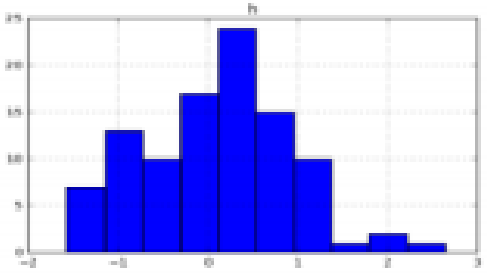
파이썬 판다스로 데이터 시각화_ 통계 그래프(2)

Scatter Plot(산점도) 그리기

Plotting

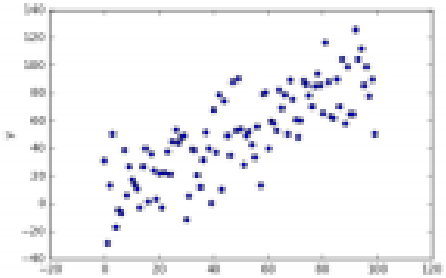
`df.plot.hist()`

Histogram for each column



`df.plot.scatter(x='w',y='h')`

Scatter chart using pairs of points



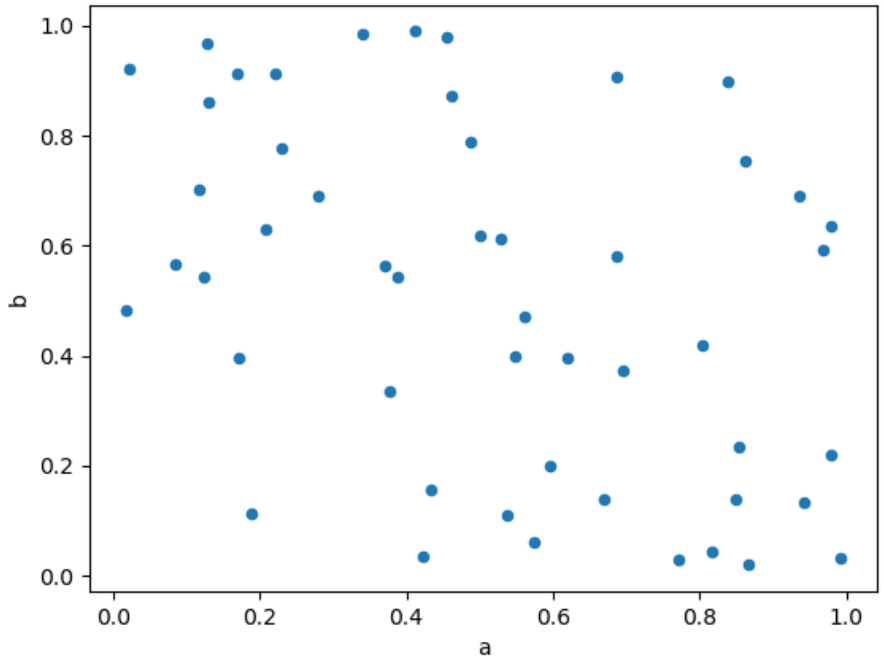
Scatter Plot(산점도) 그리기

Scatter plot

Scatter plot can be drawn by using the `DataFrame.plot.scatter()` method. Scatter plot requires numeric columns for the x and y axes. These can be specified by the `x` and `y` keywords.

```
In [63]: df = pd.DataFrame(np.random.rand(50, 4), columns=['a', 'b', 'c', 'd'])

In [64]: df.plot.scatter(x='a', y='b');
```



