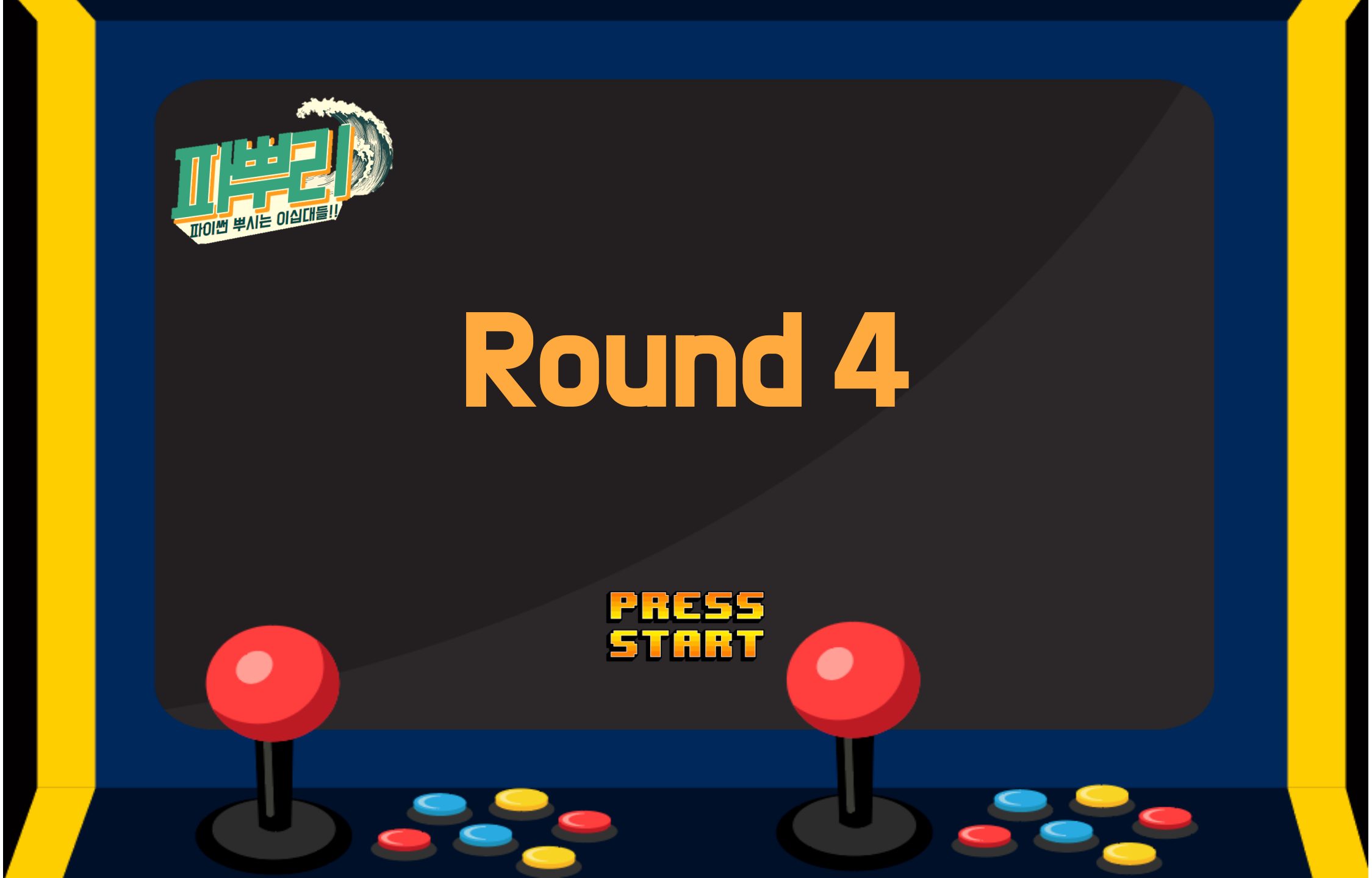

파뿌리
파이썬 뿌시는 이십대들!!
Round 4
PRESS
START

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 4 》

- 데이터 시각화 개요
- 시각화 실습
- 이상치(outliar)

New Assignment

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 4 》

- 데이터 시각화 개요 《
- 시각화 실습
- 이상치(outliar)

파뿌리
파이썬 뿌시는 이십대들!!

데이터 시각화를 하는 이유?

why ?

파뿌리
파이썬 뿌시는 이십대들!!

데이터 시각화를 하는 이유?

why ?

파뿌리
파이썬 뿌시는 이십대들!!

데이터 시각화를 하는 이유?

I		II		III		IV	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Mean of X	11.0	**Correlation between X and Y**	0.875
Variance of X	10.0	**Linear regression**	y=3.0+0.5x
Mean of Y	7.5		
Variance of Y	3.75		

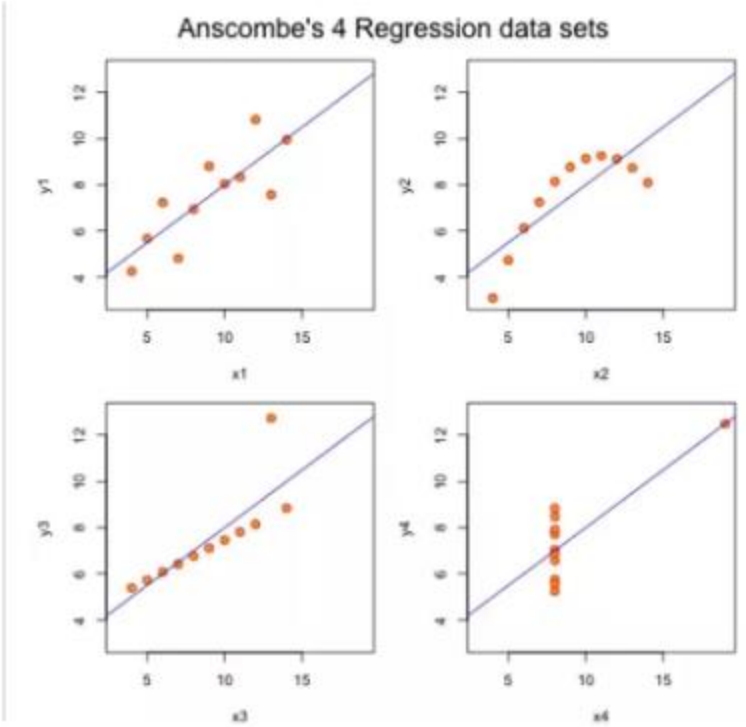
각 데이터셋은 얼마나 비슷할까?

