한빛아카데미 Hanbit Academy, Inc.

난생처음

데이터 분석 with 파이썬

상관관계 분석

목차

01

상관관계 분석의 개념

01. 상관관계 분석의 개념

Ⅰ. 상관관계 분석과 상관계수

- 상관관계 분석(Correlation analysis, 상관분석)
  - 두 변수 사이 관계의 강도와 방향을 파악하는 통계 기법.
  - 상관관계의 강도를 나타낸 수치를 상관계수(Correlation coefficient).
  - 변수 x와 y가 있을 때 두 변수의 상관관계는 다음 세 가지 중 하나.
    - » 양의 상관관계: 변수 x가 커질수록 변수 y도 커짐.
    - » 음의 상관관계: 변수 x가 커질수록 변수 y는 작아짐.
    - » 상관관계 없음: 변수 x가 커질 때 변수 y는 커질 수도, 작아질 수도 있음.

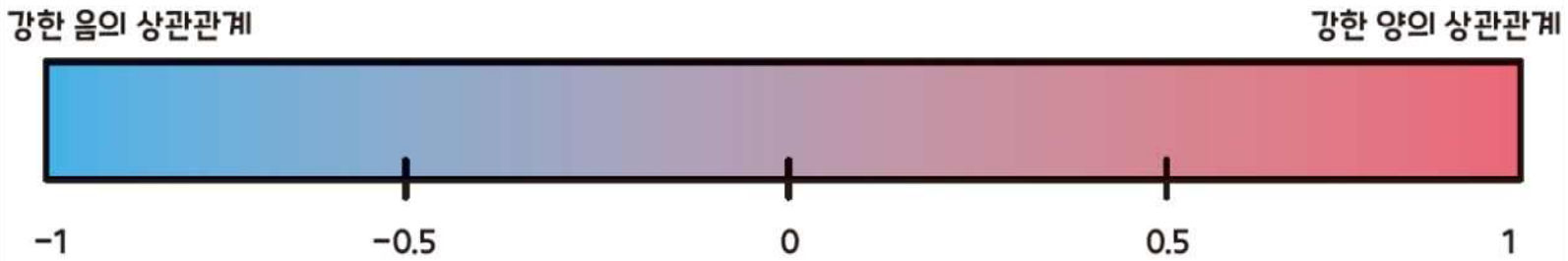


그림 9-1 상관계수

01. 상관관계 분석의 개념

Ⅰ. 상관관계 분석과 상관계수

