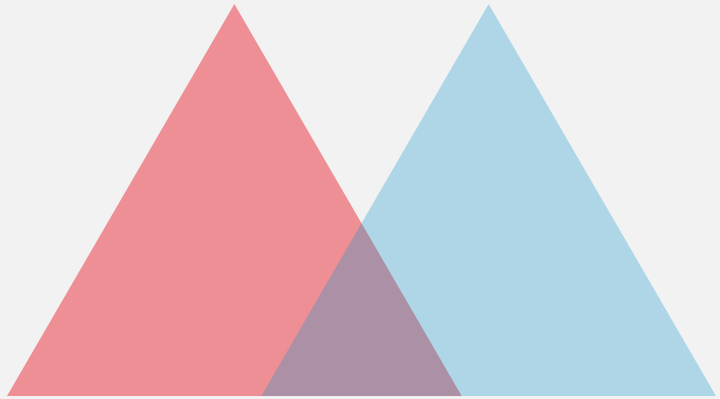
출산율과 요소관계 분석



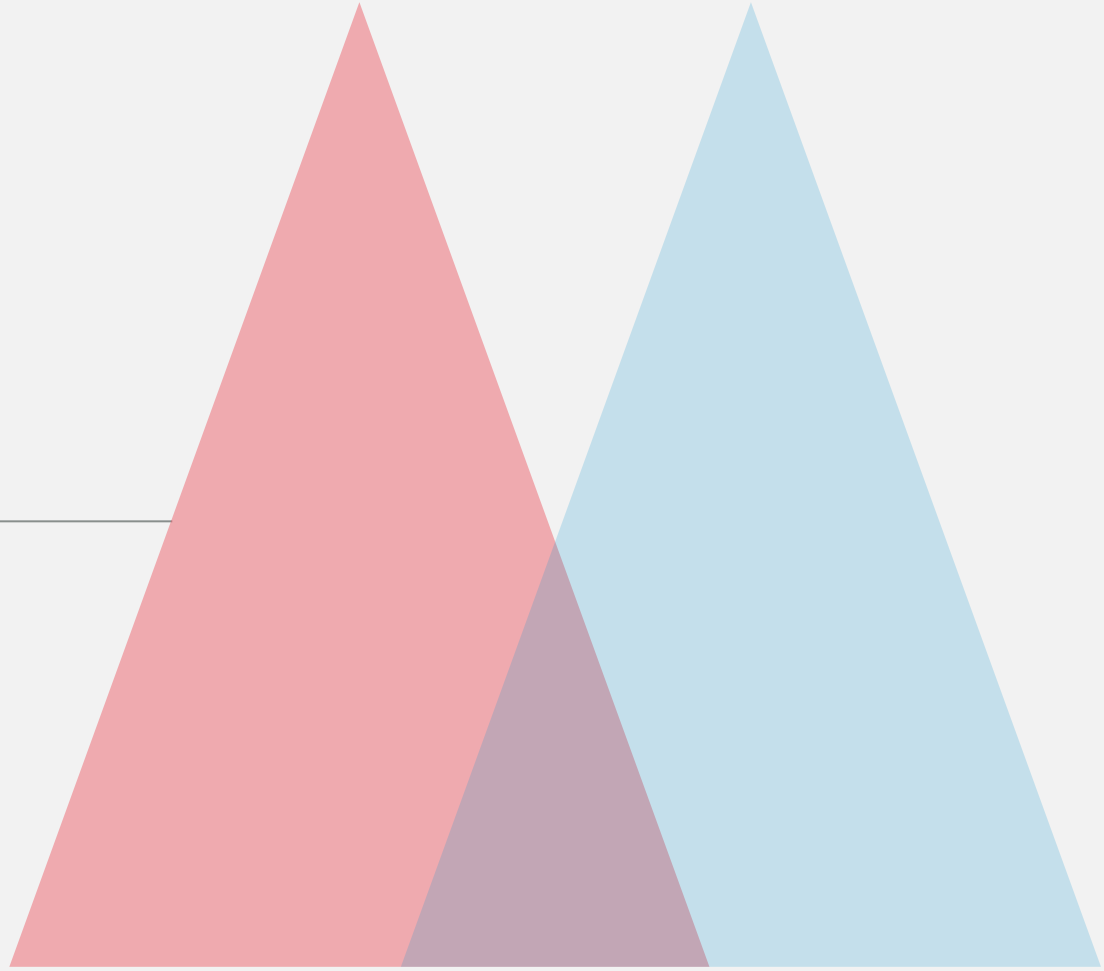
-출산율을 위주로 분석해본 인구증가의 요소

조장: 손가현
조원: 송현욱
배지용

Content

001

한국 출산율의 문제



지속적인 출산율 저하

한국의 출산율 저하 문제

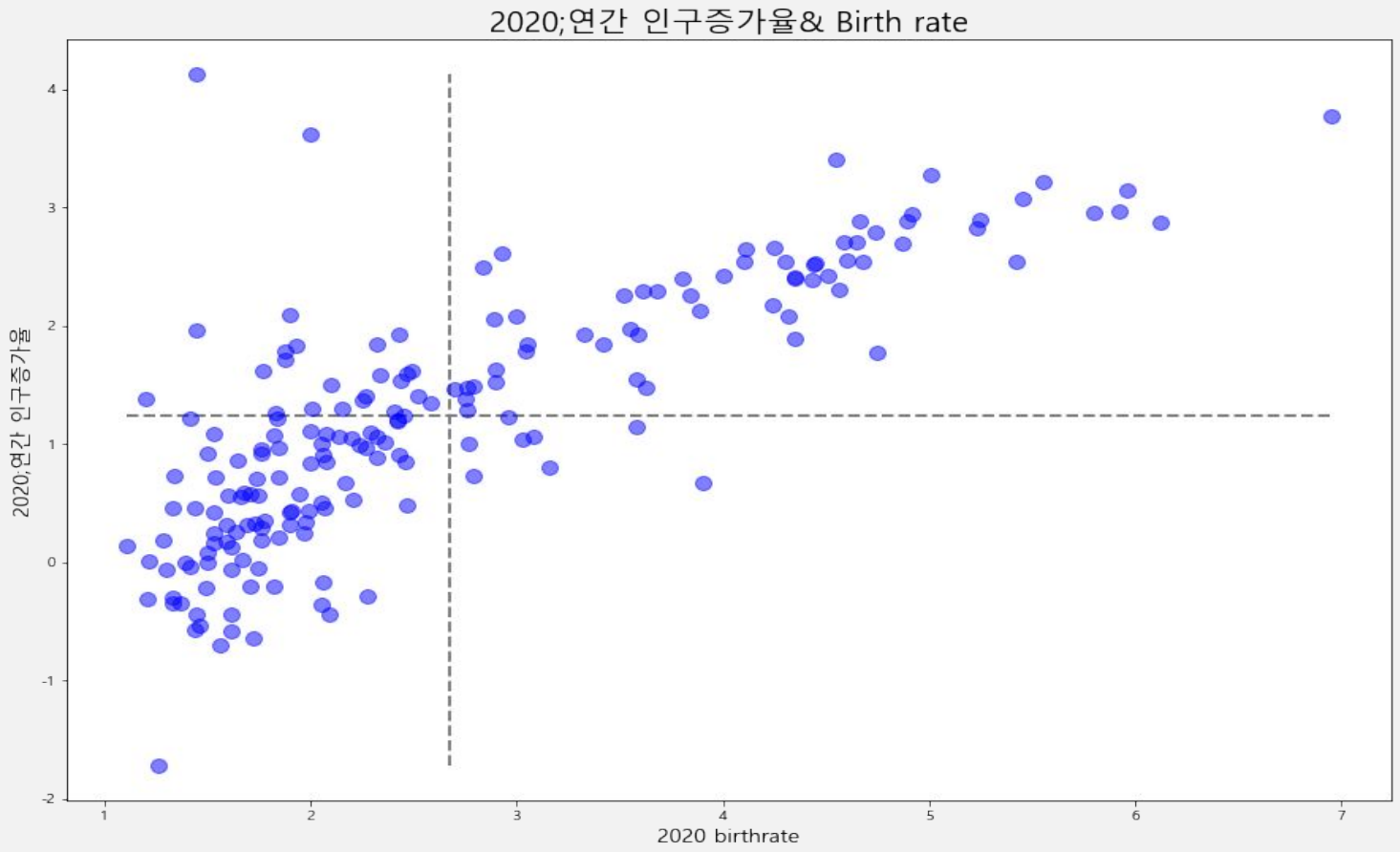
지난해 **우리나라 합계출산율이 0.84**명까지 떨어져 또다시 역대 최저를 기록했다. 출생아보다 사망자가 더 많은 인구 자연감소도 처음 시작됐다. ……, **저출생 심화는 고령화를 앞당겨 연금·의료비 등 복지지출 급증으로 이어지고, 반면 생산연령인구 감소로 경제성장 및 재정수입이 약화할 수 있다.** 정부는 2년 전 인구정책대응 태스크포스를 꾸려 인구감소에 대응하는 장기대책을 마련하겠다고 공언했지만, 뚜렷한 성과를 내지 못하고 있다. …..

-이경미 기자, kmlee@hani.co.kr, 한겨레, 2021.02.24, [“0.84명’ 출산율 세계최저 한국, 또 역대최저’]
https://www.segye.com/newsView/20210722515834

한국의 저출산고령화는 **첫째 실물시장에서의 수요와 공급 측면, 둘째 요소시장에서의 노동공급 측면, 셋째 금융시장에서의 저축률저하 및 이자율 측면, 총요소생산성 측면 등** 다양한 경로를 통해 잠재경제성장률에 부정적인 영향을 미치는 것으로 분석된다. 초저출산이 본격화된 시점에 태어난 어린이가 2016년 15세가 되면서 생산가능인구가 이미 줄기 시작했고…