파뿌리
파이썬 뿌시는 이십대들!!

데이터 시각화를 하는 이유?

I		II		III		IV	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

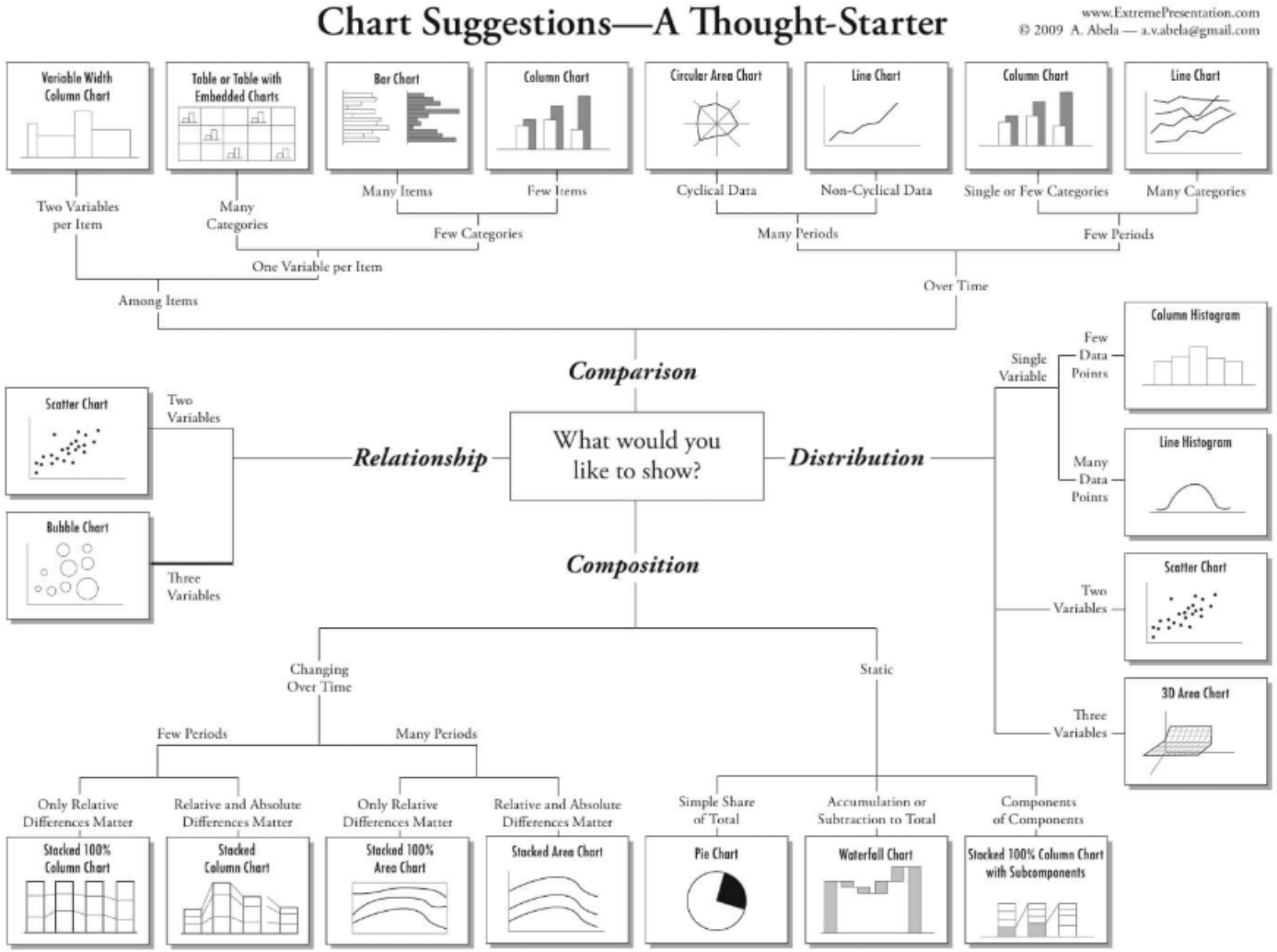
Mean of X	11.0	**Correlation between X and Y**	0.875
Variance of X	10.0	**Linear regression**	y=3.0+0.5x
Mean of Y	7.5		
Variance of Y	3.75		



요약 통계 정보만으로는 데이터를 정확하게 볼 수 없다.

파뿌리
파이썬 뿌시는 이십대들!!

데이터 시각화를 하는 이유?



http://project.newsjel.ly/depressed/

파뿌리
파이썬 뿌시는 이십대들!!

《 Round 4 》

- 데이터 시각화 개요 - complete
- 시각화 실습 《
- 이상치(outliar)

```python
import pandas as pd
import matplotlib as plt

# 데이터 읽어오기
df = pd.read_excel('남북한발전전력량.xlsx')

# 합계 데이터만 추출하기
df_ns = df.iloc[[0, 5], 2:]

# row index에 이름 붙혀주기
df_ns.index = ['south', 'north']

# columns의 데이터타입을 int형으로 변환
df_ns.columns = df_ns.columns.map(int)

# df_ns의 선그래프 그리기
df_ns.plot()

# 행 인덱스를 x축 데이터로 쓰기때문에 년도와 국가인 x y축을 바꿔줌
tdf_ns = df_ns.T
tdf_ns.plot()

# 찍은 plot을 모두 보여줌
plt.pyplot.show()
```

```python
import pandas as pd
import matplotlib as plt

# 데이터 읽어오기
df = pd.read_excel('남북한발전전력량.xlsx')

# 합계 데이터만 추출하기
df_ns = df.iloc[[0, 5], 2:]

# row index에 이름 붙혀주기
df_ns.index = ['south', 'north']

# columns의 데이터타입을 int형으로 변환
df_ns.columns = df_ns.columns.map(int)
```

	1990	1991	1992	1993	---	2016
'south'	1077	1186	1310	1444		5404
'north'	277	266	247	221		239

