computer science

pandas for data (science)

- **syntax errors**: raised while the interpreter while executing the program reaches an instruction that does not match syntax rules (the first time the instruction is executed)

- unexpected behavior (exceptions): raised while the interpreter executes the program, due to a situation (data) not taken into account while writing the program

get data from csv file

read_csv(csvfilename):   returns a dataframe from with
the data from the csvfilename

df = pd.read_csv("SBP.csv")

```
   SBP_before  SBP_after
0         125        118
1         132        134
2         138        130
3         120        124
4         125        105
5         127        130
6         136        130
7         139        132
8         131        123
9         132        128
10        135        126
11        136        140
12        128        135
13        127        126
14        130        132
```

```
SBP_before,SBP_after
125,118
132,134
138,130
120,124
125,105
127,130
136,130
139,132
131,123
132,128
135,126
136,140
128,135
127,126
130,132
```

len(df)

returns the number of row in the dataframe

handling missing data

fillna(_value_)

```
[>>> df=df.fillna(0.0)
[>>> df
    SBP_before  SBP_after
0          125      118.0
1          132      134.0
2          138      130.0
3          120      124.0
4          125      105.0
5          127      130.0
6          136      130.0
7          139      132.0
8          131      123.0
9          132      128.0
10         135      126.0
11         136      140.0
12         128      135.0
13         127      126.0
14         130      132.0
15         134        0.0
```

replace NaN with a value

```
    SBP_before  SBP_after
0          125      118.0
1          132      134.0
2          138      130.0
3          120      124.0
4          125      105.0
5          127      130.0
6          136      130.0
7          139      132.0
8          131      123.0
9          132      128.0
10         135      126.0
11         136      140.0
12         128      135.0
13         127      126.0
14         130      132.0
15         134        NaN
```

remove when incomplete

dropna()

```
[>>> df = df.dropna()
[>>> df
    SBP_before  SBP_after
0          125      118.0
1          132      134.0
2          138      130.0
3          120      124.0
4          125      105.0
5          127      130.0
6          136      130.0
7          139      132.0
8          131      123.0
9          132      128.0
10         135      126.0
11         136      140.0
12         128      135.0
13         127      126.0
14         130      132.0
```

selecting rows (subset observation)

- by position
  ```
  df.iloc[fromrow:torow]
  ```

```
[>>> dfw = pd.read_csv("./WSRTdf.csv")
[>>> dfw
     n type  alpha  value
0    5  TTT   0.05    NaN
1    5  TTT   0.01    NaN
2    5  OTT   0.05    0.0
3    5  OTT   0.01    NaN
4    6  TTT   0.05    0.0
..  ..  ...    ...    ...
99  29  OTT   0.01  110.0
100 30  TTT   0.05  137.0
101 30  TTT   0.01  109.0
102 30  OTT   0.05  151.0
103 30  OTT   0.01  120.0

[104 rows x 4 columns]
```

```
[>>> dfw.iloc[1:5]
   n type  alpha  value
1  5  TTT   0.01    NaN
2  5  OTT   0.05    0.0
3  5  OTT   0.01    NaN
4  6  TTT   0.05    0.0
```
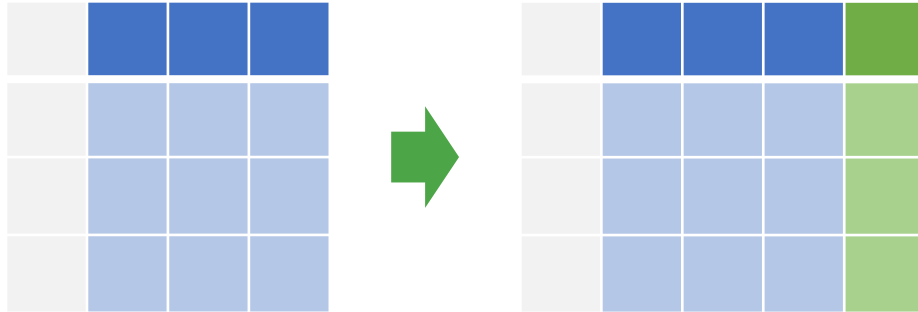
selecting rows (subset observation) 2

- by value in one …

… or more columns

creating new data

df['NewColumnName'] = df.ExistingColumn *operator* variable


```
[>>> df['Difference'] = df.SBP_before - df.SBP_after
[>>> df
    SBP_before    SBP_after    Difference
0          125          118             7
1          132          134            -2
2          138          130             8
3          120          124            -4
4          125          105            20
5          127          130            -3
6          136          130             6
7          139          132             7
8          131          123             8
9          132          128             4
10         135          126             9
11         136          140            -4
12         128          135            -7
13         127          126             1
14         130          132            -2
```

df['Difference'] = df.SBP_before - df.SBP_after

creating new data 2

```
df['NewColumnName'] = df.ExistingColumn operator variable
```

```
df['AbsDiff'] = abs(df.Difference)
```

```
[>>> df['AbsDiff'] = abs(df.Difference)
[>>> df
    SBP_before    SBP_after    Difference    AbsDiff
0          125          118             7          7
1          132          134            -2          2
2          138          130             8          8
3          120          124            -4          4
4          125          105            20         20
5          127          130            -3          3
6          136          130             6          6
7          139          132             7          7
8          131          123             8          8
9          132          128             4          4
10         135          126             9          9
11         136          140            -4          4
12         128          135            -7          7
13         127          126             1          1
14         130          132            -2          2
```

creating new data, selectively

sorting

list *of columns*

⬇

`sort_values(``by,ascending``):` sorts with respect to values in columns, in ascending (default) or descending order

it is not "in place"

