BI 코딩 실무 II

Bioinformatics Pandas Lecture

한주현

11/27/2020

kenneth.jh.han@snu.ac.kr

금일 강의

- bar plot 그리기
  - VCF 파일 InDel의 개수
- box plot 그리기
  - Gene Expression data, gene 의 차이를 보기
- scatter plot 그리기
  - Gene Expression과 Age 간의 상관관계 알아보기

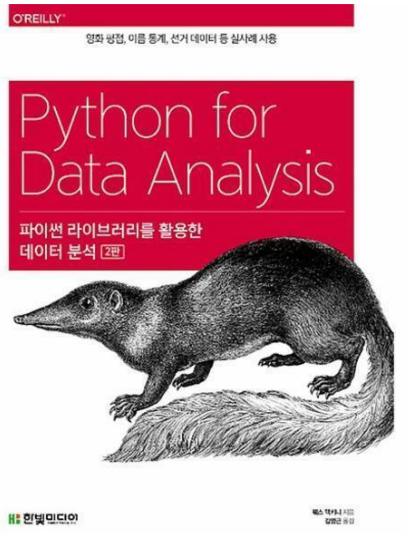
준비물

- Jupyter notebook
- Pandas

Pandas

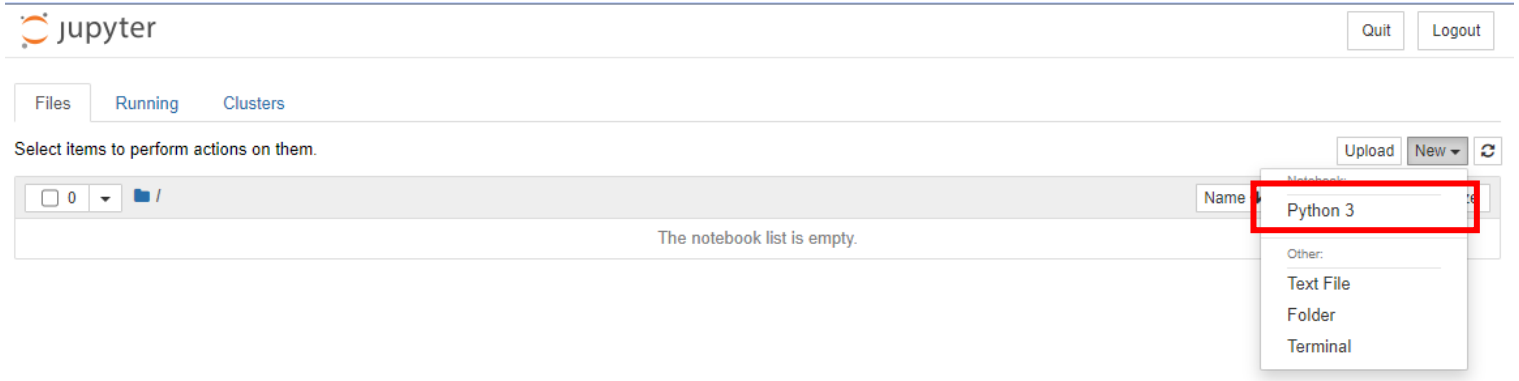
- Wes McKinney 가 만든 데이터 분석 소프트웨어 라이브러리 (2008).
- R의 data.frame 구조와 같은 DataFrame 이라는 자료구조를 만듬.
- 파이썬 사용자들이 편하게 테이블 형태의 데이터를 다룰 수 있도록 만듬.

Wes McKinney
1985.03.20 -



Warming Up

jupyter notebook 을 실행

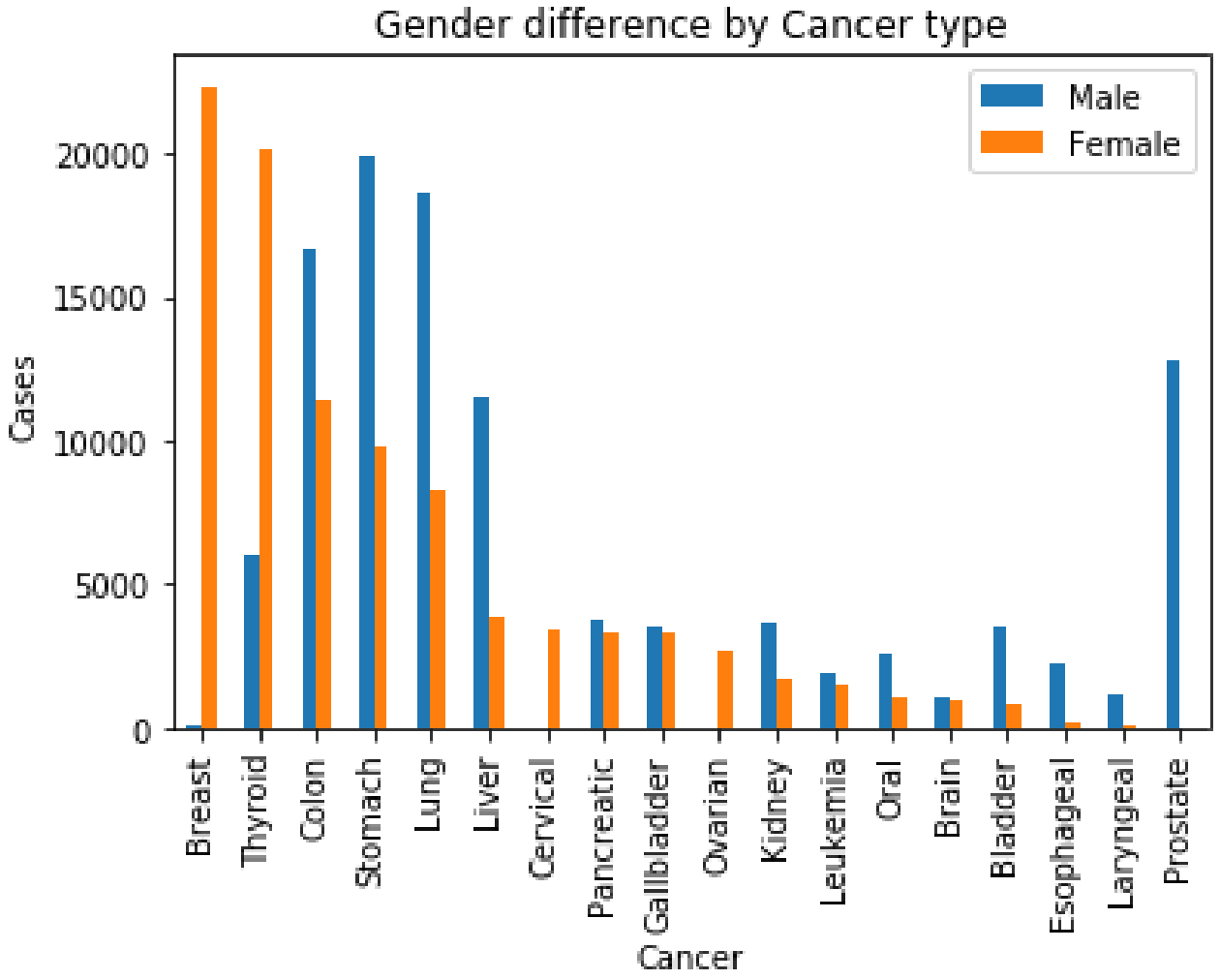


Python3 로 노트북 실행
다음을 실행하여 pandas 가 잘 로딩 되는지 확인

```
import pandas as pd
print(pd.__version__)
```

```
1.0.1
```