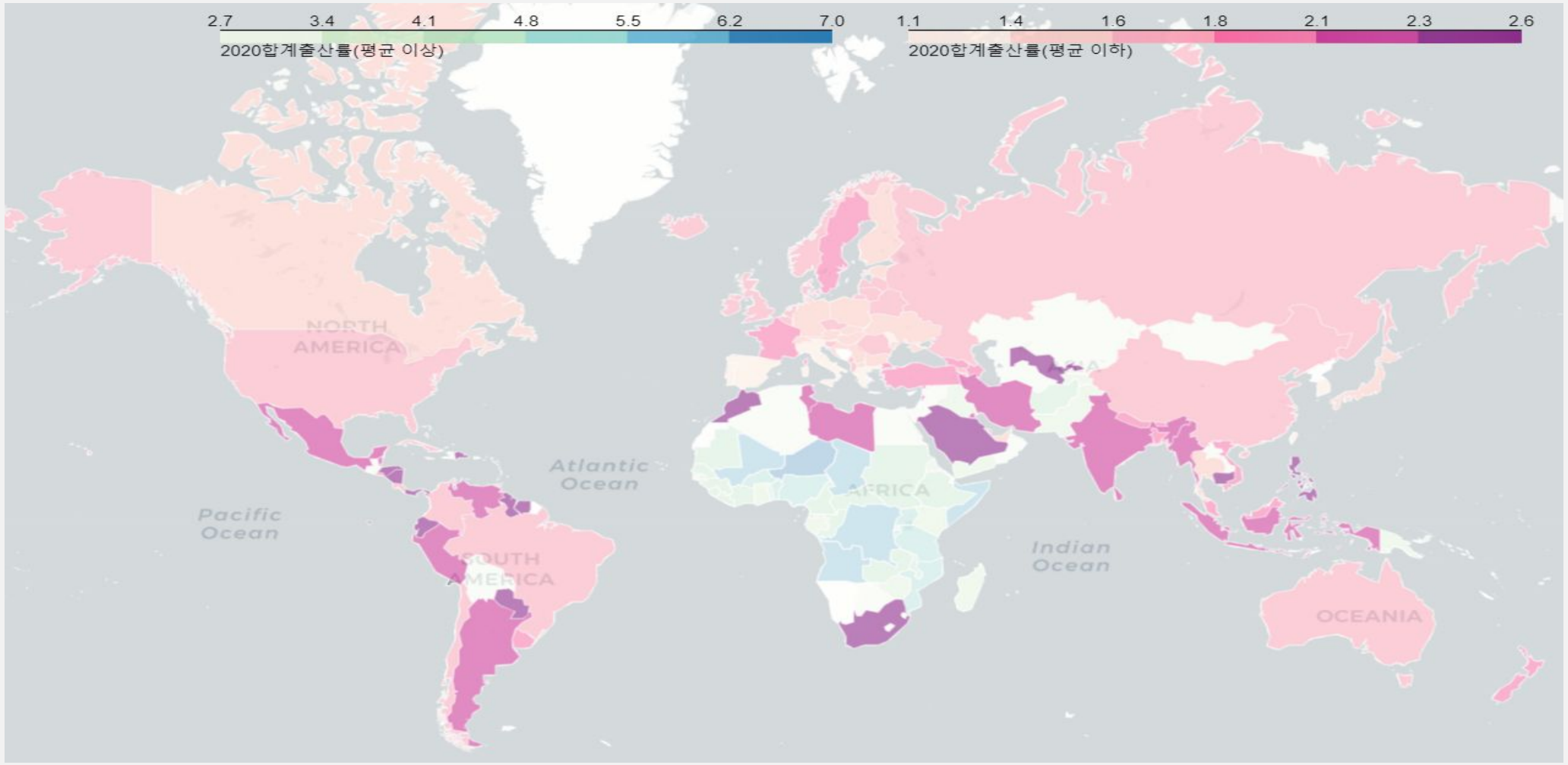
저출산.고령화가 한국경제에 미치는 영향분석 : 잠재경제성장률을 중심으로
https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002297210

출산율이 왜 중요할까 ?



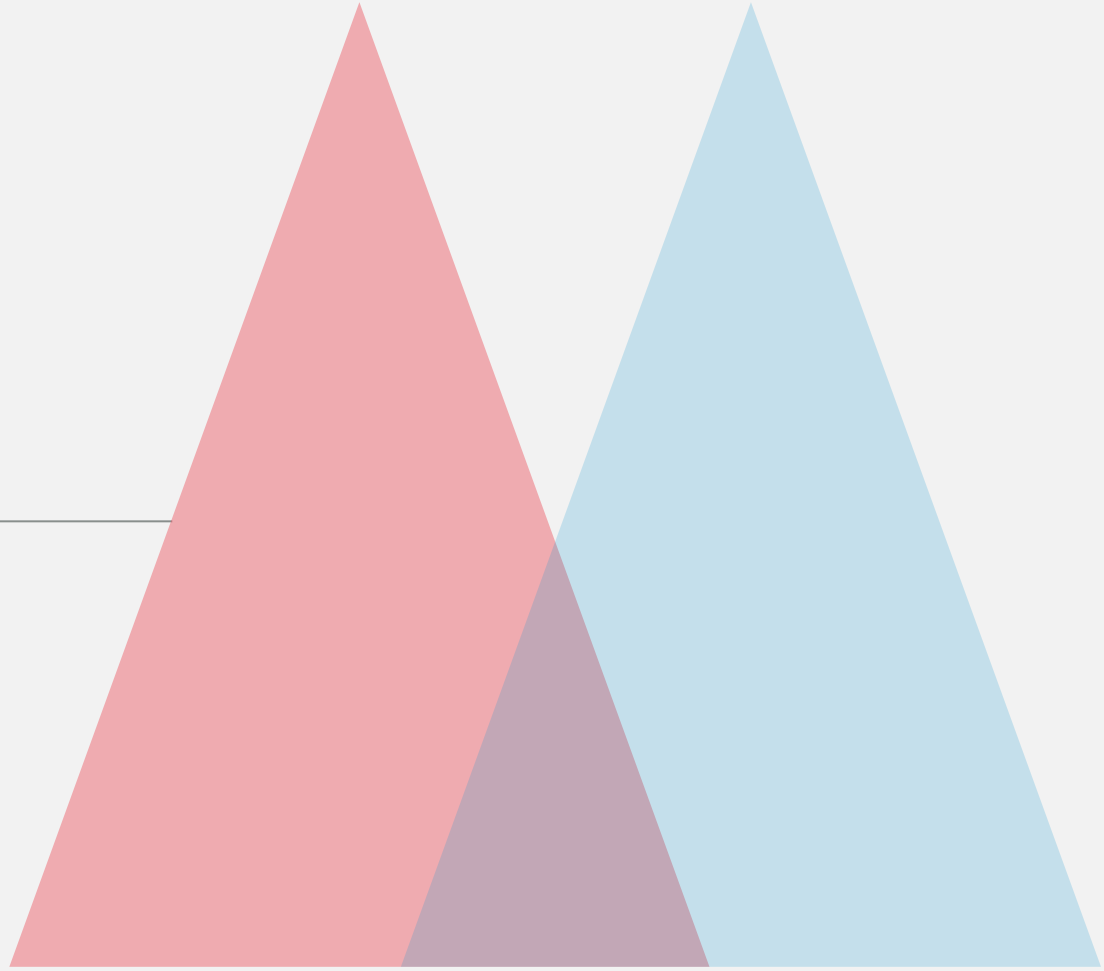
- 출산율은 인구증감에서 큰 부분을 차지한다.