`df`` = df.sort_values(``"AbsDiff"``)`

```
df = df.sort_values("AbsDiff")
```

```
[>>> df = df.sort_values("AbsDiff")
[>>> df
     SBP_before   SBP_after   Difference   AbsDiff
13        127         126            1         1
1         132         134           -2         2
14        130         132           -2         2
5         127         130           -3         3
3         120         124           -4         4
9         132         128            4         4
11        136         140           -4         4
6         136         130            6         6
0         125         118            7         7
7         139         132            7         7
12        128         135           -7         7
2         138         130            8         8
8         131         123            8         8
10        135         126            9         9
4         125         105           20        20
```

```
[>>> df = df.sort_values("Difference")
[>>> df
     SBP_before   SBP_after   Difference   AbsDiff
12        128         135           -7         7
3         120         124           -4         4
11        136         140           -4         4
5         127         130           -3         3
1         132         134           -2         2
14        130         132           -2         2
13        127         126            1         1
9         132         128            4         4
6         136         130            6         6
0         125         118            7         7
7         139         132            7         7
2         138         130            8         8
8         131         123            8         8
10        135         126            9         9
4         125         105           20        20
```

```
df = df.sort_values("Difference")
```

sorting 3

sort by this **then by this**

```
df = df.sort_values(["AbsDiff", "Difference")
```

```
[>>> df = df.sort_values(["AbsDiff", "Difference"])
[>>> df
     SBP_before   SBP_after   Difference   AbsDiff
13         127         126            1         1
1          132         134           -2         2
14         130         132           -2         2
5          127         130           -3         3
3          120         124           -4         4
11         136         140           -4         4
9          132         128            4         4
6          136         130            6         6
12         128         135           -7         7
0          125         118            7         7
7          139         132            7         7
2          138         130            8         8
8          131         123            8         8
10         135         126            9         9
4          125         105           20        20
```

ranking data

```
rank(method="…"):
```

returns a rank of every respective index of a series passed.
The rank is returned on the basis of position after sorting

```
df["AbsDiff"].rank()
```

ranking data 2

extra column

df["AbsDiff"].rank() ➡ df["Ranks"] = df["AbsDiff"].rank()

with sorting

	SBP_before	SBP_after	Difference	AbsDiff	Ranks
13	127	126	1	1	1.0
1	132	134	-2	2	2.5
14	130	132	-2	2	2.5
5	127	130	-3	3	4.0
3	120	124	-4	4	6.0
11	136	140	-4	4	6.0
9	132	128	4	4	6.0
6	136	130	6	6	8.0
12	128	135	-7	7	10.0
0	125	118	7	7	10.0
7	139	132	7	7	10.0
2	138	130	8	8	12.5
8	131	123	8	8	12.5
10	135	126	9	9	14.0
4	125	105	20	20	15.0

	SBP_before	SBP_after	Difference	AbsDiff	Ranks
0	125	118	7	7	10.0
1	132	134	-2	2	2.5
2	138	130	8	8	12.5
3	120	124	-4	4	6.0
4	125	105	20	20	15.0
5	127	130	-3	3	4.0
6	136	130	6	6	8.0
7	139	132	7	7	10.0
8	131	123	8	8	12.5
9	132	128	4	4	6.0
10	135	126	9	9	14.0
11	136	140	-4	4	6.0
12	128	135	-7	7	10.0
13	127	126	1	1	1.0
14	130	132	-2	2	2.5

without sorting

creating new data, selectively

```
df['R+'] = df['Ranks']*(df['Difference'] > 0)
```

```
[>>> df['R+'] = df['Ranks']*(df['Difference'] > 0)
[>>> df
    SBP_before  SBP_after  Difference  AbsDiff  Ranks   R+
13        127        126           1        1    1.0   1.0
1         132        134          -2        2    2.5   0.0
14        130        132          -2        2    2.5   0.0
5         127        130          -3        3    4.0   0.0
3         120        124          -4        4    6.0   0.0
11        136        140          -4        4    6.0   0.0
9         132        128           4        4    6.0   6.0
6         136        130           6        6    8.0   8.0
12        128        135          -7        7   10.0   0.0
0         125        118           7        7   10.0  10.0
7         139        132           7        7   10.0  10.0
2         138        130           8        8   12.5  12.5
8         131        123           8        8   12.5  12.5
10        135        126           9        9   14.0  14.0
4         125        105          20       20   15.0  15.0
```

1 when true
0 when false

dataframes

df[col]   select column

df.iloc[rowlabel]   select row by label