- 상관관계 분석(Correlation analysis, 상관분석)
  - [그림 9-2]는 두 변수를 각각 x축과 y축으로 하여 나타낸 그래프

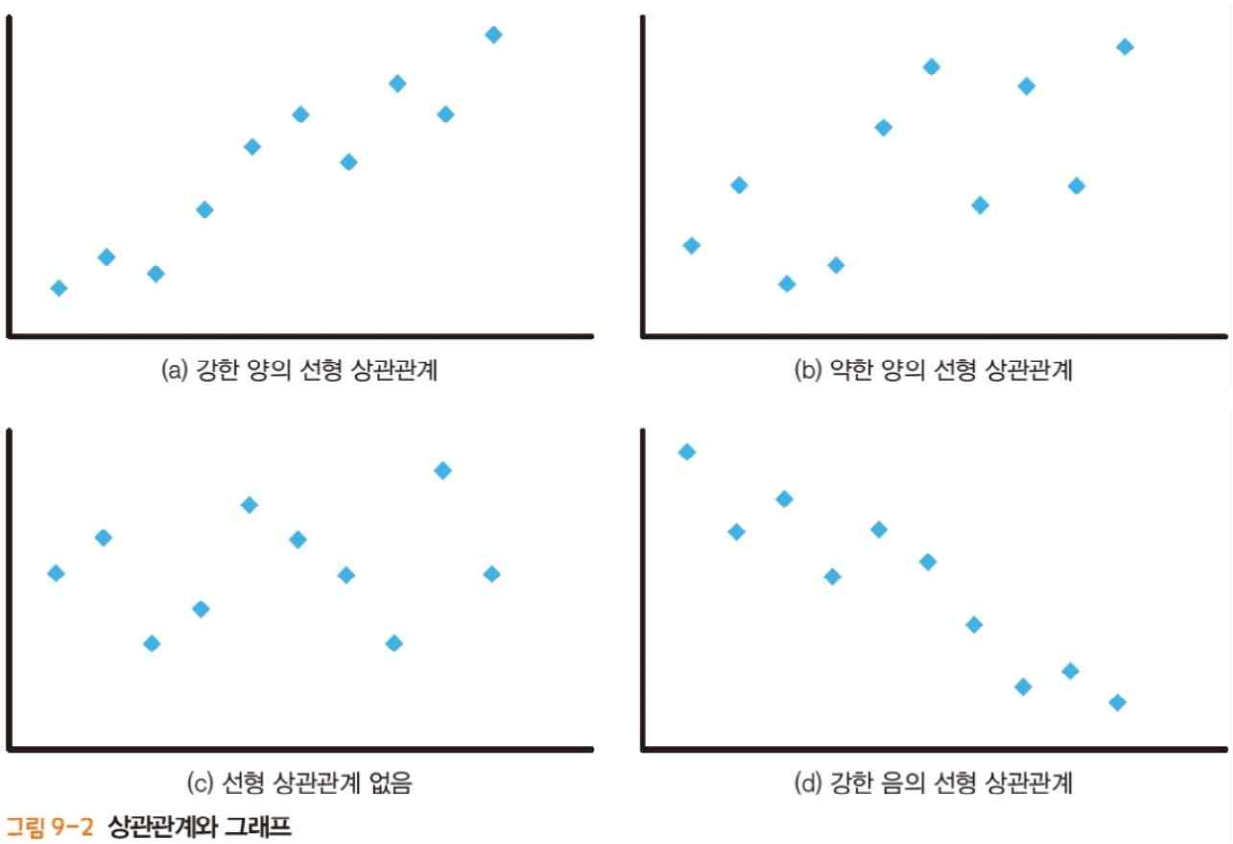


그림 9-2 상관관계와 그래프

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 상관관계 분석(Correlation analysis, 상관분석)

  ① 피어슨 상관분석(Pearson correlation analysis)은 가장 일반적인 상관분석 방법.

  ② 스피어만 상관분석(Spearman correlation analysis)은 두 변수가 정규성을 보이지 않을 때 사용하기 적합한 방법.

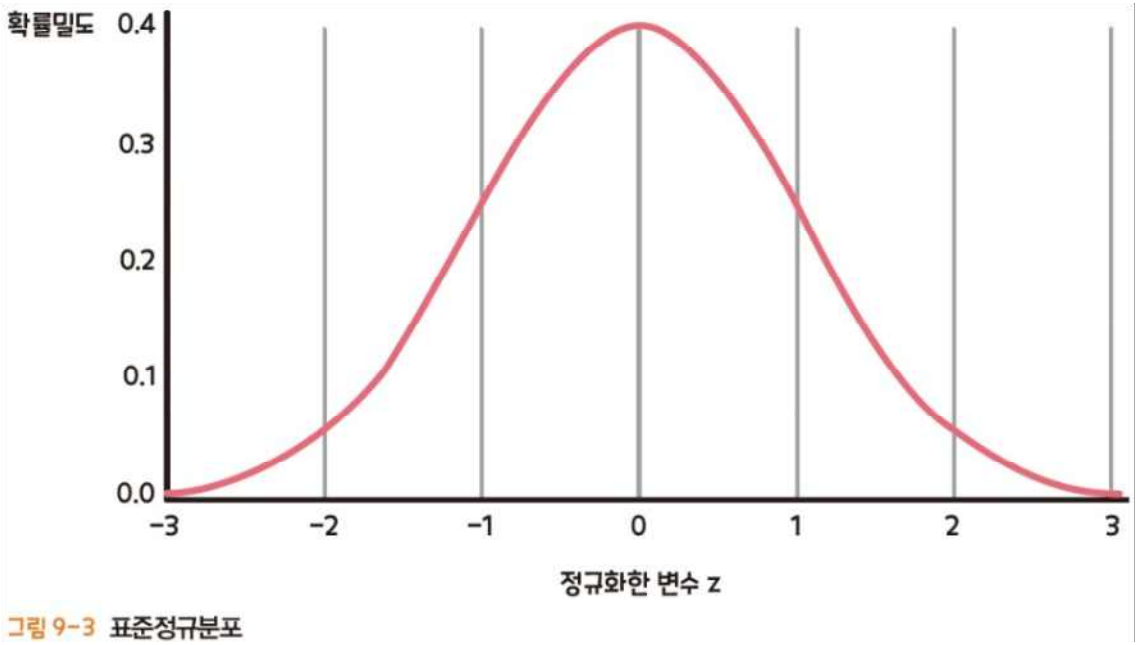


그림 9-3 표준정규분포

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 상관관계 분석(Correlation analysis, 상관분석)