전세계 출산율



002

출산율과 요소들의 관계분석



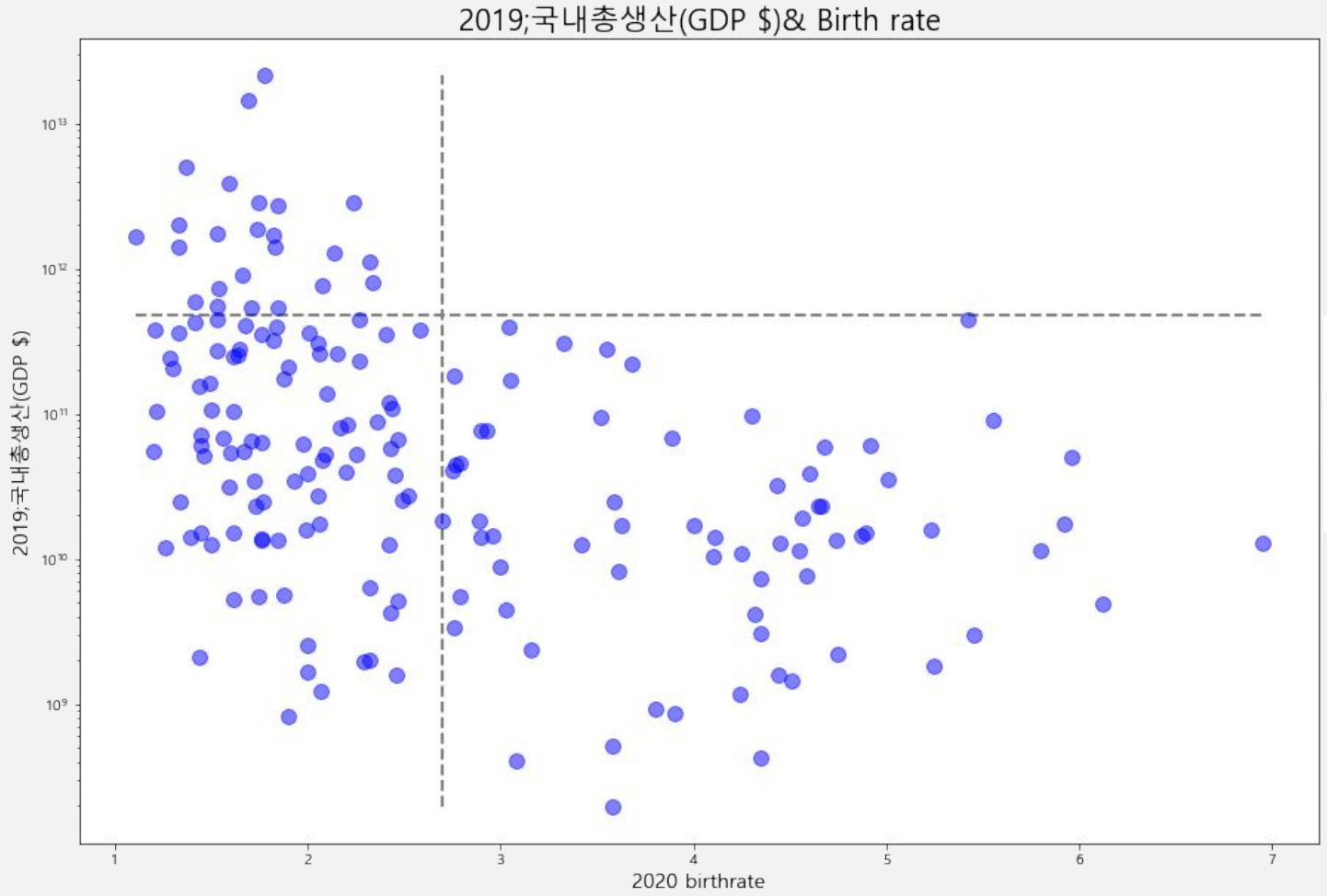
3

001 소제목

출산율과 요소 상관분석

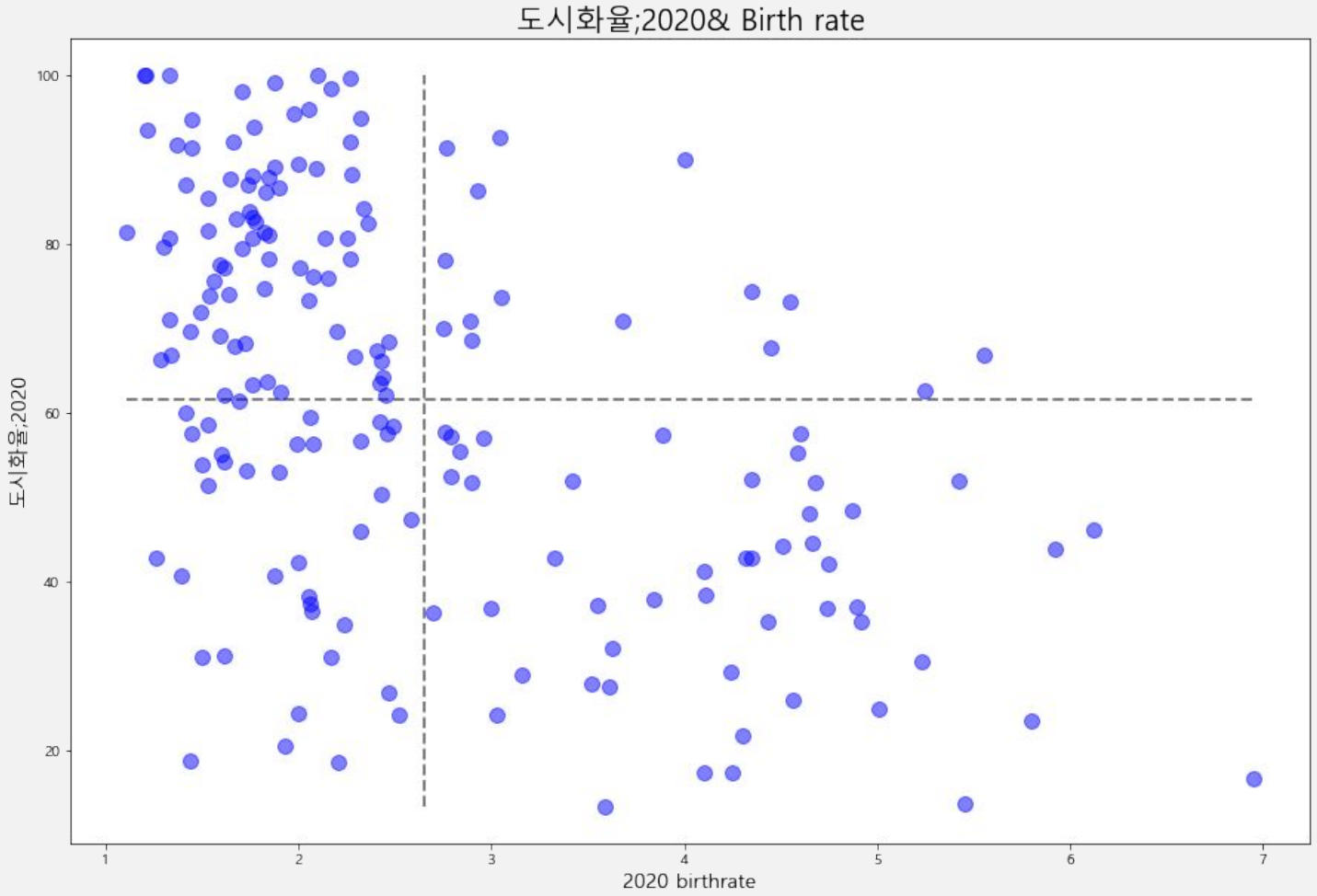
	GDP	도시화율	평균 교육기간	1인당 GNI	경제 활동 참여율(여성)	소비자 물가지수
공분산	-4.311268e+11	-14.760446	-2.978451	-1.376205e+04	1.308887	28.615940