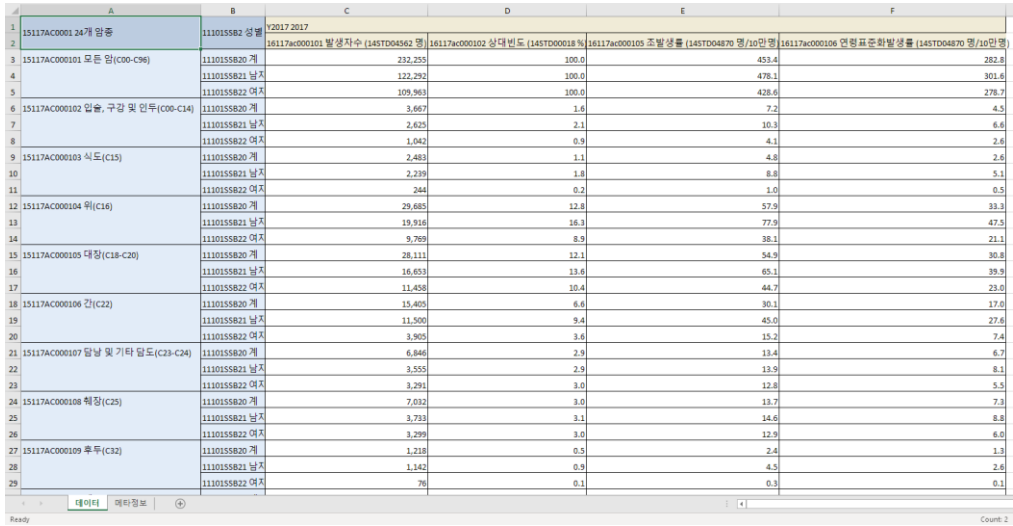
Pandas 로 bar plot 그리기



- 다음과 같은 plot 을 그려봅시다.

데이터 다운로드

- https://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_117N_A00022



	A	B	C	D	E	F
1	15117AC0001 24개 암종	11101SSB2 성별	Y2017 2017			
2			16117ac000101 발생자수 (145TD04562 명)	16117ac000102 상대빈도 (145TD00018 %)	16117ac000105 조발생률 (145TD04870 명/10만명)	16117ac000106 연령표준화발생률 (145TD04870 명/10만명)
3	15117AC000101 모든 암(C00-C96)	11101SSB20 계	232,255	100.0	453.4	282.8
4		11101SSB21 남자	122,292	100.0	478.1	301.6
5		11101SSB22 여자	109,963	100.0	428.6	278.7
6	15117AC000102 입술, 구강 및 인두(C00-C14)	11101SSB20 계	3,667	1.6	7.2	4.5
7		11101SSB21 남자	2,625	2.1	10.3	6.6
8		11101SSB22 여자	1,042	0.9	4.1	2.6
9	15117AC000103 식도(C15)	11101SSB20 계	2,483	1.1	4.8	2.6
10		11101SSB21 남자	2,239	1.8	8.8	5.1
11		11101SSB22 여자	244	0.2	1.0	0.5
12	15117AC000104 위(C16)	11101SSB20 계	29,685	12.8	57.9	33.3
13		11101SSB21 남자	19,916	16.3	77.9	47.5
14		11101SSB22 여자	9,769	8.9	38.1	21.1
15	15117AC000105 대장(C18-C20)	11101SSB20 계	28,111	12.1	54.9	30.8
16		11101SSB21 남자	16,653	13.6	65.1	39.9
17		11101SSB22 여자	11,458	10.4	44.7	23.0
18	15117AC000106 간(C22)	11101SSB20 계	15,405	6.6	30.1	17.0
19		11101SSB21 남자	11,500	9.4	45.0	27.6
20		11101SSB22 여자	3,905	3.6	15.2	7.4
21	15117AC000107 담낭 및 기타 담도(C23-C24)	11101SSB20 계	6,846	2.9	13.4	6.7
22		11101SSB21 남자	3,555	2.9	13.9	8.1
23		11101SSB22 여자	3,291	3.0	12.8	5.5
24	15117AC000108 췌장(C25)	11101SSB20 계	7,032	3.0	13.7	7.3
25		11101SSB21 남자	3,733	3.1	14.6	8.8
26		11101SSB22 여자	3,299	3.0	12.9	6.0
27	15117AC000109 후두(C32)	11101SSB20 계	1,218	0.5	2.4	1.3
28		11101SSB21 남자	1,142	0.9	4.5	2.6
29		11101SSB22 여자	76	0.1	0.3	0.1

kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_117N_A00022

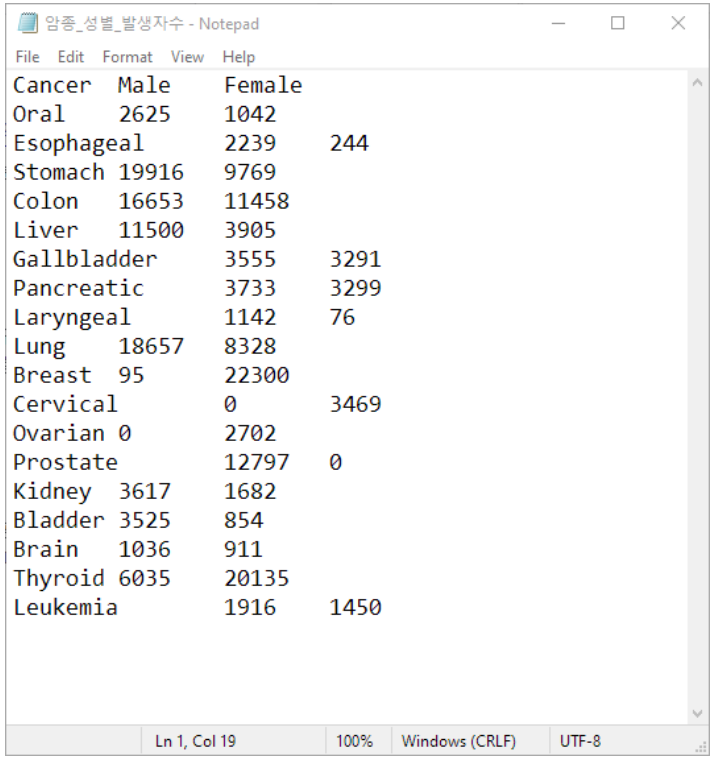
KOSIS

1) 24개 암종/성별 암발생자수, 상대빈도, 조발생률, 연령표준화발생률

자료갱신일: 2020-01-29 / 수록기간: 년 1999 ~ 2017 / 자료문의처 : 044-202-2513

일괄설정 | 항목[4/4] | 성별[3/3] | 24개 암종[25/25] | 시점[1/19]