  ③ 세 번째 켄달 상관분석(Kendall correlation analysis)은 스피어만 상관분석과 비슷하나 표본 데이터가 적고 동점이 많을 때 사용하기 적합한 방법.

✓**하나 더 알기: 상관관계 분석의 의미**

상관계수만 가지고 두 변수 사이의 상관성이 있는지 없는지 판단할 수는 없음.
두 변수에 선형 상관관계가 아닌 다른 상관관계가 있을 수 있음.

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

그림 9-4 영어 동요

- 피어슨 상관분석
  - 어린이가 영어 동요에 노출된 시간과 영어 점수와의 상관관계 분석해보기.

[코드 9-1] 피어슨 상관분석(1)

```
import pandas as pd
#리스트에 데이터 삽입하기
engListening = [30, 60, 90]
engScore = [70, 80, 90]

#리스트를 데이터프레임으로 변환하기
data = {'engListening':engListening, 'engScore':engScore}
df = pd.DataFrame(data)

#상관분석 수행하기
coef = df.corr(method='pearson')
print(coef)
```

```
             engListening  engScore
engListening           1.0       1.0
engScore               1.0       1.0
```

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 피어슨 상관분석
  - 데이터를 조금 더 추가하여 상관계수를 구하기.
  - 리스트 engListening과 engScore에 [표 9-1]의 데이터 5쌍을 추가.

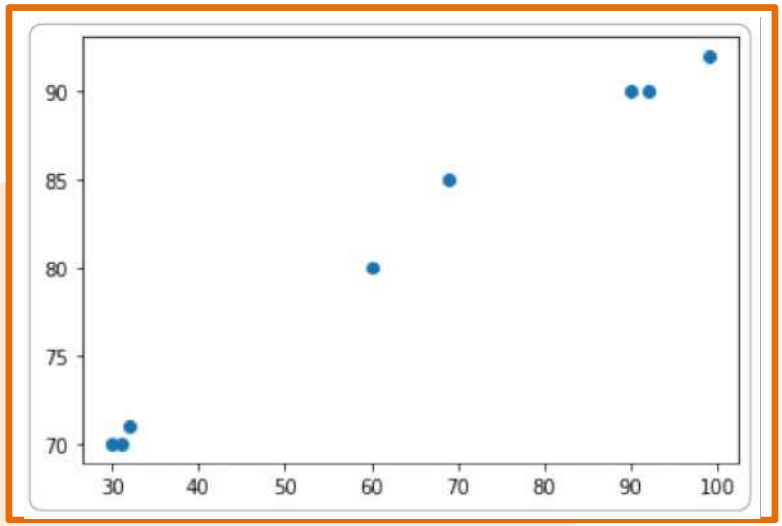
표 9-1 추가할 데이터

engListening	31	32	69	92	99
engScore	70	71	85	90	92

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 피어슨 상관분석
  - 데이터의 산포도를 산점도 그래프로 확인.



[코드 9-2] 산포도 확인

```
import matplotlib.pyplot as plt

#데이터 추가하기
engListening = [30, 60, 90, 31, 32, 69, 92, 99]
engScore = [70, 80, 90, 70, 71, 85, 90, 92]
data2 = {'engListening':engListening, 'engScore':engScore}
df2 = pd.DataFrame(data2)

#산점도 그래프의 x좌표와 y좌표 설정하기
plt.scatter(df2['engListening'], df2['engScore'])
plt.show()
```

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 피어슨 상관분석

[코드 9-3] 피어슨 상관분석(2)

```
coef = df2.corr(method='pearson')
print(coef)
```

```
             engListening   engScore
engListening     1.000000   0.995829
engScore         0.995829   1.000000
```

- 데이터를 추가한 데이터프레임 data2의 선형 상관도는 0.995829.
- 매우 강한 선형 상관성이 있다고 말할 수 있음.