상관계수	-0.166606	-0.509245	-0.759941	-0.572801	0.082832	0.188945

출산율과 GDP의 상관관계



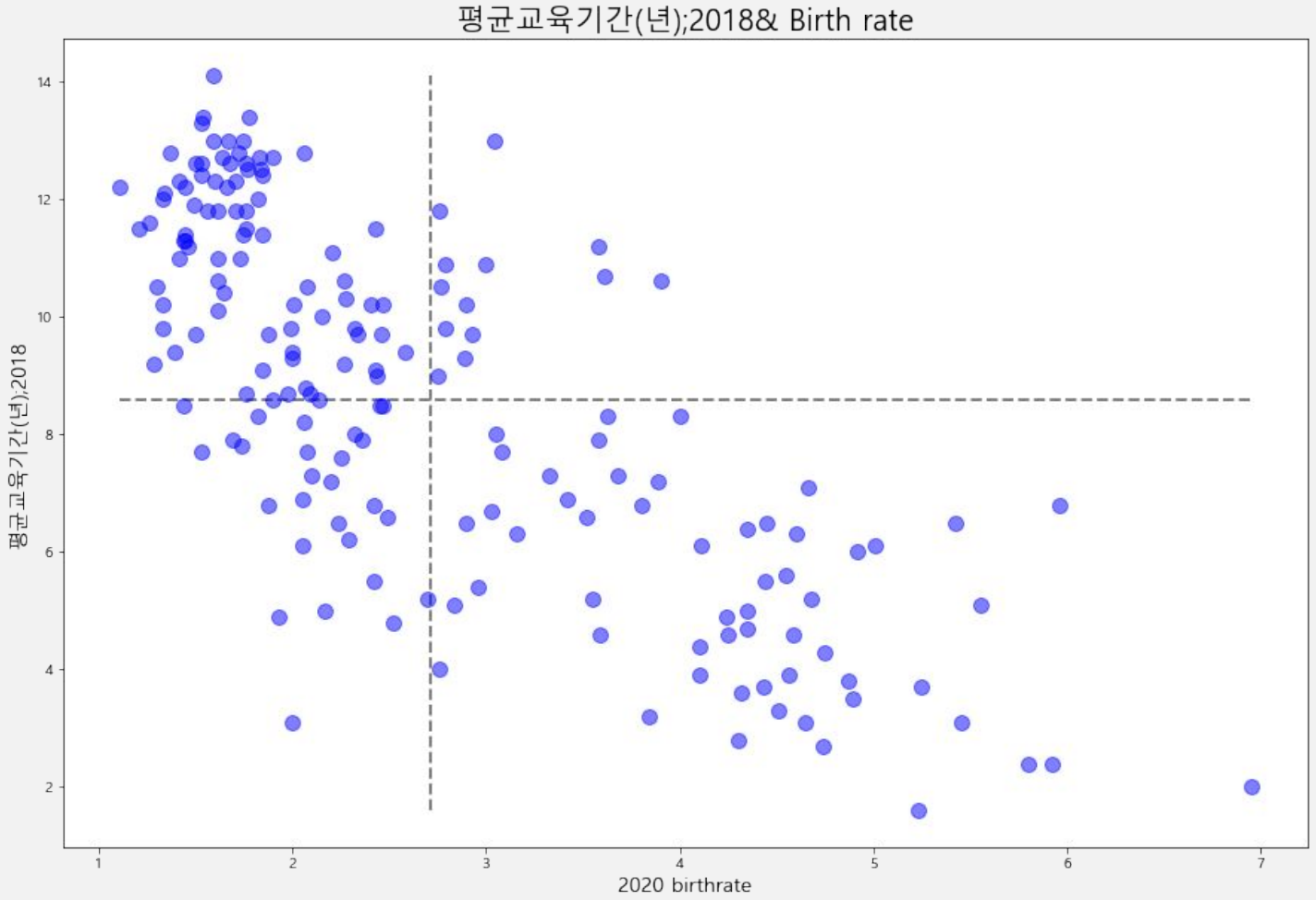
	2019;국내총생산(GDP $)	합계출산;2020
2019;국내총생산(GDP $)	1.000000	-0.166184
합계출산;2020	-0.166184	1.000000