24개 암종	성별	2017			
		발생자수 (명)	상대빈도 (%)	조발생률 (명/10만명)	연령표준화발생률 (명/10만명)
모든 암(C00-C96)	계	232,255	100.0	453.4	282.8
	남자	122,292	100.0	478.1	301.6
	여자	109,963	100.0	428.6	278.7
입술, 구강 및 인두(C00-C14)	계	3,667	1.6	7.2	4.5
	남자	2,625	2.1	10.3	6.6
	여자	1,042	0.9	4.1	2.6
식도(C15)	계	2,483	1.1	4.8	2.6
	남자	2,239	1.8	8.8	5.1
	여자	244	0.2	1.0	0.5
위(C16)	계	29,685	12.8	57.9	33.3
	남자	19,916	16.3	77.9	47.5
	여자	9,769	8.9	38.1	21.1
대장(C18-C20)	계	28,111	12.1	54.9	30.8
	남자	16,653	13.6	65.1	39.9
	여자	11,458	10.4	44.7	23.0
간(C22)	계	15,405	6.6	30.1	17.0
	남자	11,500	9.4	45.0	27.6
	여자	3,905	3.6	15.2	7.4
담낭 및 기타 담도(C23-C24)	계	6,846	2.9	13.4	6.7
	남자	3,555	2.9	13.9	8.1
	여자	3,291	3.0	12.8	5.5
췌장(C25)	계	7,032	3.0	13.7	7.3
	남자	3,733	3.1	14.6	8.8
	여자	3,299	3.0	12.9	6.0
후두(C32)	계	1,218	0.5	2.4	1.3



암종_성별_발생자수 - Notepad

File Edit Format View Help

```
Cancer  Male    Female
Oral    2625    1042
Esophageal      2239    244
Stomach 19916   9769
Colon   16653   11458
Liver   11500   3905
Gallbladder     3555    3291
Pancreatic      3733    3299
Laryngeal       1142    76
Lung    18657   8328
Breast  95      22300
Cervical        0       3469
Ovarian 0       2702
Prostate        12797   0
Kidney  3617    1682
Bladder 3525    854
Brain   1036    911
Thyroid 6035    20135
Leukemia        1916    1450
```

Ln 1, Col 19 | 100% | Windows (CRLF) | UTF-8

csv, tsv 파일 읽기

```
import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t")
df
```

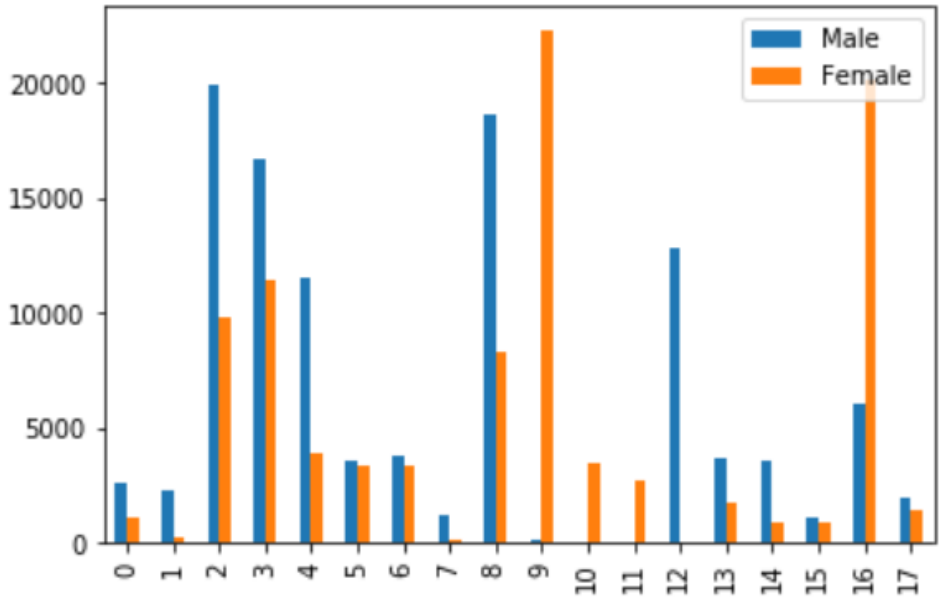
- pandas.read_csv() 메서드를 사용하여 파일을 읽습니다.
- sep 으로 separator 를 지정할 수 있습니다.
  기본적으로는 comma (,) 이나 “\t” 을 하게 되면 탭으로 나눌 수 있습니다.

	Cancer	Male	Female
0	Oral	2625	1042
1	Esophageal	2239	244
2	Stomach	19916	9769
3	Colon	16653	11458
4	Liver	11500	3905
5	Gallbladder	3555	3291
6	Pancreatic	3733	3299
7	Laryngeal	1142	76
8	Lung	18657	8328
9	Breast	95	22300
10	Cervical	0	3469
11	Ovarian	0	2702
12	Prostate	12797	0
13	Kidney	3617	1682
14	Bladder	3525	854
15	Brain	1036	911
16	Thyroid	6035	20135
17	Leukemia	1916	1450

bar plot 그리기

```
%matplotlib inline

import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t")
df.plot(kind="bar")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc873f80090>
```



%matplotlib inline

→ 페이지에서 바로 그림 확인

df.plot(kind="bar")

→ DataFrame 에서 barplot 을 그릴 수 있게 해줌

그런데.. x 축이 좀 이상하다..

bar plot 그리기

```
import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t")
df
```

index 라고 함

	Cancer	Male	Female
0	Oral	2625	1042
1	Esophageal	2239	244
2	Stomach	19916	9769
3	Colon	16653	11458
4	Liver	11500	3905
5	Gallbladder	3555	3291
6	Pancreatic	3733	3299
7	Laryngeal	1142	76
8	Lung	18657	8328
9	Breast	95	22300
10	Cervical	0	3469
11	Ovarian	0	2702
12	Prostate	12797	0
13	Kidney	3617	1682
14	Bladder	3525	854
15	Brain	1036	911
16	Thyroid	6035	20135
17	Leukemia	1916	1450

```
import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t", index_col=0)
df
```

Cancer	Male	Female
Oral	2625	1042
Esophageal	2239	244
Stomach	19916	9769
Colon	16653	11458
Liver	11500	3905
Gallbladder	3555	3291
Pancreatic	3733	3299
Laryngeal	1142	76
Lung	18657	8328
Breast	95	22300
Cervical	0	3469
Ovarian	0	2702
Prostate	12797	0
Kidney	3617	1682
Bladder	3525	854
Brain	1036	911
Thyroid	6035	20135
Leukemia	1916	1450

index_col = 0 으로

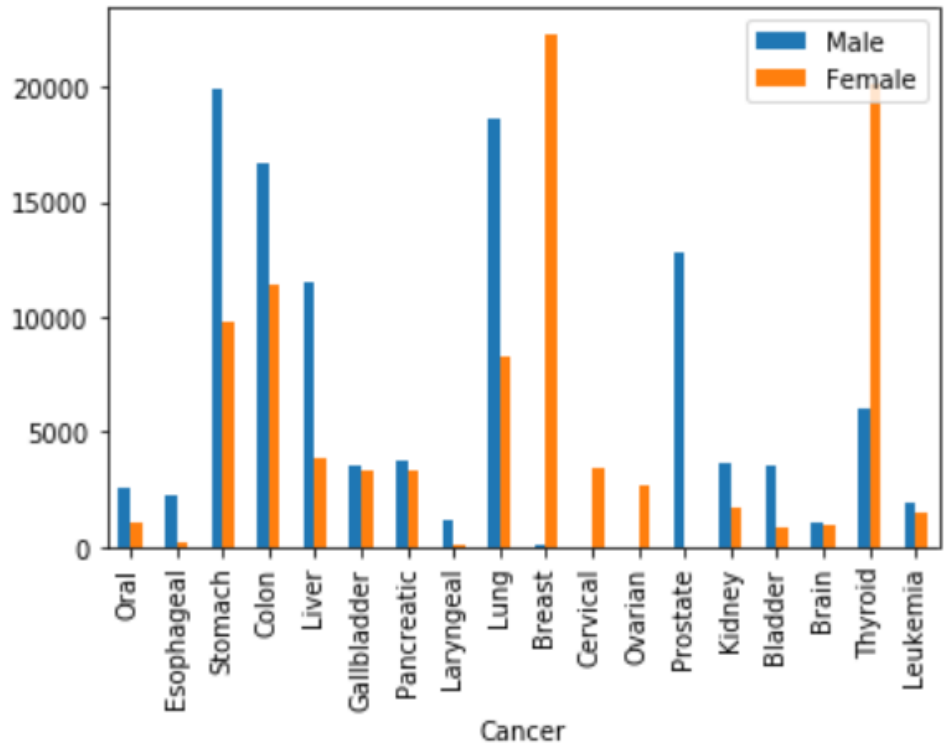
0 번째 컬럼을 index로 지정했다.