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 스피어만 상관분석과 켄달 상관분석

[코드 9-4] 스피어만 상관분석과 켄달 상관분석

```
#스피어만 상관분석
spearmanCoef = df.corr(method='spearman')
print(spearmanCoef)

#켄달 상관분석
kendallCoef = df.corr(method='kendall')
print(kendallCoef)
```

```
               engListening engScore
engListening            1.0      1.0
engScore                1.0      1.0
               engListening engScore
engListening            1.0      1.0
engScore                1.0      1.0
```

– 판다스의 corr() 함수에서 method 인자 'pearson'을 'spearman'과 'kendall'로 변경.

– 상관계수의 값은 분석 방법 종류에 따라 조금씩 다를 수 있음.

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 스피어만 상관분석과 켄달 상관분석

✓**하나 더 알기: 스피어만 상관분석과 피어슨 상관분석의 차이점**

피어슨 상관분석은 두 연속 변수 간의 선형 관계를 측정하는 반면, 스피어만 상관분석은 선형인지 여부에 관계없이 변수 간의 단조 연관성(Monotonic relationship)을 측정.

피어슨 상관관계가 스피어만 상관관계보다 데이터의 이상치(Outlier)에 민감하게 반응.

– 더 많은 변수 활용.

표 9-2 추가할 변수 및 원소

engReading	40	45	60	20	15	70	60	80
engClass	60	120	120	60	60	180	120	120

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 스피어만 상관분석과 켄달 상관분석

[코드 9-5] 데이터 추가

```
engListening = [30, 60, 90, 31, 32, 69, 92, 99]
engReading = [40, 45, 60, 20, 15, 70, 60, 80]
engClass = [60, 120, 120, 60, 60, 180, 120, 120]
engScore = [70, 80, 90, 70, 71, 85, 90, 92]

data3 = {'engListening':engListening, 'engReading':engReading,
'endClass':engClass, 'engScore':engScore}
df3 = pd.DataFrame(data3)
```

  - 리스트 engReading과 engClass에 데이터를 할당, 리스트 4개를 열이 4개인 데이터 프레임으로 변환.

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 스피어만 상관분석과 켄달 상관분석
  - 데이터프레임 df3으로 피어슨/스피어만/켄달 상관분석을 각각 수행.

[코드 9-6] 상관분석 결과

```
#피어슨 상관분석
pearsonCoef = df3.corr(method='pearson')
print(pearsonCoef)

#스피어만 상관분석
spearmanCoef = df3.corr(method='spearman')
print(spearmanCoef)

#켄달 상관분석
kendallCoef = df3.corr(method='kendall')
print(kendallCoef)
```

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 스피어만 상관분석과 켄달 상관분석

[코드 9-6] 실행결과

```
             engListening engReading endClass engScore
engListening    1.000000   0.877201 0.703028 0.995829
engReading      0.877201   1.000000 0.808755 0.894111
endClass        0.703028   0.808755 1.000000 0.759453
engScore        0.995829   0.894111 0.759453 1.000000
             engListening engReading endClass engScore
engListening    1.000000   0.826362 0.717256 0.988024
engReading      0.826362   1.000000 0.852757 0.848500
endClass        0.717256   0.852757 1.000000 0.725950
engScore        0.988024   0.848500 0.725950 1.000000
             engListening engReading endClass engScore
engListening    1.000000   0.618284 0.563621 0.963624
engReading      0.618284   1.000000 0.750568 0.679366
endClass        0.563621   0.750568 1.000000 0.584898
engScore        0.963624   0.679366 0.584898 1.000000
```

- 피어슨 상관분석 결과 engListening과 engScore의 상관계수가 0.995829로 가장 큰 선형 상관성을 보였으며, engListening과 engClass의 상관계수가 0.703028로 가장 작은 선형 상관성을 보였음.
- 스피어만 상관분석과 켄달 상관분석에서도 engListening과 engScore의 선형 상관성이 가장 크고 engListening과 engClass의 선형 상관성이 가장 작게 나타남.