출산율과 도시화율의 상관관계



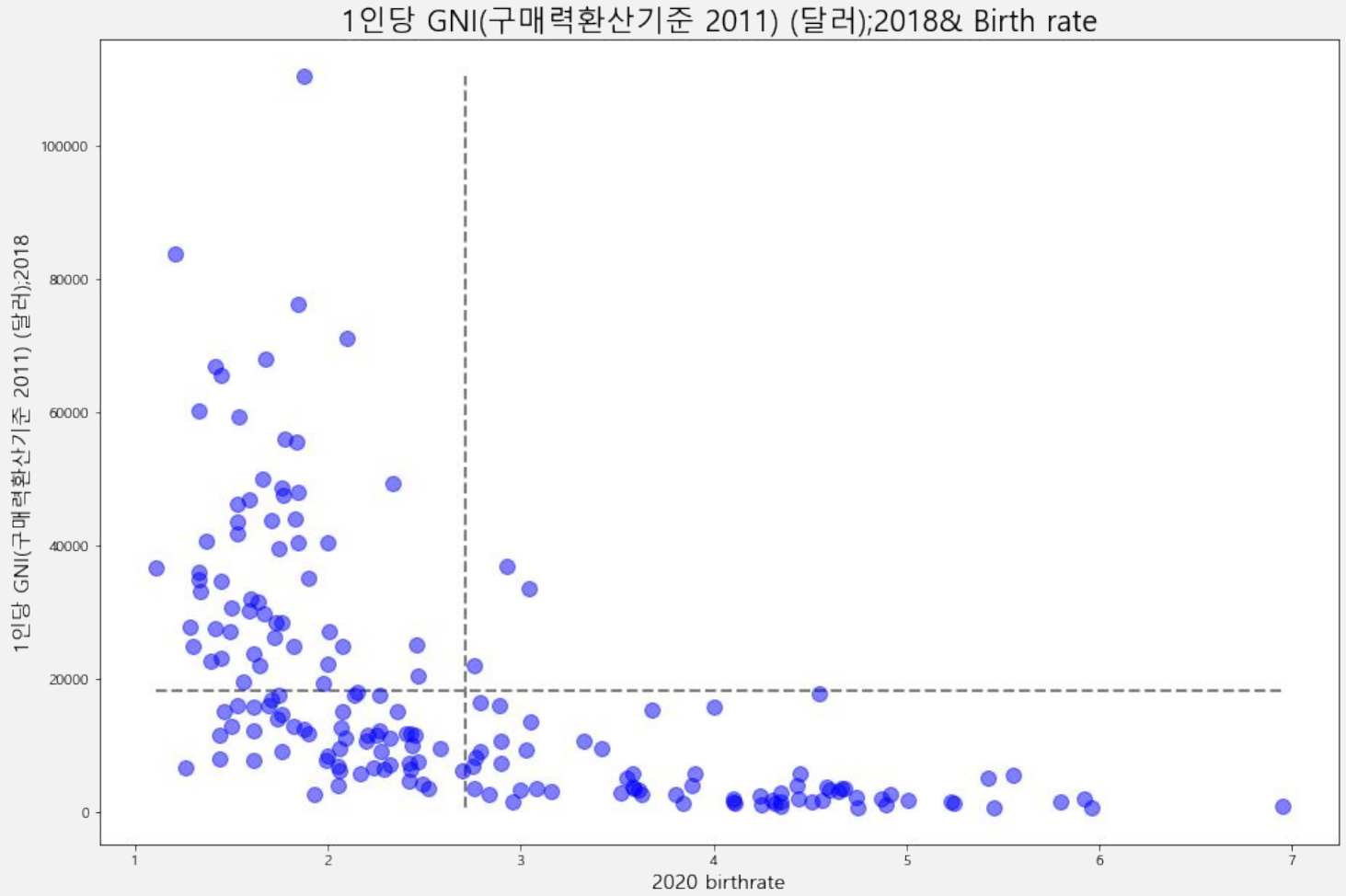
	도시화율;2020	합계출산;2020
도시화율;2020	1.000000	-0.508729
합계출산;2020	-0.508729	1.000000

출산율과 평균교육기간의 상관관계



	평균교육기간(년);2018	합계출산;2020
평균교육기간(년);2018	1.00000	-0.75965
합계출산;2020	-0.75965	1.00000