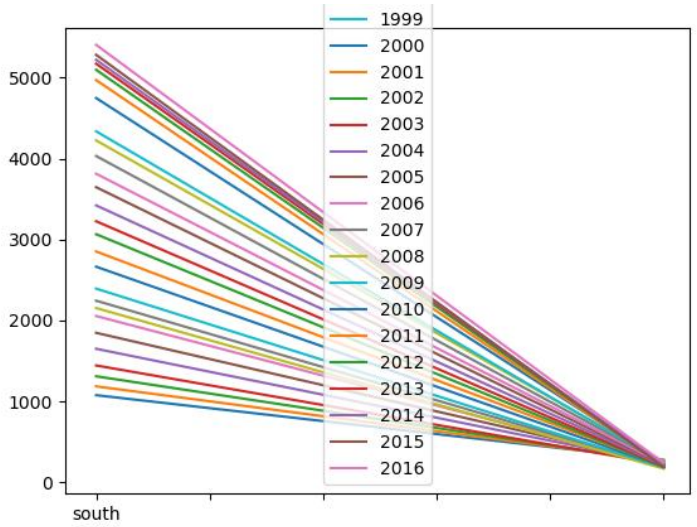
파뿌리
파이썬 뿌시는 이십대들!!

```
# df_ns의 선그래프 그리기
df_ns.plot()
```

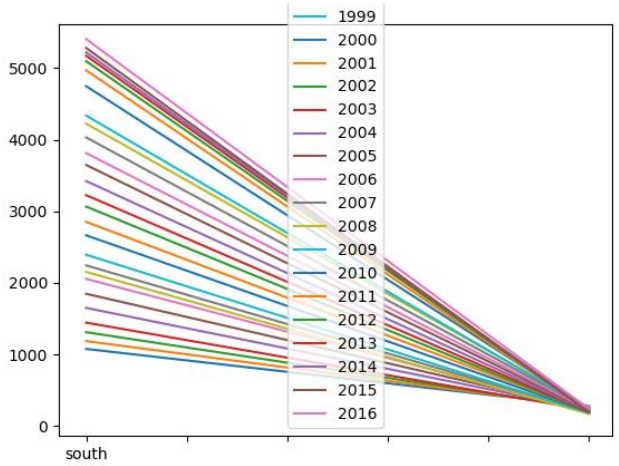
pandas에는 matplotlib의 기능을 일부 내장
DataFrame.plot()은 그래프를 찍어줌

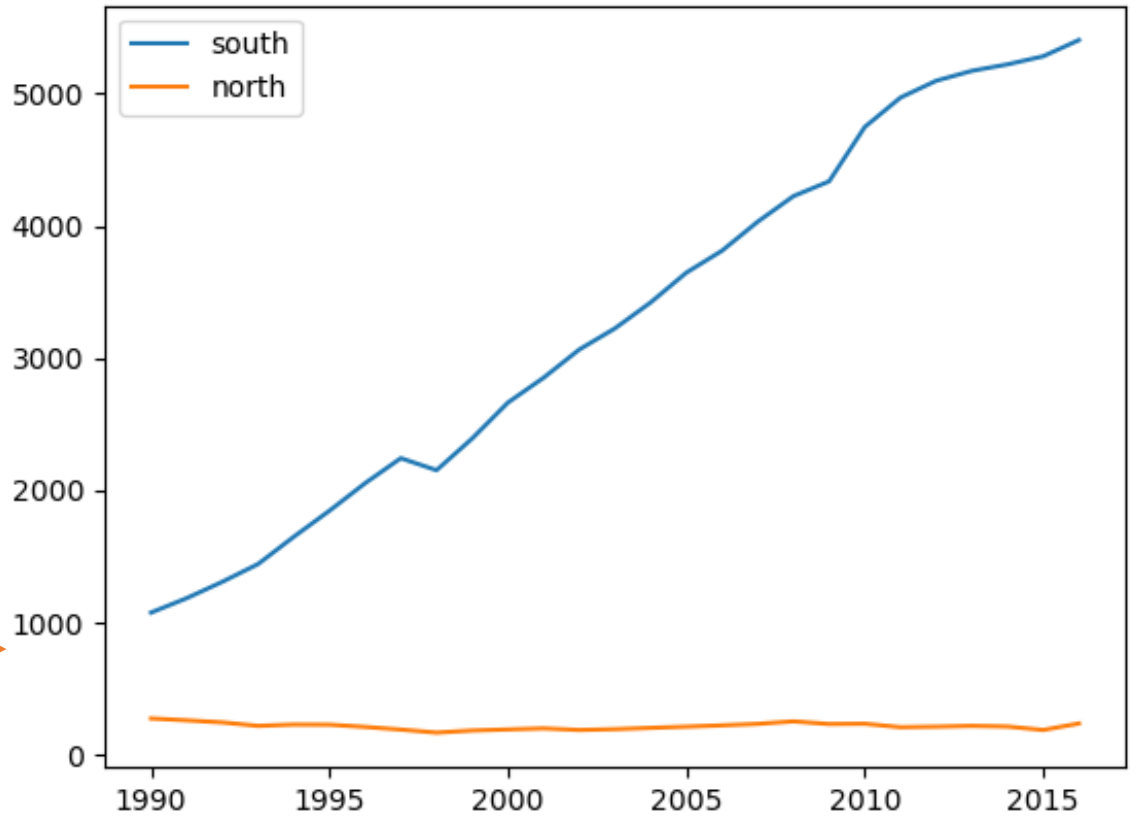
- row index x축으로 전달

- kind 옵션으로 그래프의 종류 선택가능
(default는 선 그래프)



```
# 행 인덱스를 x축 데이터로 쓰기때문에 년도와 국가인 x y축을 바꿔줌
tdf_ns = df_ns.T
tdf_ns.plot()

# 찍은 plot을 모두 보여줌
plt.pyplot.show()
```





Q. 다음 데이터셋으로 그리기 부적합한 그래프는?

	1990	1991	1992	1993	---	2016
'south'	1077	1186	1310	1444		5404
'north'	277	266	247	221		239