Scatter Plot(산점도) 그리기

```
df.plot.scatter(x='a', y='b', s=50, grid=True)
```

- x, y축의 값 반드시 지정(지정하지 않으면 오류 발생)
- s=50(s: Size 크기), grid : 격자

Scatter Plot(산점도) 그리기

이론 영상 후 실습 영상 제시

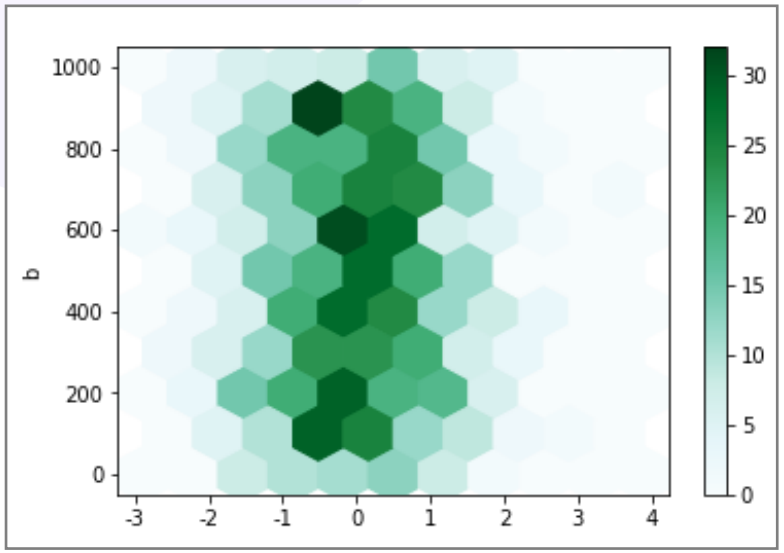
Hexbin Plot 그리기

01 Hexbin Plot(Hexagonal Bin Plot)

- 데이터가 클 때 각각의 점을 산점도(Scatter Plot)로 표현할 때의 단점을 보완할 수 있는 그래프
- 육각형 모양의 Bin을 생성하여 그래프로 표현
  - 데이터의 크기 비교 가능
- Histogram과 산점도를 혼합한 형태

Hexbin Plot 그리기

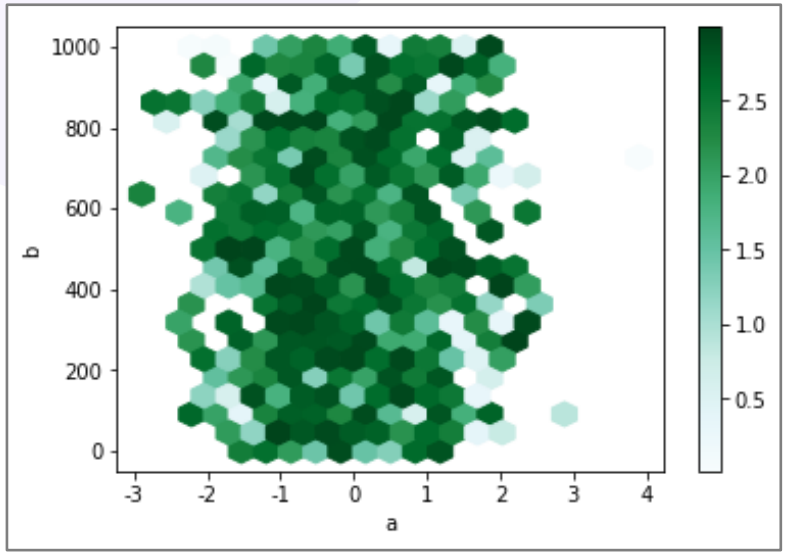
02 df.plot.hexbin(x='a', y='b', gridsize=10)



- x, y축의 값 반드시 지정
  - 지정하지 않으면 오류 발생
- gridsize=10 : 격자의 Size(크기)를 조절 가능

Hexbin Plot 그리기

03 df.plot.hexbin(x='a', y='b', C='z', reduce_C_function=np.max, gridsize=20)



- 기본적으로 각 (x, y) 점 주변의 개수에 대한 히스토그램이 계산됨
- C 및 reduce_C_function 인수에 값을 전달하여 대체 집계를 지정할 수 있음
- C는 각 (x, y) 점에서 값을 지정하고 reduce_C_function은 Bin의 모든 값을 단일 숫자로 줄이는 하나의 인수의 함수

  예 평균, 최대 값, 합계, 표준

  ➡ 예에서 위치는 열(Column) a 및 b에 의해 주어지며, 값은 열 (z)에 의해 주어진다.