이제 다시 plot 을 그려보자.

bar plot 그리기

```
%matplotlib inline

import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t", index_col=0)
df.plot(kind="bar")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc878578510>
```

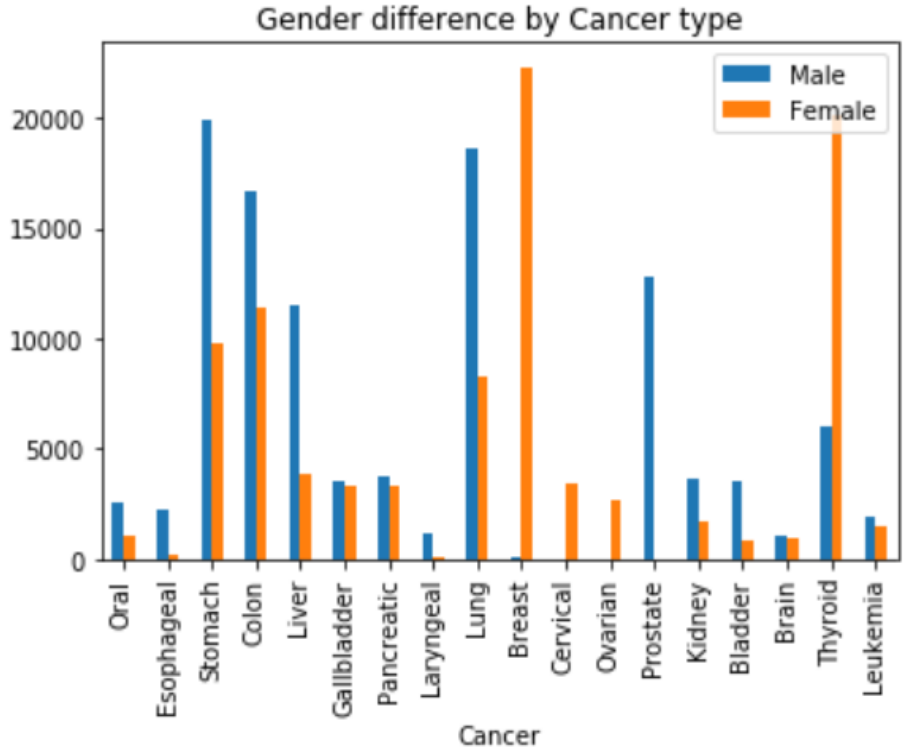


- 이제 각 암종별로 Male/Female 의 발생 건수에 대한 bar plot 을 그렸다.
- plot 에 제목이 있었으면 좋겠다..

bar plot 그리기

```
%matplotlib inline

import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t", index_col=0)
df.plot(kind="bar", title="Gender difference by Cancer type")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc87287c090>
```



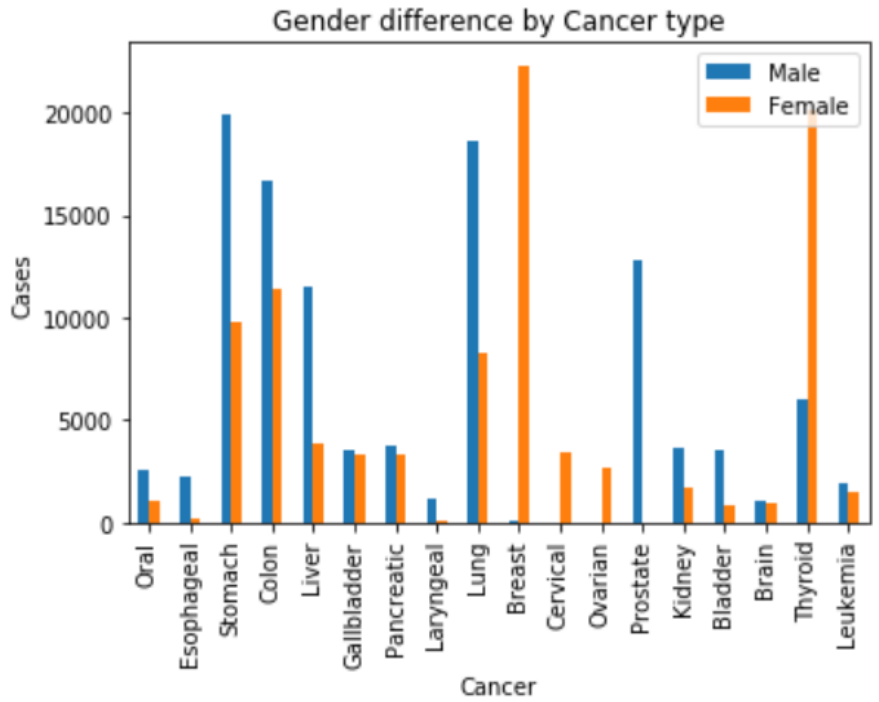
- df.plot() 메서드에서 title="" 로 값을 지정하면 plot 에서 제목을 넣을 수 있다.

- x 축에는 라벨이 있는데, y 축에는 라벨이 없네?

bar plot 그리기

```
%matplotlib inline

import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t", index_col=0)
ax = df.plot(kind="bar")
ax.set_title("Gender difference by Cancer type")
ax.set_ylabel("Cases")
```

```
Text(0, 0.5, 'Cases')
```



- df.plot() 메서드의 반환값을 ax 로 받고,

ax.set_title() 로 타이틀을 지정

ax.set_ylabel() 로 y축 라벨을 지정 할 수 있다.

- 정렬을 했으면 좋겠다.

bar plot 그리기

```python
import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t", index_col=0)
df = df.sort_values(by=['Male'])
df
```

```python
import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t", index_col=0)
df = df.sort_values(by=['Male'], ascending=False)
df
```

Cancer	Male	Female
Ovarian	0	2702
Cervical	0	3469
Breast	95	22300
Brain	1036	911
Laryngeal	1142	76
Leukemia	1916	1450
Esophageal	2239	244
Oral	2625	1042
Bladder	3525	854
Gallbladder	3555	3291
Kidney	3617	1682
Pancreatic	3733	3299
Thyroid	6035	20135
Liver	11500	3905
Prostate	12797	0
Colon	16653	11458
Lung	18657	8328
Stomach	19916	9769

Cancer	Male	Female
Stomach	19916	9769
Lung	18657	8328
Colon	16653	11458
Prostate	12797	0
Liver	11500	3905
Thyroid	6035	20135
Pancreatic	3733	3299
Kidney	3617	1682
Gallbladder	3555	3291
Bladder	3525	854
Oral	2625	1042
Esophageal	2239	244
Leukemia	1916	1450
Laryngeal	1142	76
Brain	1036	911
Breast	95	22300
Ovarian	0	2702
Cervical	0	3469