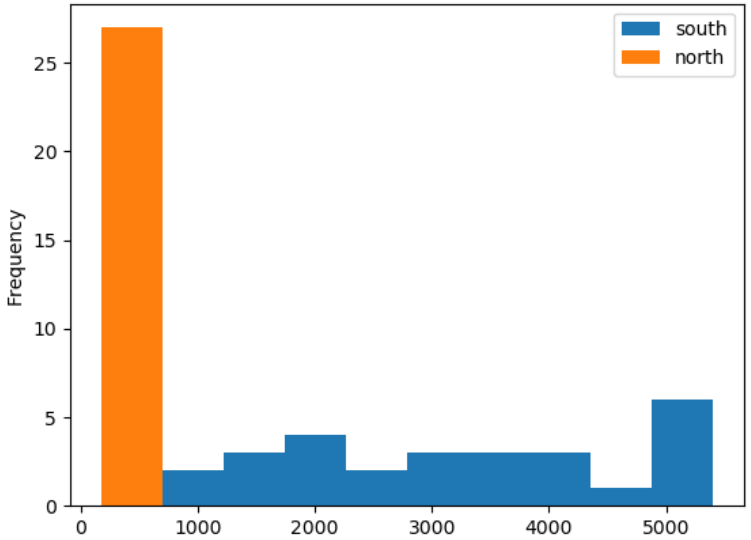
1) DataFrame.plot(kind = "bar") # 막대 그래프

2) DataFrame.plot(kind = "scatter") # 산점도

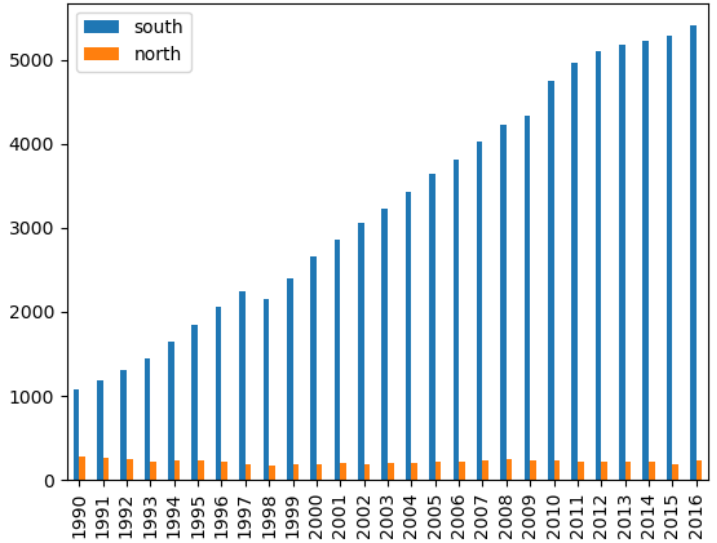
3) DataFrame.plot(kind = "hist") # 히스토그램

4) DataFrame.plot(kind = "box") # 박스플롯

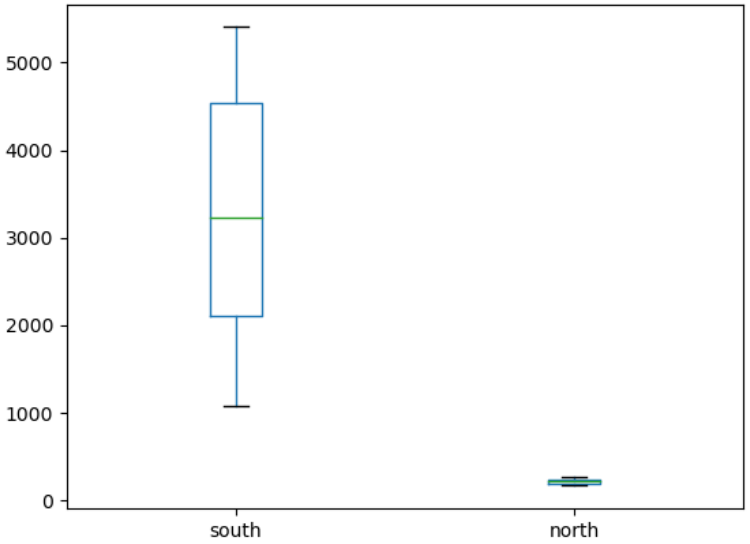
5) DataFrame.plot(kind = "pie") # 원 그래프



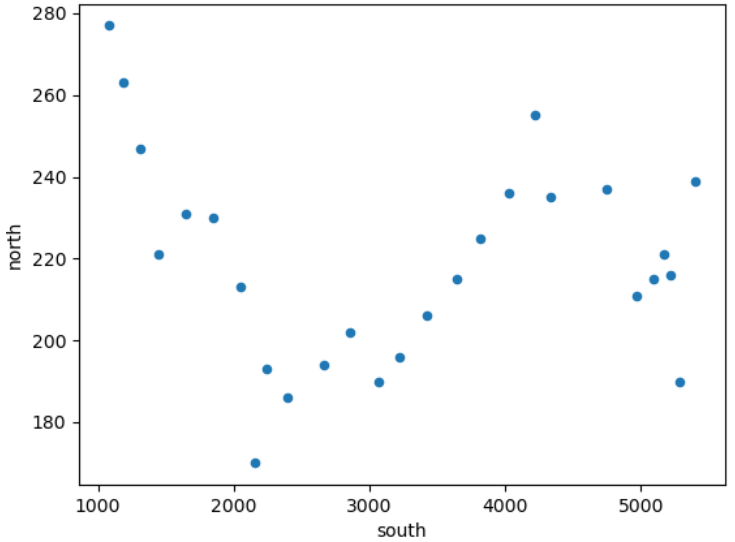
DataFrame.plot(kind = "hist")



DataFrame.plot(kind = "bar")



DataFrame.plot(kind = "box")



DataFrame.plot(kind = "scatter", x= "south", y = "north")