출산율과 1인당 GNI의 상관관계

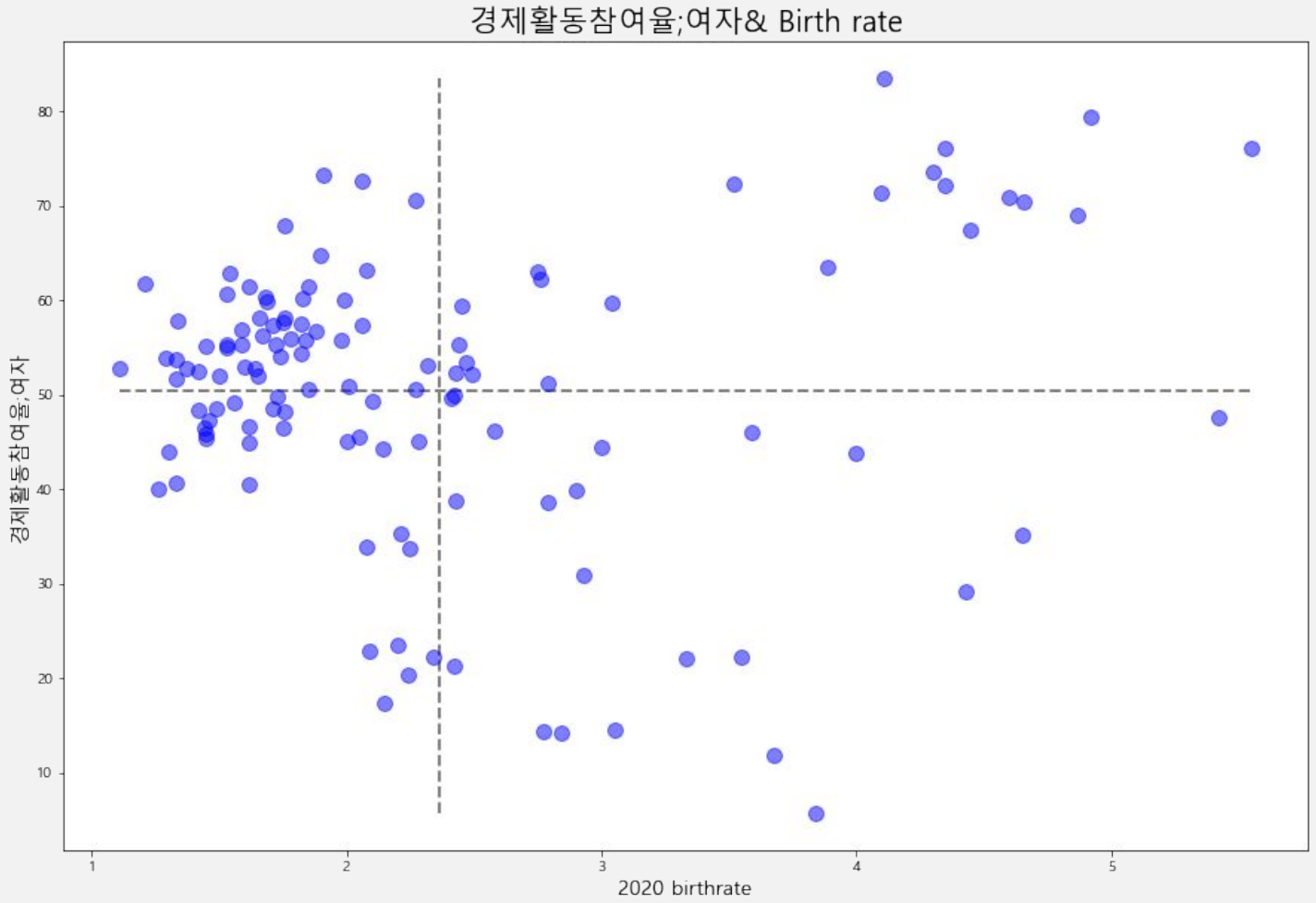


	1인당 GNI(구매력환산기준 2011) (달러);2018	합계출산;2020
1인당 GNI(구매력환산기준 2011) (달러);2018	1.000000	-0.572353
합계출산;2020	-0.572353	1.000000