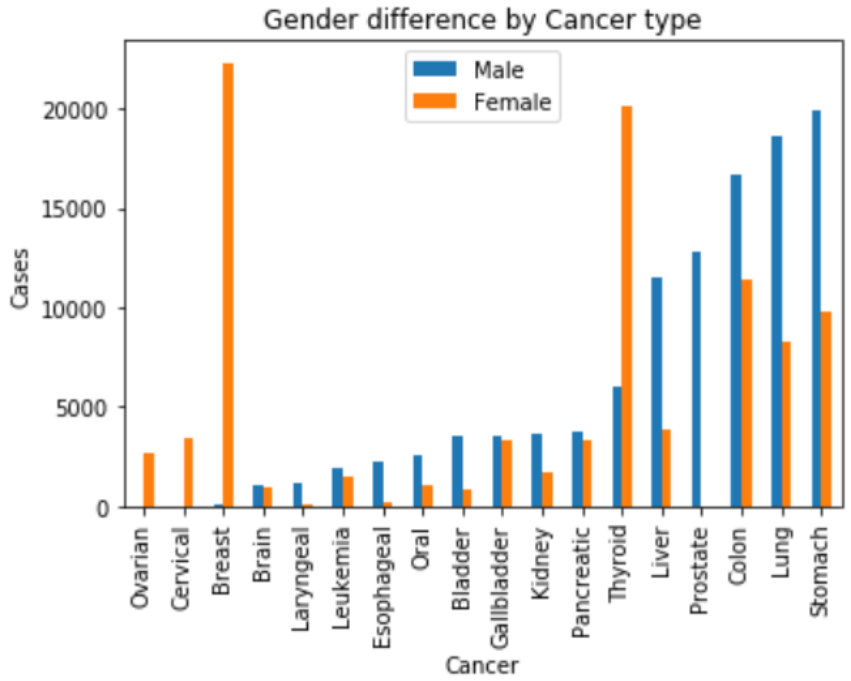
df.sort_values() 메서드로 정렬할 수 있다.

ascending 은 기본적으로 True 로 되어있어서, False 로 지정하면 내림차순이 된다.

bar plot 그리기

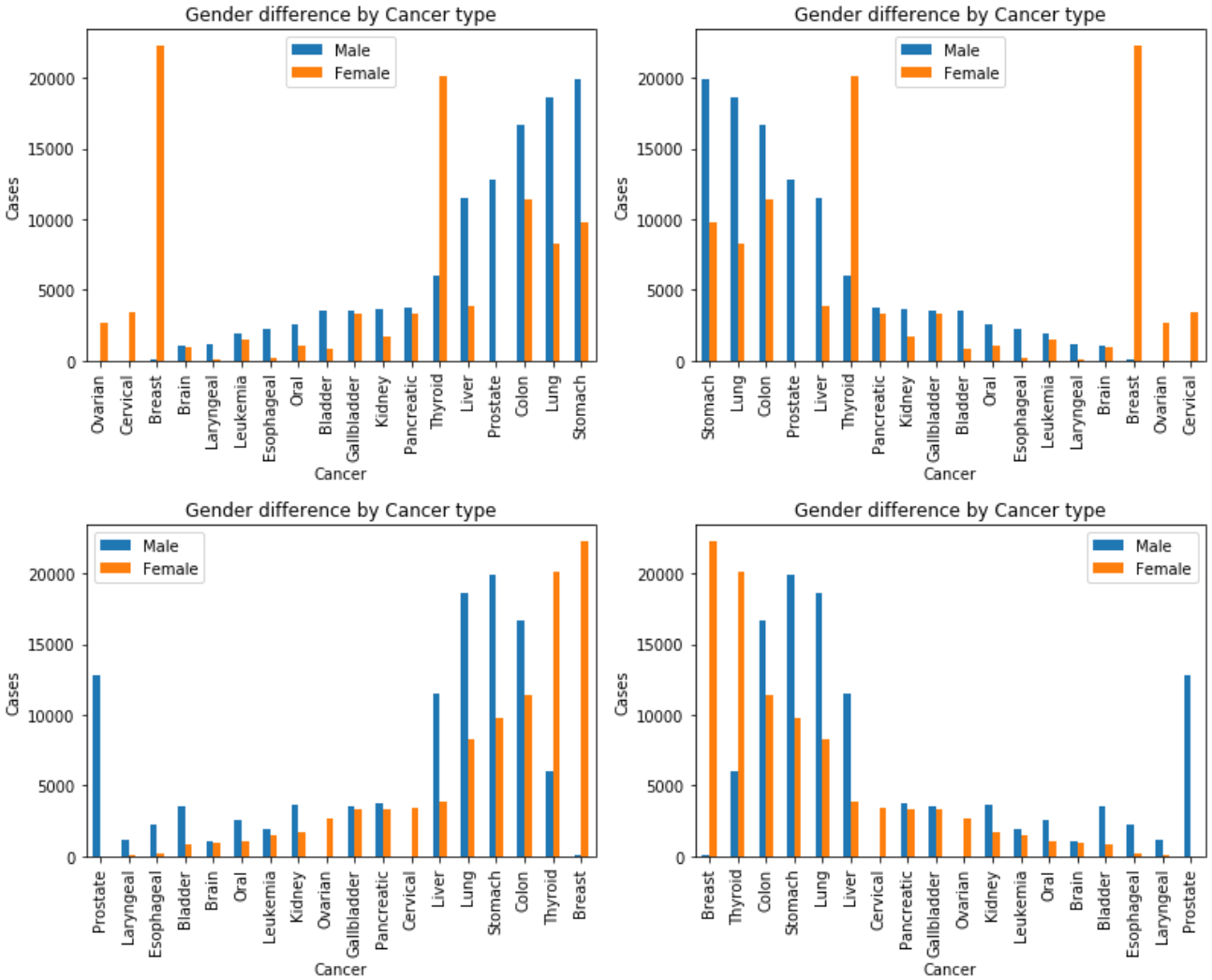
```
%matplotlib inline

import pandas as pd

in_file = "암종_성별_발생자수.txt"
df = pd.read_csv(in_file, sep="\t", index_col=0)
df = df.sort_values(by=['Male'])
ax = df.plot(kind="bar")
ax.set_title("Gender difference by Cancer type")
ax.set_ylabel("Cases")
```

```
Text(0, 0.5, 'Cases')
```



- Male 로 증가하는 plot 을 그려보았다.