- matplotlib을 기반으로 하는 고급 시각화 도구

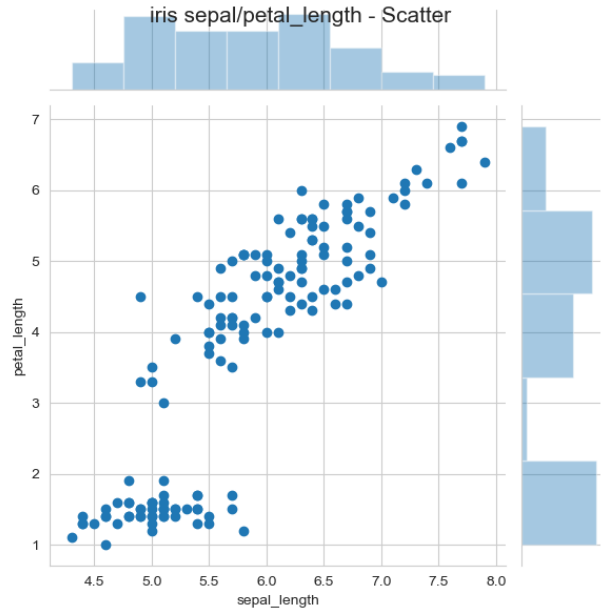
- 다양한 실습용 데이터 내장

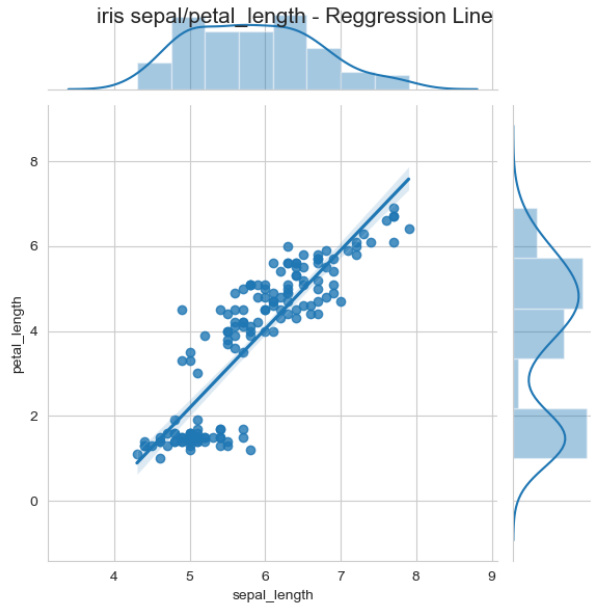
- pip install seaborn
import seaborn as sns

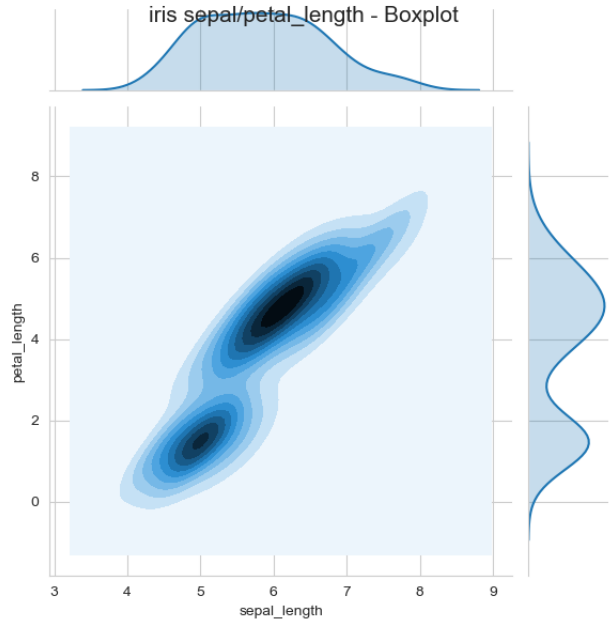
파뿌리
파이썬 뿌시는 이십대들!!

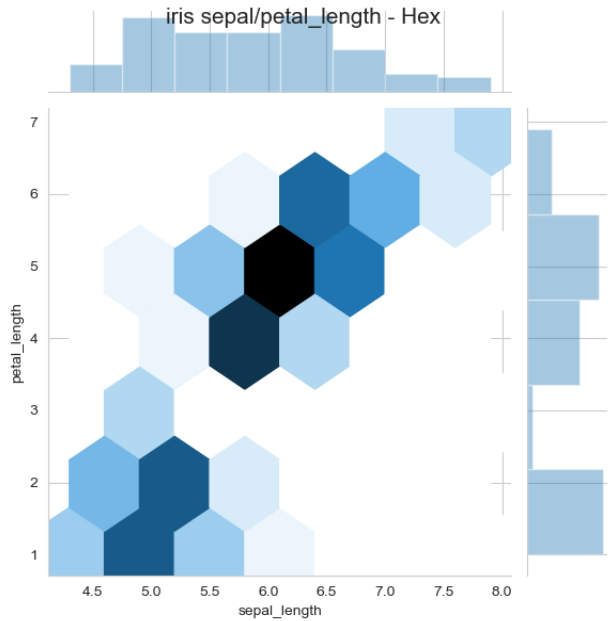
조인트 그래프(jointplot)

```
# seaborn의 iris(붓꽃) 데이터셋 불러오기
iris = sns.load_dataset("iris")

# 스타일 테마 설정(darkgrid, whitegrid, dark, white, ticks)
sns.set_style('whitegrid')

# 조인트 그래프1 - 산점도(기본값)
joi_1 = sns.jointplot(x='sepal_length', y='petal_length', data=iris)

# 조인트 그래프2 - 회귀선
joi_2 = sns.jointplot(x='sepal_length', y='petal_length', kind='reg', data=iris)

# 조인트 그래프3 - 커널밀집 그래프
joi_3 = sns.jointplot(x="sepal_length", y='petal_length', kind="kde", data=iris)

# 조인트 그래프4 - 육각 그래프
joi_4 = sns.jointplot(x='sepal_length', y='petal_length', kind="hex", data=iris)

# 그래프 제목 표시
joi_1.fig.suptitle('iris sepal/petal_length - Scatter', size=15)
joi_2.fig.suptitle('iris sepal/petal_length - Reggression Line', size=15)
joi_3.fig.suptitle('iris sepal/petal_length - Boxplot', size=15)
joi_4.fig.suptitle('iris sepal/petal_length - Hex', size=15)

plt.pyplot.show()
```