01. 상관관계 분석의 개념

Ⅱ. 상관관계 분석의 세 가지 방법

- 스피어만 상관분석과 켄달 상관분석

✓ **하나 더 알기: 켄달 상관분석의 특징**

켄달 상관분석은 두 변수 간의 순위를 비교하여 연관성을 계산함.

한 변수가 증가할 때 다른 변수가 함께 증가하는 횟수와 감소하는 횟수를 측정하여 횟수의 차이를 상관계수로 표현하는 방법.

순위로 표현할 수 있는 데이터이거나, 표본 크기가 작거나, 데이터의 순위에 동률이 많을 때 활용.

02

상관관계 분석의 활용

02. 상관관계 분석의 활용

Ⅰ. 기준금리와 부동산 매매가격

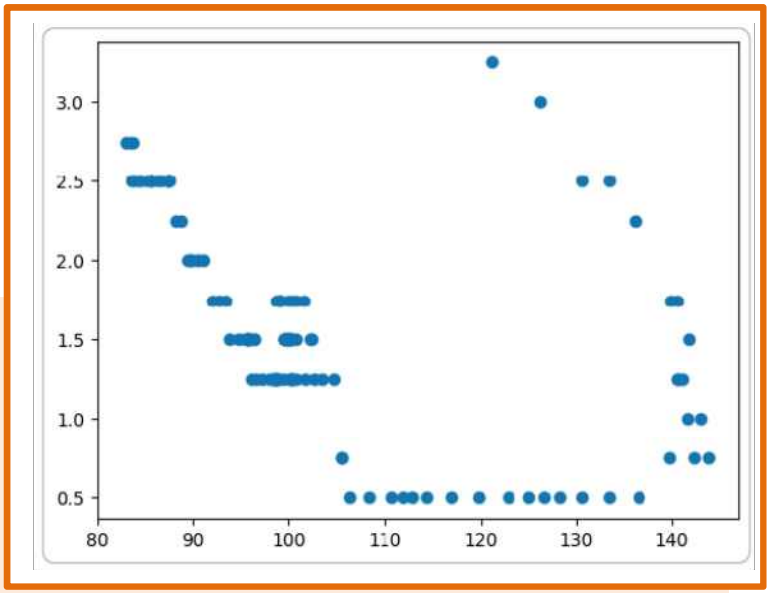
- 기준금리와 아파트 가격의 관계를 분석.

표 9-3 월별 부동산 지수와 기준금리 ©한국부동산원, 한국은행

연월(yymm)	부동산 지수	기준금리(%)
1301	83	2.75
1302	83	2.75
1303	83.5	2.75
1304	83.8	2.75
1305	83.9	2.5
…	…	…
2207	136.1	2.25
2208	133.4	2.5
2209	130.6	2.5
2210	126.2	3
2211	121.1	3.25

02. 상관관계 분석의 활용

Ⅰ. 기준금리와 부동산 매매가격



- 기준금리와 아파트 가격의 관계를 분석.

[코드 9-7] 데이터 준비

```
import pandas as pd
import matplotlib.pyplot as plt

realEstate = [83, 83, 83.5, 83.8, 83.9,
                  (중략), 136.1, 133.4, 130.6, 126.2, 121.1]
interestRate = [2.75, 2.75, 2.75, 2.75, 2.5,
                  (중략), 2.25, 2.5, 2.5, 3, 3.25]
data = {'부동산':realEstate, '금리':interestRate}

plt.scatter(df['부동산'], df['금리'])
plt.show()
```

- 데이터를 딕셔너리 data에 할당하고, 키는 '부동산'과 '금리'로 설정.
- '부동산'이 120 초과일 때는 선형 상관도가 낮을 것으로 예상.

02. 상관관계 분석의 활용

Ⅰ. 기준금리와 부동산 매매가격

- 기준금리와 아파트 가격의 관계를 분석.
  - 월별 부동산 실거래 매매가격 지수와 기준금리 전체 데이터의 상관관계를 피어슨 방식으로 분석.

[코드 9-8] 피어슨 상관분석

```
df = pd.DataFrame(data)
coef = df.corr(method='pearson')
print(coef)
```

```
              부동산        금리
부동산      1.000000  -0.497677
금리        -0.497677   1.000000
```

  - 두 변수의 상관계수 값은 -0.497677로, 절댓값이 0.5에 가까운 음의 선형 상관관계.
  - 기준금리가 오를수록 부동산 가격이 낮아지고, 기준금리가 내릴수록 부동산 가격이 높아진다는 것으로 해석.