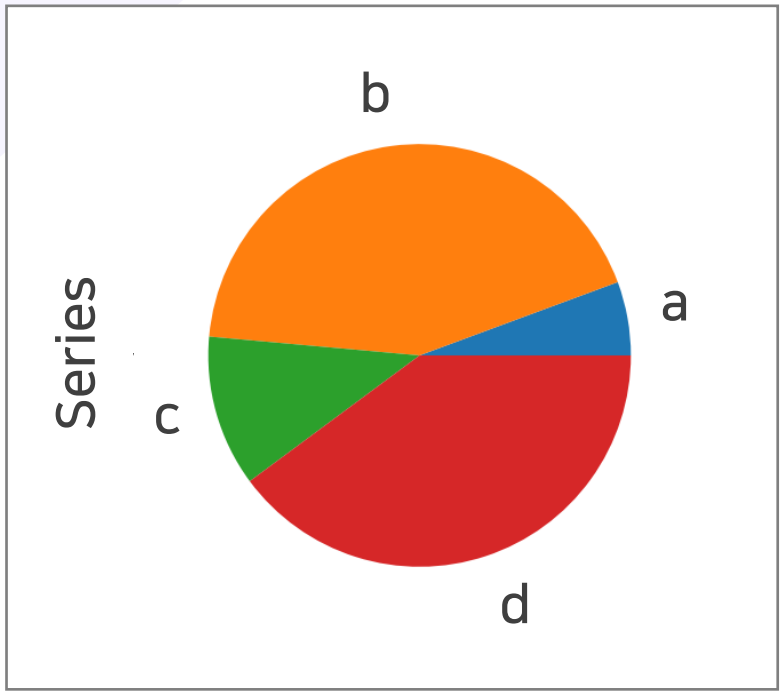
- Bin은 NumPy의 max 함수로 집계됨

Hexbin Plot 그리기

이론 영상 후 실습 영상 제시

Pie Plot 그리기

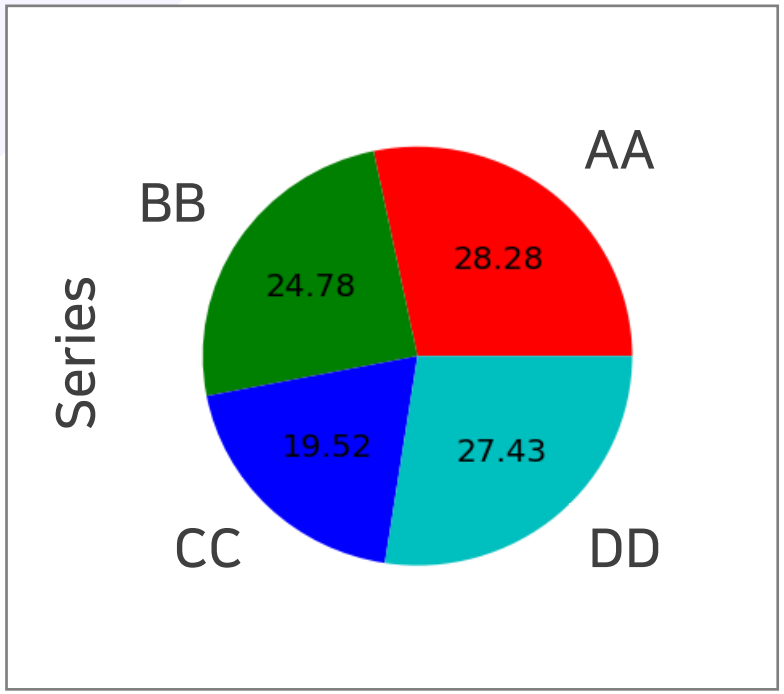
01 series.plot.pie(figsize=(6, 6))



- Seaborn 등 다른 모듈에서는 파이 차트 지원하지 않음
- 면적으로 나타나 오해의 소지가 존재하기 때문
- Data Frame 뒤에 plot.pie() 함수를 호출하면 구현됨