iris sepal/petal_length - Scatter
petal_length
sepal_length

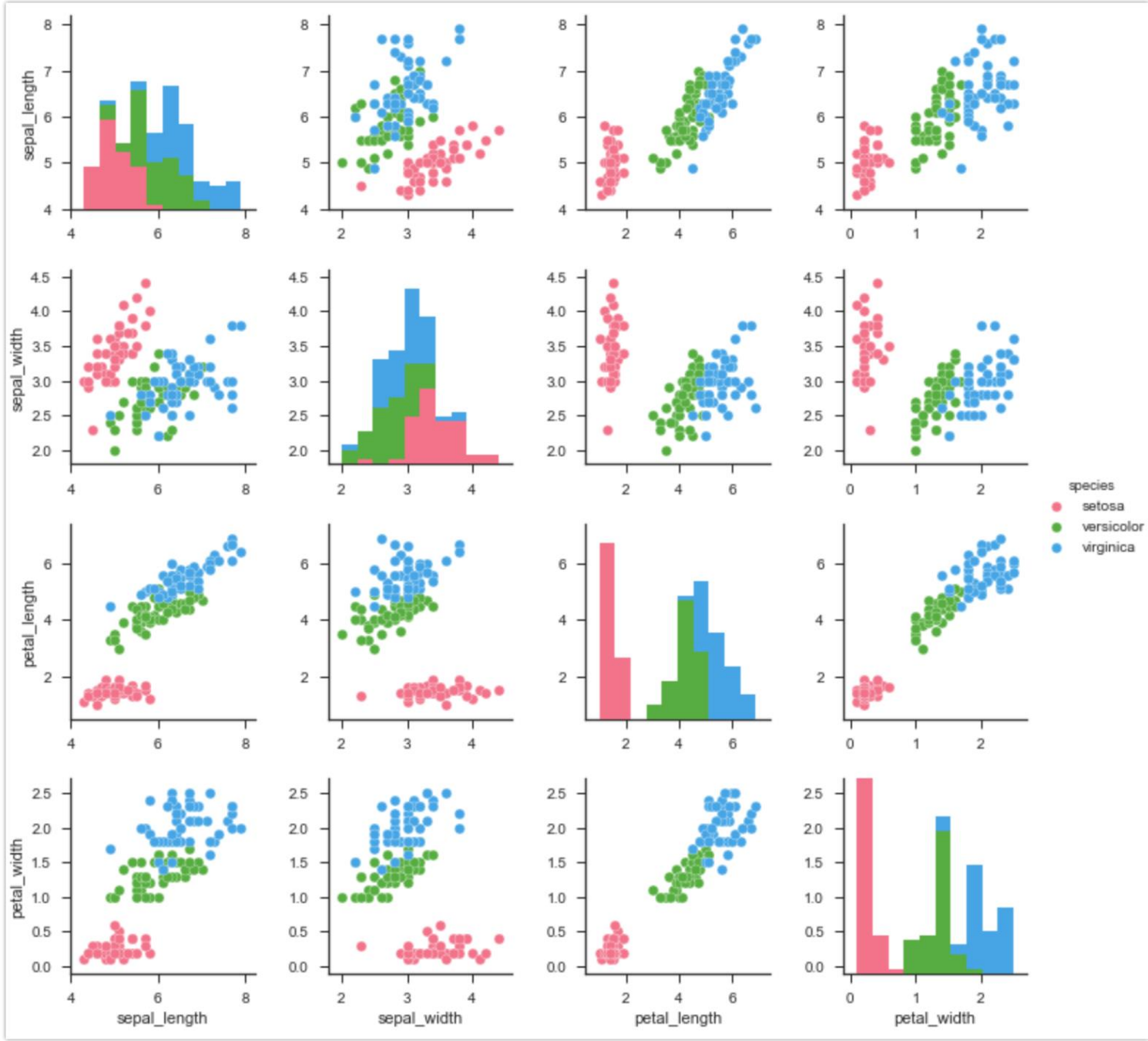

iris sepal/petal_length - Reggression Line
petal_length
sepal_length


iris sepal/petal_length - Boxplot
petal_length
sepal_length


iris sepal/petal_length - Hex
petal_length
sepal_length

이변수 데이터 분포(pairplot)

```
# seaborn의 iris(붓꽃) 데이터셋 불러오기
iris = sns.load_dataset("iris")

# 스타일 테마 설정(darkgrid, whitegrid, dark, white, ticks)
sns.set(style = "whitegrid")

# 조건에 따라 그리드 나누기
# hue : Variable in data to map plot aspects to different colors
# palette : set of colors for mapping the hue variable
# more info of parameters... -> https://seaborn.pydata.org/generated/seaborn.pairplot.html
g = sns.pairplot(iris, hue = "species", palette="husl")

plt.pyplot.show()
```


sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica

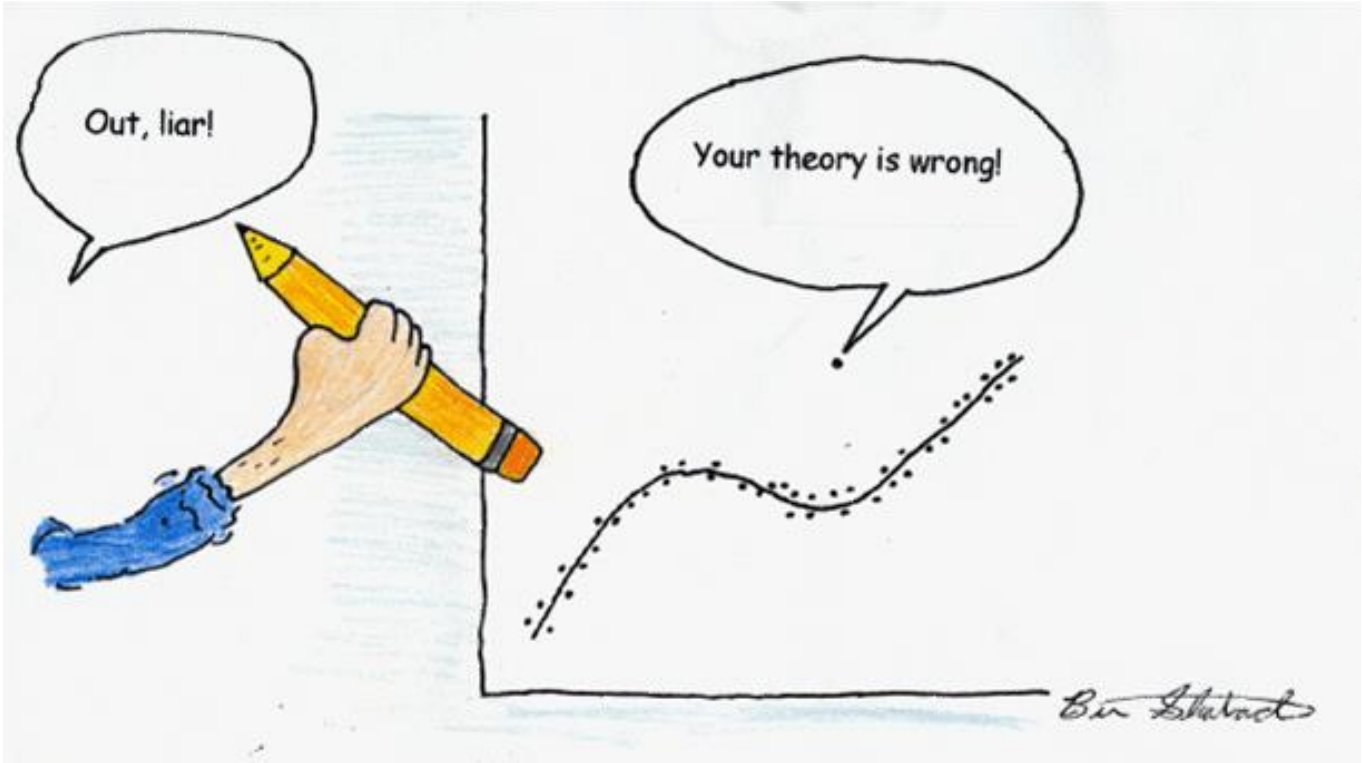
파뿌리
파이썬 뿌시는 이십대들!!

《 Round 4 》

- 데이터 시각화 개요 - complete
- 시각화 실습 - complete
- 이상치(outliar) 《

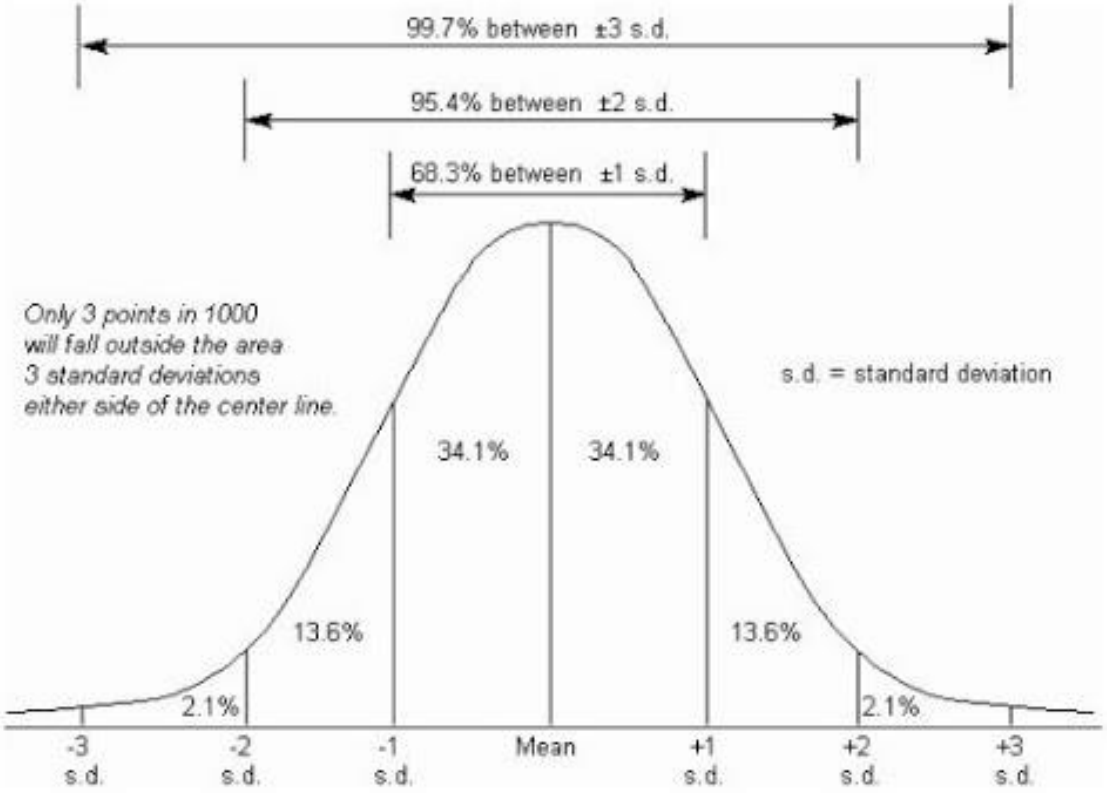
파뿌리
파이썬 뿌시는 이십대들!!

이상치(Outliar)



전체적인 데이터/샘플 범위에서 동떨어진 관측값으로,
모델을 크게 왜곡시킬 가능성이 있음.

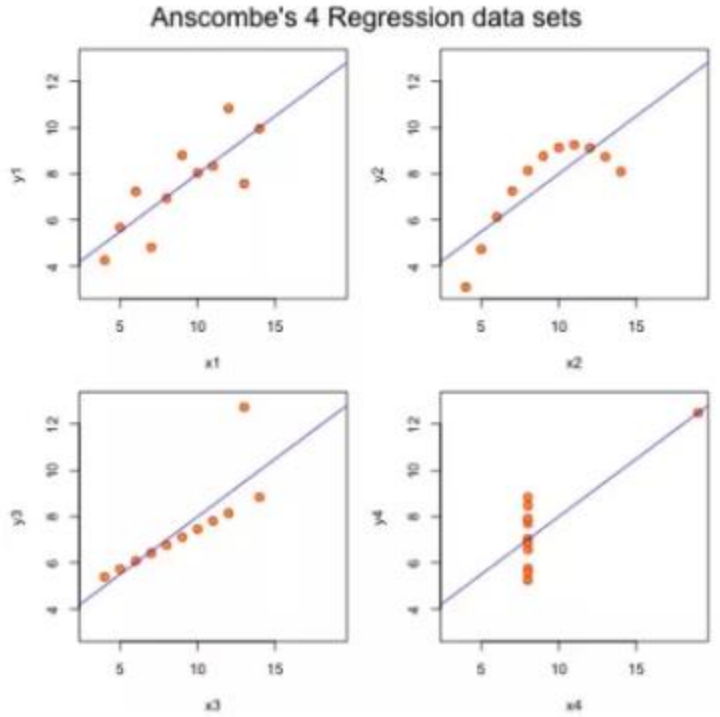
Outliar를 결정하는 방법?