02. 상관관계 분석의 활용

Ⅰ. 기준금리와 부동산 매매가격

- 기준금리와 아파트 가격의 관계를 분석.

[코드 9-9] 부동산 상승기 상관분석

```
originalData = {'부동산':realEstate, '금리':interestRate}

realEstateIndexList = []
interestList = []
lastIndex = -1

#부동산 지수가 143.8이 될 때까지만 리스트에 데이터 추가하기
for key, value in originalData.items():
   if key == '부동산':
      for i in range(0, len(value)):
         if value[i] == 143.8:
           break
         else:
           realEstateIndexList.append(value[i])
           lastIndex = i
   else:
      for i in range(0, lastIndex + 1):
         interestList.append(value[i])
```

02. 상관관계 분석의 활용

Ⅰ. 기준금리와 부동산 매매가격

- 기준금리와 아파트 가격의 관계를 분석

[코드 9-9] 부동산 상승기 상관분석

```
data = {'지수':realEstateIndexList, '금리':interestList}
df = pd.DataFrame(data)
coef = df.corr(method='pearson')
print(coef)
```

```
            지수        금리
지수    1.000000  -0.854603
금리   -0.854603   1.000000
```

– 피어슨 상관계수는 -0.854603으로 '지수'와 '금리' 두 변수는 강한 음의 선형 상관관계.

02. 상관관계 분석의 활용

Ⅱ. 영어 성적과 수학 성적

- 학생 10명의 영어 시험과 수학 시험 등수로 스피어만 상관분석을 수행.

표 9-4 영어 시험과 수학 시험 등수

학생	학생1	학생2	학생3	학생4	학생5	학생6	학생7	학생8	학생9	학생10
영어 시험 등수	4	2	1	3	10	8	9	7	6	5
수학 시험 등수	2	1	3	4	8	7	10	5	9	6

[코드 9-10] 영어와 수학 등수 상관분석

```
import pandas as pd

data = {'영어':[4, 2, 1, 3, 10, 8, 9, 7, 6, 5],
'수학':[2, 1, 3, 4, 8, 7, 10, 5, 9, 6]}
df = pd.DataFrame(data)
coef = df.corr(method='spearman')
print(coef)
```

```
          영어        수학
영어  1.000000  0.818182
수학  0.818182  1.000000
```

– 스피어만 상관계수 0.818182로 두 변수는 양의 상관관계.

실전분석 GDP 성장률과 인구수의 상관관계 분석

[문제]

인구가 많을수록 GDP 성장 잠재력이 높다고 합니다. 실제로 인구수가 세계 1위인 중국과 2위인 인도의 경제 성장률은 2000년 이후 매우 높은 수준을 유지하고 있습니다. 분석 대상을 G20 회원국으로 한정하여 GDP 성장률과 인구수의 상관분석을 수행하세요.

G20는 G7 국가인 미국, 일본, 영국, 프랑스, 독일, 이탈리아, 캐나다를 비롯하여 신흥경제 12개 국가, 그리고 유럽연합(EU)으로 총 20개 국가입니다. 이 국가들 의 GDP 성장률과 인구수 데이터는 [표 9-5]과 같습니다. 인구가 많은 나라가 GDP 성장률도 더 높은 것처럼 보입니다. 그렇다면 인구수와 GDP 성장률이 정말 비례하는지 분석해 보겠습니다.

실전분석 GDP 성장률과 인구수의 상관관계 분석

표 9-5 G20 국가의 GDP 성장률과 인구수 데이터

국가	GDP 성장률 (2022년, %)	인구수 (2021년, 백만 명)
미국	0.9	334
일본	0.6	125
영국	0.4	67.53
프랑스	0.5	67.65
독일	1.1	83.16
이탈리아	1.7	59.24
캐나다	3.9	38.44
대한민국	1.4	51.74
러시아	−3.7	146
중국	2.9	1,412
인도	6.3	1,380
인도네시아	5.01	273
아르헨티나	5.9	45.81
브라질	3.6	213
멕시코	3.5	126
호주	5.9	25.77
남아프리카공화국	4.1	60.14
사우디아라비아	5.4	34.11
터키	3.9	84.68
유럽연합(EU)	1.9	343