Pie Plot 그리기

02 series.plot.pie(labels=['AA', 'BB', 'CC', 'DD'], colors= ['r', 'g', 'b', 'c'], autopct='%.2f', fontsize=20, figsize=(6, 6))



- autopct='%.2f'
  - 소수점 두 번째 자리까지 표현

Pie Plot 그리기

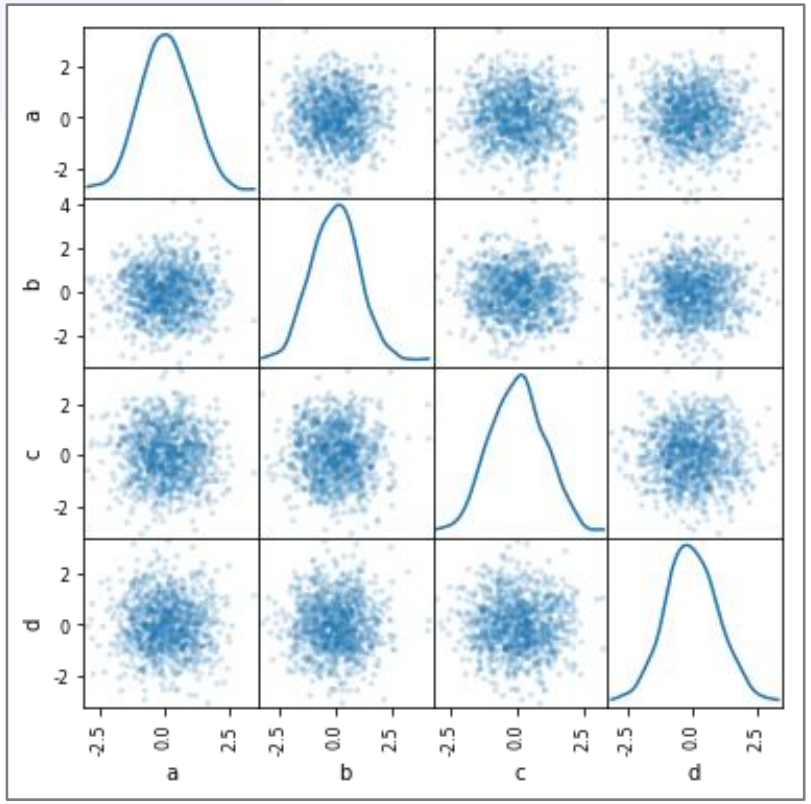
이론 영상 후 실습 영상 제시

파이썬 판다스로 데이터 시각화_ 통계 그래프(3)

Scatter Matrix Plot 산점도 행렬

KDE(커널밀도함수)

```
scatter_matrix(df, alpha=0.2, figsize=(6, 6), diagonal='kde')
```



- alpha
  - float, optional, amount of transparency applied(투명도)
- diagonal=kde
- diagonal
  - {'hist', 'kde'} 대각선에 히스토그램 또는 커널밀도함수 그리기
- pick between 'kde' and 'hist' for either Kernel Density Estimation or Histogram

Scatter Matrix Plot 산점도 행렬

이론 영상 후 실습 영상 제시

Kernal Density Estimate Plot 커널밀도함수

커널밀도추정 (KDE)

- 통계에서 임의 변수의 확률밀도함수(PDF)를 추정하는 비모수적 방법임
  - 이 함수는 가우스 커널을 사용하며 자동 대역폭 결정을 포함함

비모수 통계

- 정규분포로 표현하지 못함

Kernal Density Estimate Plot 커널밀도함수

커널밀도 추정치

- 히스토그램과 밀접한 관련이 있지만 적절한 커널을 사용하여 매끄럽고 연속성과 같은 속성을 부여 할 수 있음
  - 이산적으로 끊어져 있는 히스토그램을 부드럽게 연결함