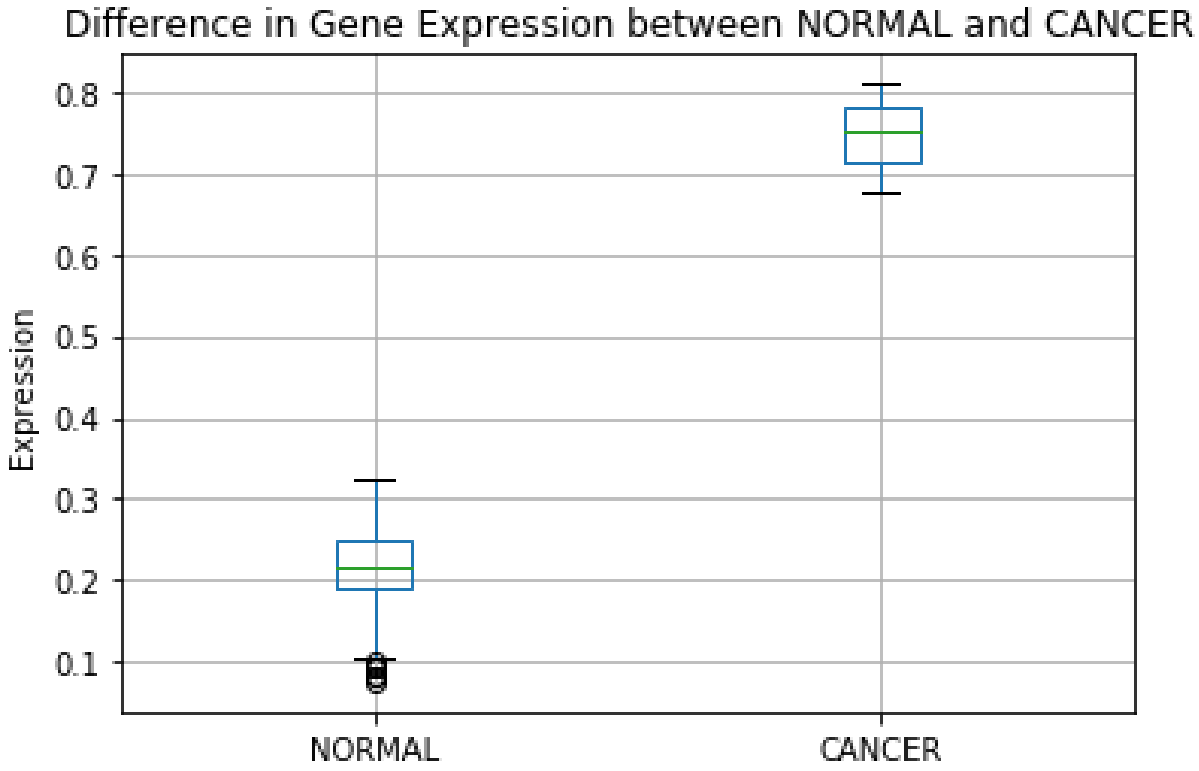
연습문제



- 다음과 같이 성별로 오름차순, 내림차순으로 barplot 을 그려보세요.

- 성별 별로 수가 가장 많고 적은 암은 무엇인가요?

Pandas 로 box plot 그리기



- 다음과 같은 plot 을 그려봅시다.

box plot 그리기

```
import pandas as pd

df_expr = pd.read_csv("cancer_expression.txt",sep="\t")
df_expr
```

	NORMAL	CANCER
0	0.074617	0.678911
1	0.082337	0.679004
2	0.083483	0.679280
3	0.091111	0.680160
4	0.094738	0.680228
...	...	...
423	0.312627	0.809795
424	0.316885	0.809860
425	0.319162	0.810271
426	0.323173	0.810312
427	0.323960	0.810514

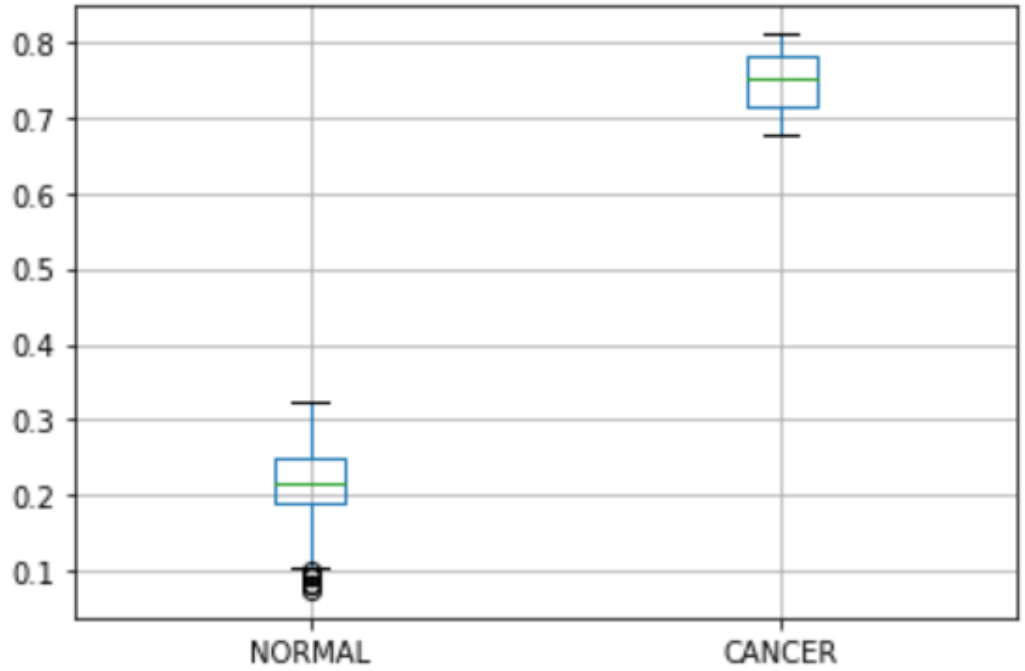
428 rows × 2 columns

- 다음과 같이 데이터를 불러온다.

box plot 그리기

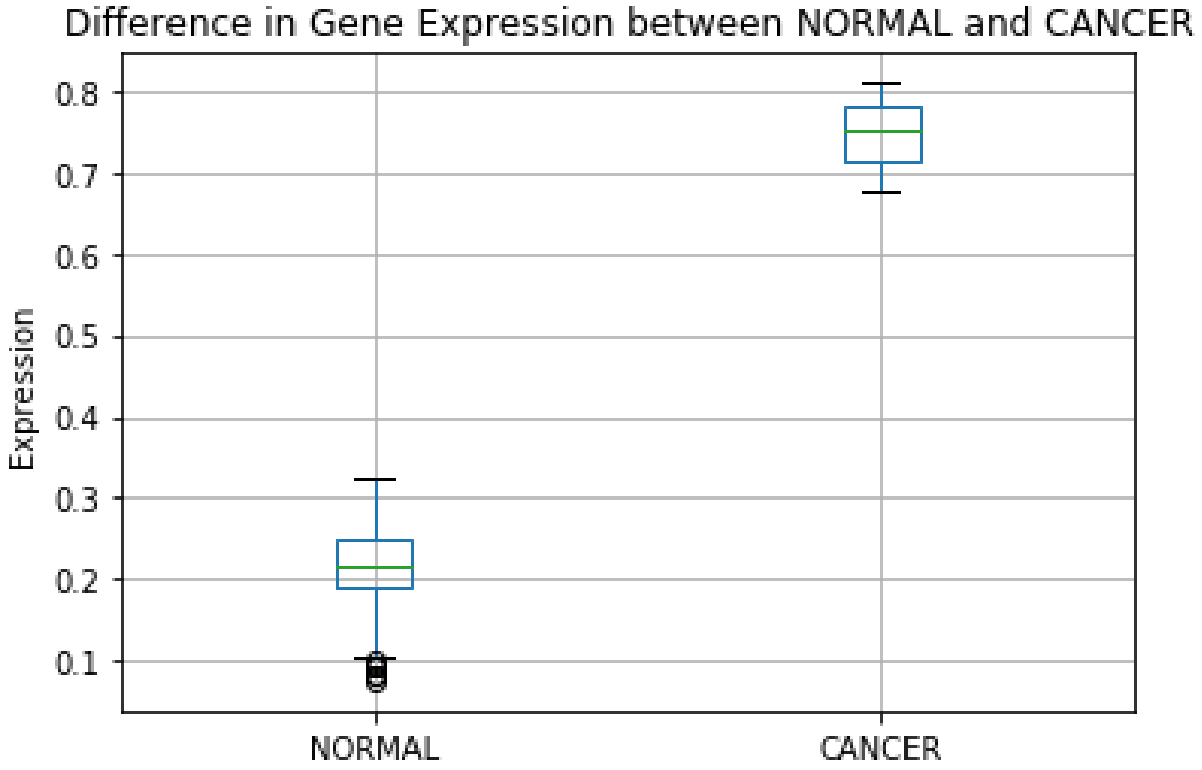
```
%matplotlib inline

import pandas as pd

df_expr = pd.read_csv("cancer_expression.txt",sep="\t")
df_expr.boxplot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc7f55f1dd0>
```



- df 의 boxplot() 메서드로 boxplot 을 그릴 수 있다.