실전분석 GDP 성장률과 인구수의 상관관계 분석

[해결]

1. 데이터를 딕셔너리에 할당하고, 키는 국가, GDP 성장률, 인구수로 함.
   딕셔너리를 데이터프레임으로 변경하여 출력.

```
import pandas as pd

data = {'국가': ['미국', '일본', '영국', '프랑스', '독일', '이탈리아', '캐나다’,
              '대한민국', '러시아', '중국', '인도', '인도네시아', '아르헨티나’,
              '브라질', '멕시코', '호주', '남아프리카공화국', '사우디아라비아’,
              '튀르키예', '유럽연합(EU)’],
        'GDP 성장률': [0.9, 0.6, 0.4, 0.5, 1.1, 1.7, 3.9, 1.4, -3.7, 2.9, 6.3,
                    5.01, 5.9, 3.6, 3.5, 5.9, 4.1, 5.4, 3.9, 1.9],
        '인구수': [334, 125, 67.53, 67.65, 83.16, 59.24, 38.44, 51.74, 146, 1412,
                 1380, 273, 45.81, 213, 126, 25.77, 60.14, 34.11, 84.68, 343]}

df = pd.DataFrame(data)
print(df)
```

실전분석 GDP 성장률과 인구수의 상관관계 분석

[해결]

1. 데이터를 딕셔너리에 할당하고, 키는 국가, GDP 성장률, 인구수로 함.
   딕셔너리를 데이터프레임으로 변경하여 출력.

	국가	GDP 성장률	인구수
0	미국	0.90	334.00
1	일본	0.60	125.00
2	영국	0.40	67.53
3	프랑스	0.50	67.65
4	독일	1.10	83.16
5	이탈리아	1.70	59.24
6	캐나다	3.90	38.44
7	대한민국	1.40	51.74
8	러시아	-3.70	146.00
9	중국	2.90	1412.00
10	인도	6.30	1380.00
11	인도네시아	5.01	273.00
12	아르헨티나	5.90	45.81
13	브라질	3.60	213.00
14	멕시코	3.50	126.00
15	호주	5.90	25.77
16	남아프리카공화국	4.10	60.14
17	사우디아라비아	5.40	34.11
18	튀르키예	3.90	84.68
19	유럽연합(EU)	1.90	343.00

실전분석 GDP 성장률과 인구수의 상관관계 분석

[해결]

2. 데이터프레임 df로 피어슨, 스피어만, 켄달 상관분석을 모두 수행.

```
pearsonCoef = df.corr(method='pearson')
print("Pearson Correlation Analysis")
print(pearsonCoef)

spearmanCoef = df.corr(method='spearman')
print("\nSpearman Correlation Analysis")
print(spearmanCoef)

kendallCoef = df.corr(method='kendall')
print("\nKendall Correlation Analysis")
print(kendallCoef)
```

실전분석 GDP 성장률과 인구수의 상관관계 분석

[해결]

2. 데이터프레임 df로 피어슨, 스피어만, 켄달 상관분석을 모두 수행.

```
Pearson Correlation Analysis
            GDP 성장률      인구수
GDP 성장률    1.000000   0.198924
인구수        0.198924   1.000000

Spearman Correlation Analysis
            GDP 성장률      인구수
GDP 성장률    1.000000  -0.196388
인구수       -0.196388   1.000000

Kendall Correlation Analysis
            GDP 성장률      인구수
GDP 성장률    1.000000  -0.137568
인구수       -0.137568   1.000000
```

실전분석 GDP 성장률과 인구수의 상관관계 분석

[해결]

3. 세 가지 상관분석 방법으로 분석한 결과 상관계수의 절댓값이 모두 0.5에 훨씬 못 미침.
   따라서 GDP 상승률과 인구수는 선형 상관관계가 없음. 그러나이 분석에는 두 가지 한계가 있음.

   첫째는 전 세계에 코로나19라는 특수한 요인이 작용한 시기의 전년대비 GDP 성장률이라는 점.
   둘째는 표본 수가 적다는 점.

Thank You!

한빛아카데미 Hanbit Academy, Inc.