覇ー福岡、福岡ー札幌、札幌ー那覇)の和と差をそれぞれ求めよう (引き算の順序は問わない) 。
- 上記のベクトル間の距離も求めてみよう。
- 求めた差のベクトルを折れ線グラフやヒストグラムで可視化してみよう。
- ※) `matplotlib.pyplot` を利用する