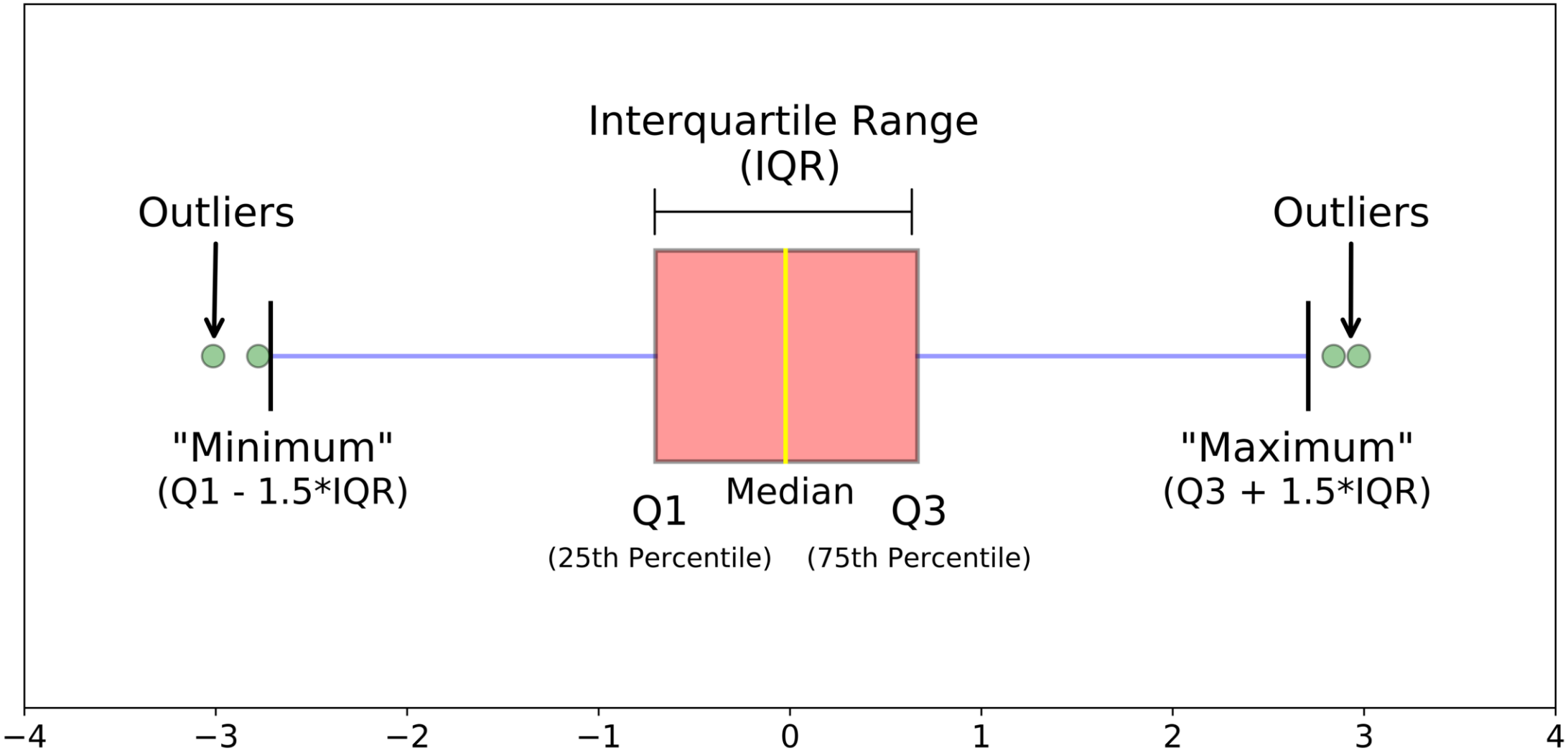
box plot 그리기



- 왼쪽 그림과 같이
제목과 y 축 label 을 달아보자.

hint: 앞서 배운 barplot 을
참고한다.

box plot 이론


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

Pandas 로 Scatter plot 그리기

- 다음과 같은 plot 을 그려봅시다.

scatter plot 그리기

```
import pandas as pd

df_age_expr = pd.read_csv("Age_Expression.txt",sep="\t")
df_age_expr
```

	Age	Expression
0	74	0.228342
1	65	2.293392
2	55	4.330165
3	57	4.295806
4	33	3.092767
...	...	...
500	35	1.931817
501	20	3.695942
502	41	1.384062
503	74	1.932777
504	50	4.001779

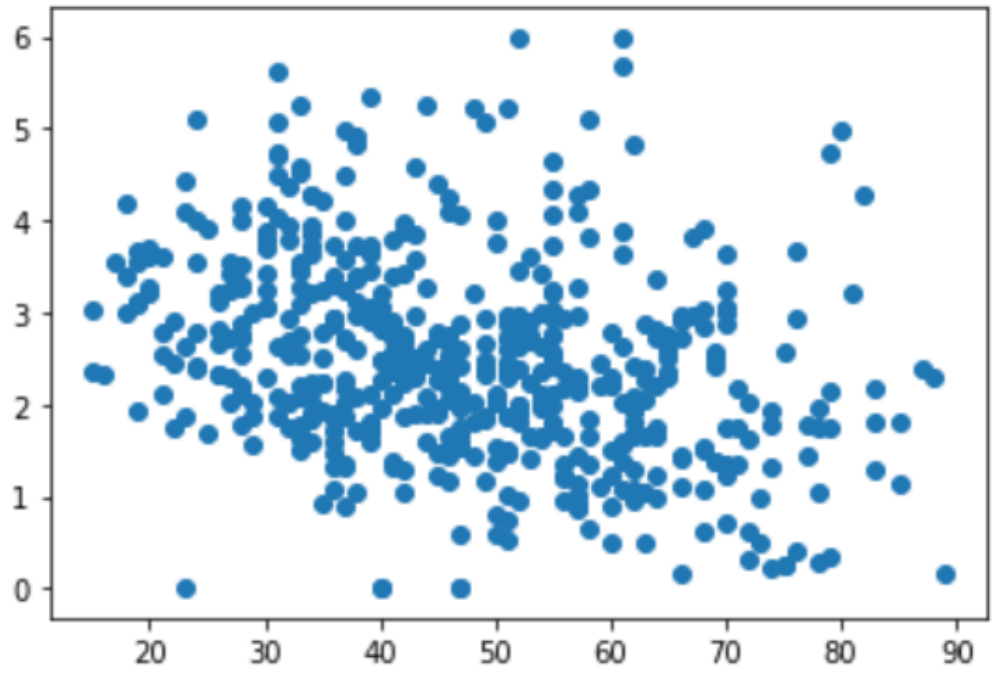
505 rows × 2 columns

- 다음과 같이 데이터를 불러온다.

scatter plot 그리기

```
import matplotlib.pyplot as plt
import numpy as np

x = df_age_expr["Age"]
y = df_age_expr["Expression"]

plt.scatter(x, y)
```

```
<matplotlib.collections.PathCollection at 0x7fc85d127450>
```



- plt.scatter() 메서드로 scatter plot 을 그립니다.
- x 축 값과 y 축 값을 각각 DataFrame에서 가져옵니다.

scatter plot 그리기

```
import matplotlib.pyplot as plt
import numpy as np

x = df_age_expr["Age"]
y = df_age_expr["Expression"]

plt.scatter(x, y)

plt.title("Age to Gene Expression")
plt.ylabel("Gene Expression")
plt.xlabel("Age")
```

```
Text(0.5, 0, 'Age')
```

- plt.title() 메서드로 plot 의 제목을 설정합니다.
- plt.xlabel(), plt.ylabel() 메서드로 x축과 y축의 라벨을 설정합니다.