일반적으로 6σ, 즉 +- 3표준편차에 해당하는 값을 outliar라고 봄
목적과 자료에 따라 3σ, 4σ, 5σ로도 설정

IQR방식, 앤드류스 그림, 마하라노비스 거리로도 이상치 결정가능

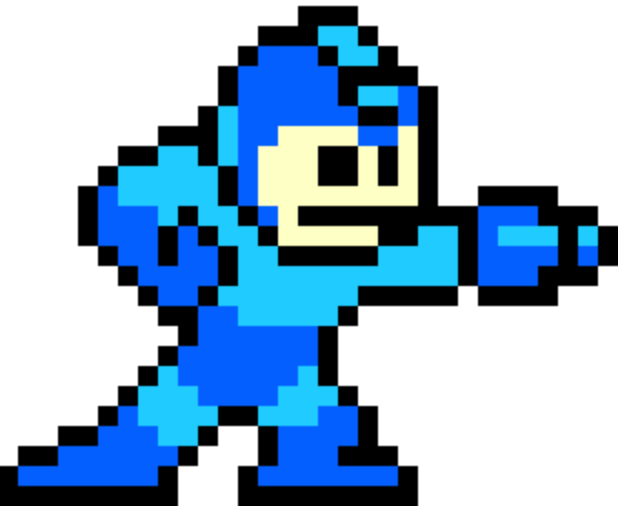
Outliar를 결정하는 방법?



가장 직관적인 방법은 시각화!
시각화를 통해서 이상치를 직관적으로 확인할 수 있음.

파뿌리
파이썬 뿌시는 이십대들!!

WARNING

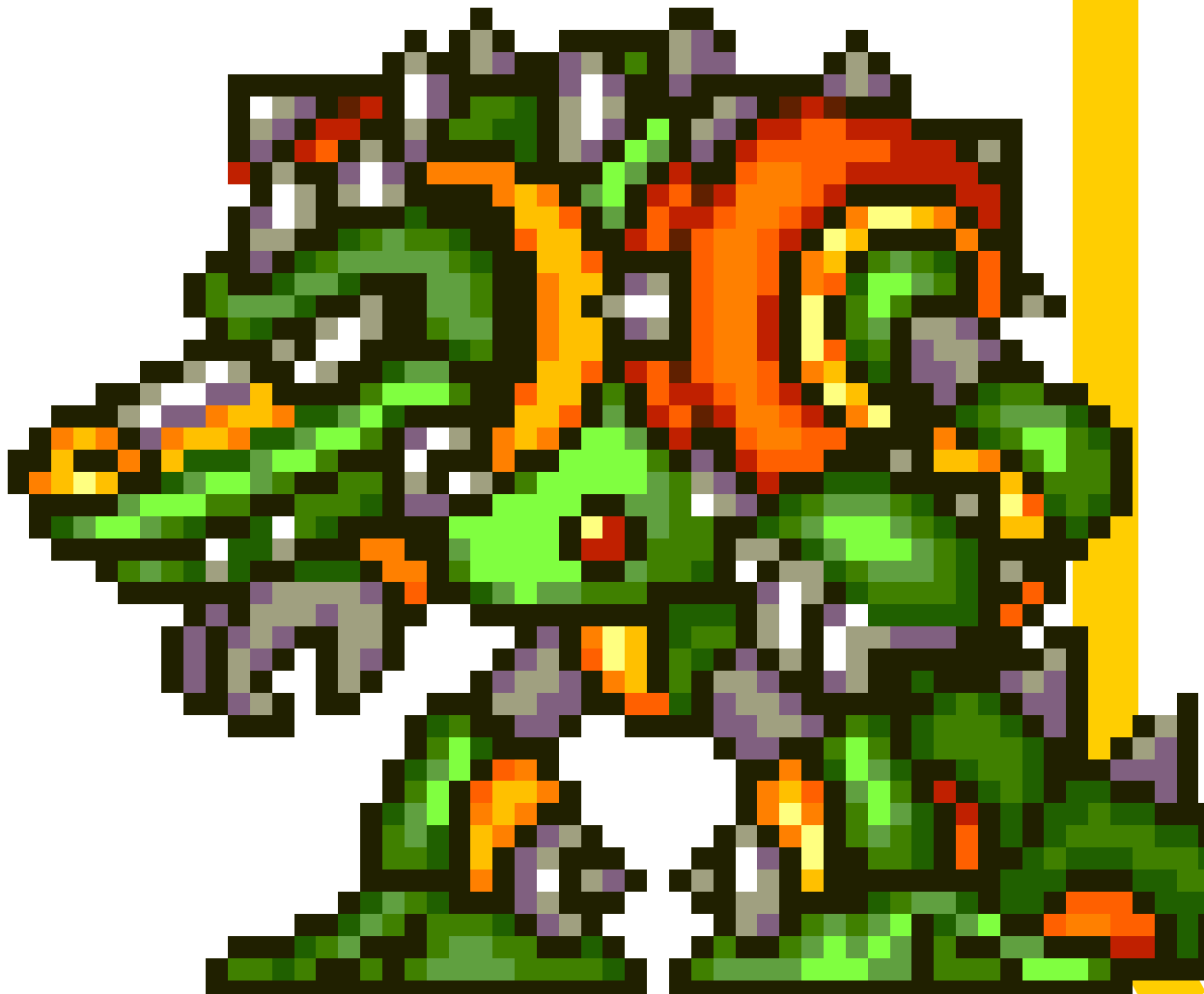


파뿌리
파이썬 뿌시는 이십대들!!

WARNING

《 QUEST 》

1. outlier 결정법 조사
2. tips 데이터 시각화 문제 풀기



파뿌리
파이썬 뿌시는 이십대들!!

NEXT STAGE