커널 함수

- 원점을 중심으로 대칭이며 적분 값이 1인 함수임

Kernal Density Estimate Plot 커널밀도함수

정규분포

정규 분포(Normal Distribution) 또는
가우시안 분포(Gaussian Distribution)는 연속 확률 분포의 하나임

정규분포는 수집된 자료의 분포를 근사하는 데에 자주 사용됨

- 이것은 중심극한정리에 의하여 독립적인 확률변수들의 평균은 정규분포에 가까워지는 성질이 있기 때문임

Kernal Density Estimate Plot 커널밀도함수

이론 영상 후 실습 영상 제시

실데이터 실습(1)

서울 코로나 19 실데이터 수집

<출처 : https://www.seoul.go.kr/coronaV/coronaStatus.do>

서울 코로나 19 실데이터 수집

01 url = http://www.seoul.go.kr/coronaV/coronaStatus.do

- 수집할 데이터가 있는 url 주소를 변수 url에 대입함

02 table = pd.read_html(url)

- 판다스의 read_html() 모듈로 url 주소의 테이블로 구성된 데이터를 가져옴

03 table[0].T

- T : Transpose(전치)

서울 코로나 19 실데이터 수집

01 df.to_csv(file_name, index=False)

- df.to_csv() : 데이터프레임을 csv 파일로 변환
- index=False : index는 저장하지 않기 위함

02 df.to_csv(file_name, index=False, encoding='cp949')

- Excel에서 읽으려고 할 때, encoding='cp949'
- UnicodeEncodeError : unicode 에러 발생

서울 코로나 19 실데이터 수집

03 last_day = df.loc[0, "확진일"]

- loc : 행의 순서로 접근, 데이터 추출
- 여기서는 0번째 행의 “확진일” 데이터를 last_day 변수로 대입함

04 last_day = last_day.replace(".", "_")

- replace(“/", "_") : “/” 표기를 “_”로 변경

05 file_name = f"seoul_covid_{last_day}.csv"

- f”{}.csv” : 파일 이름 {}로 자동화하여 csv 파일을 업데이트해주기 위함
  - 지속적으로 데이터를 업데이트할 때 유용

서울 코로나 19 실데이터 수집

이론 영상 후 실습 영상 제시

학습평가

학습한 내용을 바탕으로 다음 문제를 풀어 보세요.

번호	문제	정답	해설
1	데이터가 클 때 각각의 점을 산점도(Scatter Plot)로 표현할 때의 단점을 보완할 수 있는 그래프로 Histogram과 산점도를 혼합한 형태에 해당하는 그래프는 무엇인가? ① 산점도(Scatter Plot) ② Hexbin Plot ③ Pie Plot ④ Scatter Matrix Plot	2	Hexbin Plot 에 대한 설명이다.
2	다음 코드에서 autopct='%.2f' 가 의미하는 것은? `series.plot.pie(labels=['AA', 'BB', 'CC', 'DD'], colors=['r', 'g', 'b', 'c'], autopct='%.2f', fontsize=20, figsize=(6, 6))` ① 실수형으로 0.2를 출력하라 ② 정수형으로 0.2를 출력하라 ③ 소수점 두 번째 자리까지 출력하라 ④ 소수점 0.2%를 출력하라	3	autopct='%.2f' 는 소수점 두 번째 자리까지 출력하라는 의미이다.
3	커널 밀도 함수에 대한 설명으로 올바르지 않은 것은? ① 커널 밀도 추정치는 히스토그램 과 밀접한 관련이 있지만 적절한 커널을 사용하여 매끄럽고 연속성과 같은 속성을 부여 할 수 있다. ② 통계에서 커널 밀도 추정(KDE)은 임의 변수의 확률 밀도 함수(PDF)를 추정하는 모수적 방법이다. ③이 함수는 가우스 커널을 사용하며 자동 대역폭 결정을 포함한다. ④ 커널 함수는 원점을 중심으로 대칭이며 적분 값이 1인 함수다.	2	통계에서 커널 밀도 추정(KDE)은 임의 변수의 확률 밀도 함수(PDF)를 추정하는 비모수적 방법이다.

화면설명

[학습평가 페이지]
- 페이지 퀴즈 컴포넌트 사용하여 페이지 개발