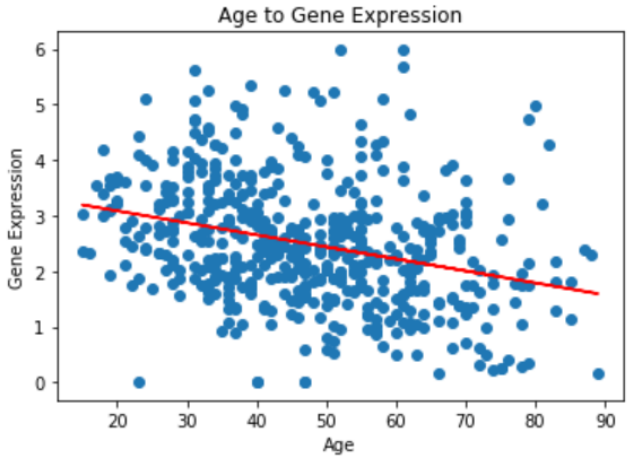
scatter plot 그리기

- Gene expression은 Age 에 따라 값이 감소하는 것일까?
- 이를 검정하려면 어떻게 해야 하는가?

scatter plot 그리기

```
import matplotlib.pyplot as plt
import scipy

x = df_age_expr["Age"]
y = df_age_expr["Expression"]

plt.scatter(x, y)
plt.title("Age to Gene Expression")
plt.ylabel("Gene Expression")
plt.xlabel("Age")

slope, intercept, r_value, p_value, std_err = scipy.stats.linregress(x, y)

# print(slope, intercept, r_value, p_value, std_err)

m, b = np.polyfit(x, y, 1)
plt.plot(x, slope*x + intercept, "r")
print(f"y = {slope}x + {intercept}")
print(f"r2 = {r_value}")
print(f"pval = {p_value}")
```

```
y = -0.021596438821670914x + 3.5194134458502697
r2 = -0.30721710397345064
pval = 1.6874777354402026e-12
```



- scipy.stats.linregress() 메서드로

slope: 기울기

intercept: 절편

r_value: 상관계수 (correlation coefficient)

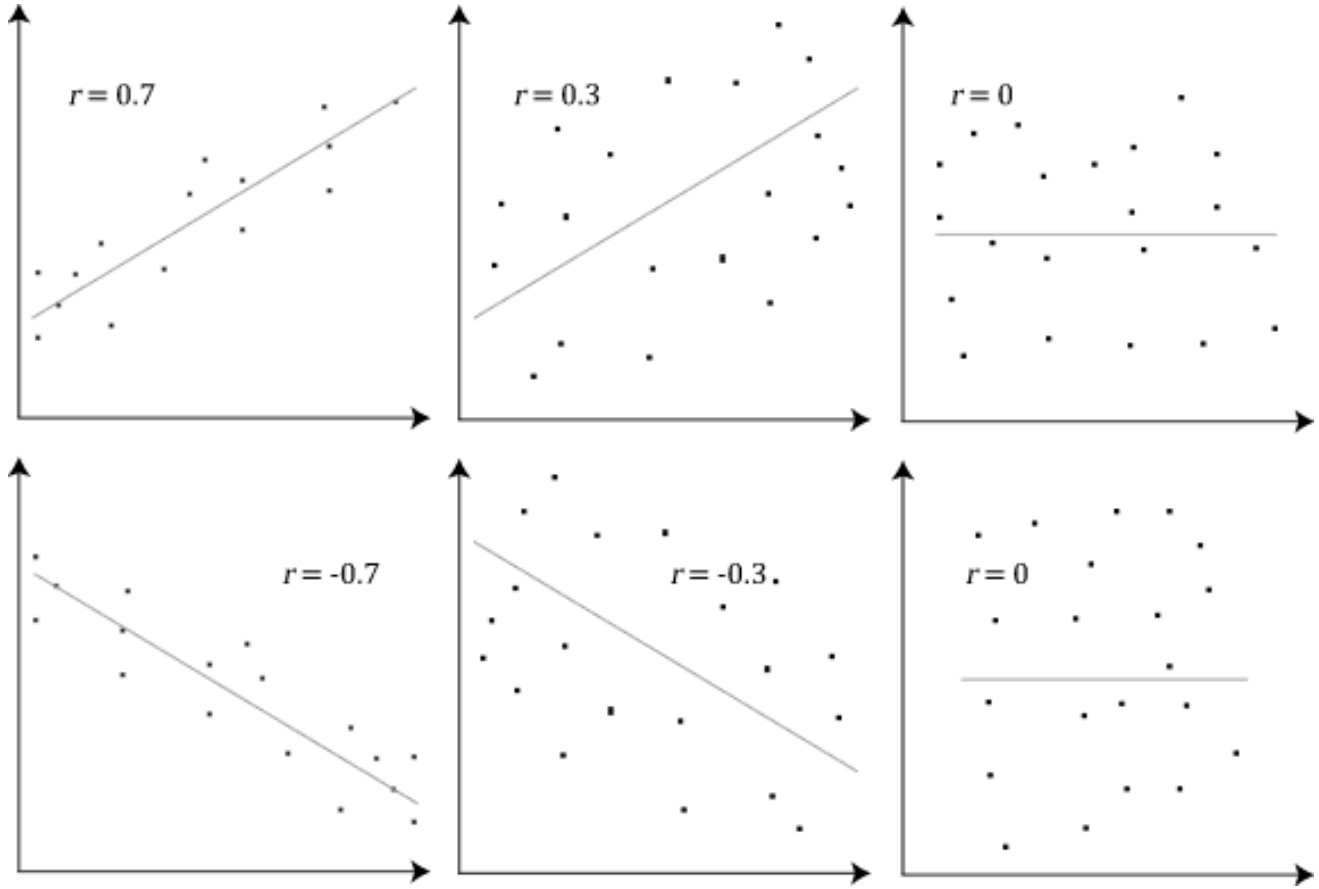
p_value:

H0: 기울기가 0이다.

Ha: 기울기가 0이 아니다.

std_err: 오차

상관계수 correlation coefficient 에 대하여



$$r_{XY} = \frac{\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\sum_i^n \left(Y_i - \overline{Y}\right)^2}}$$

과제

- gender_recurrence.txt 파일로 부터 M, F 에 따른 recurrence 를 bar plot 으로 그려보세요.
- box_exercise.txt 파일로 부터 boxplot 을 그려보세요.
- age_expression_exercise.txt 파일로 부터 scatter plot 을 그려보세요. scatter plot 은 어떤 상관관계를 보이나요?