2

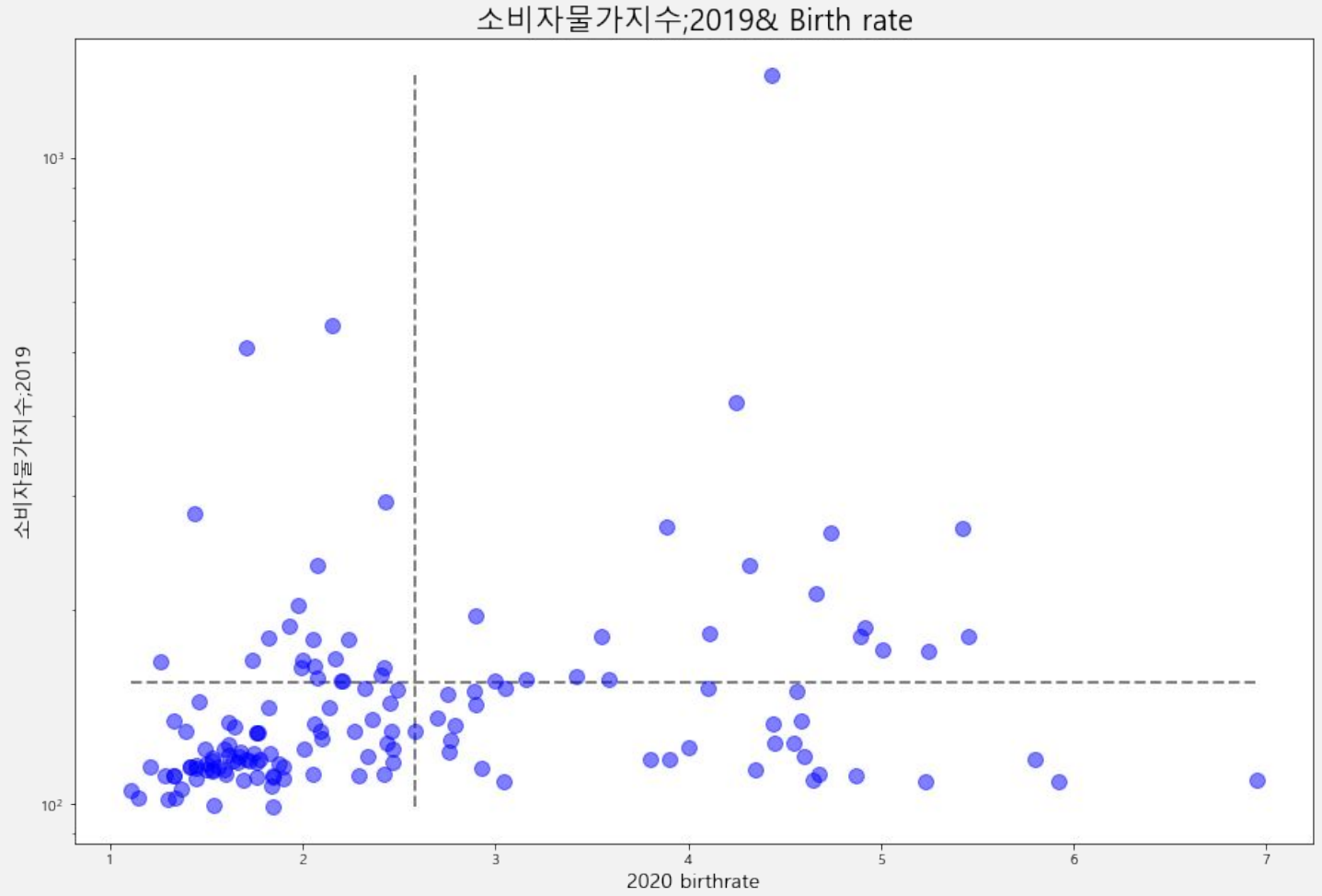
출산율과 세계

출산율과 경제활동 참여율(여성)의 상관관계



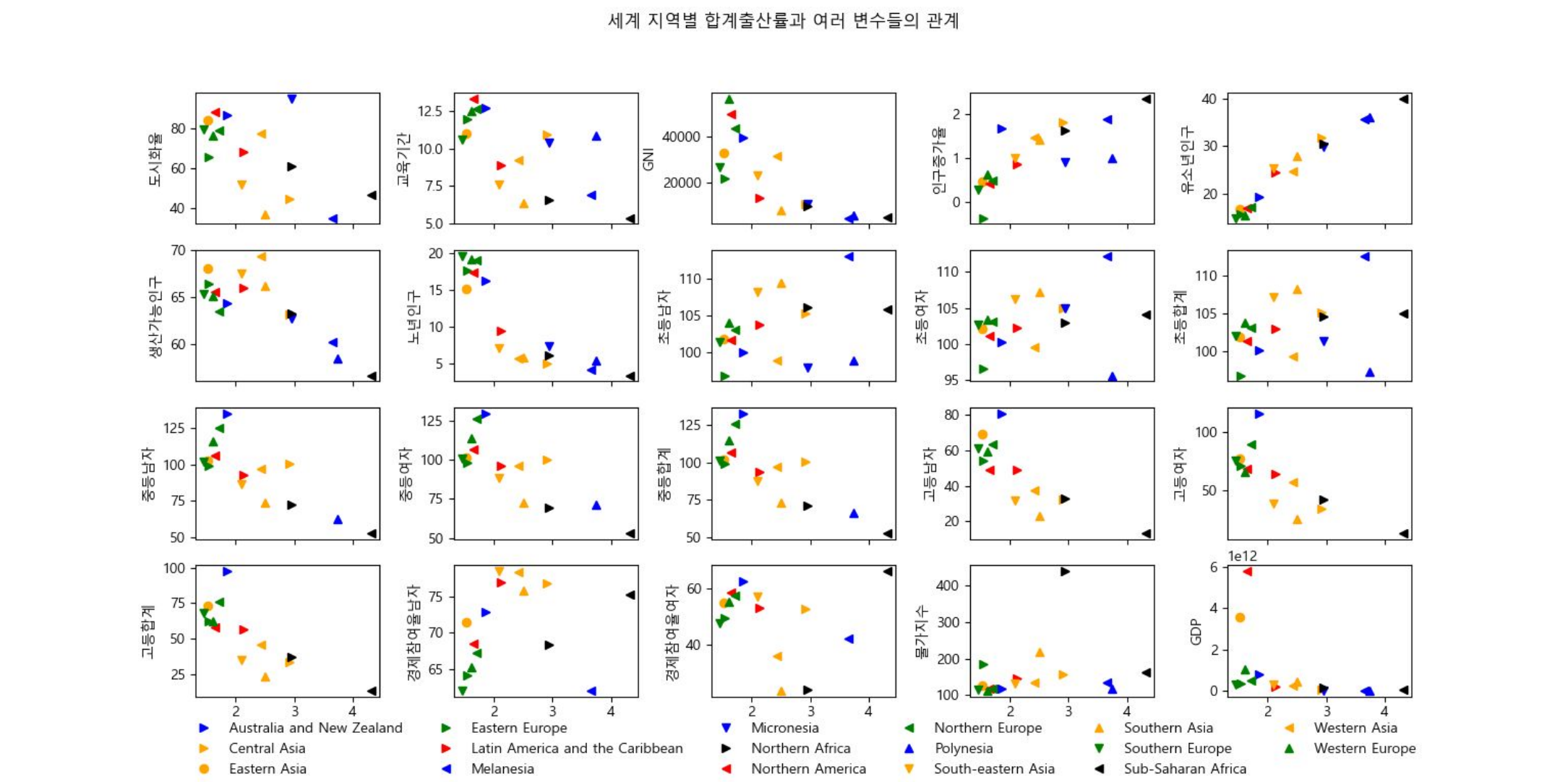
	경제활동참여율;여자	합계출산;2020
경제활동참여율;여자	1.000000	0.083403
합계출산;2020	0.083403	1.000000

출산율과 소비자 물가지수의 상관관계



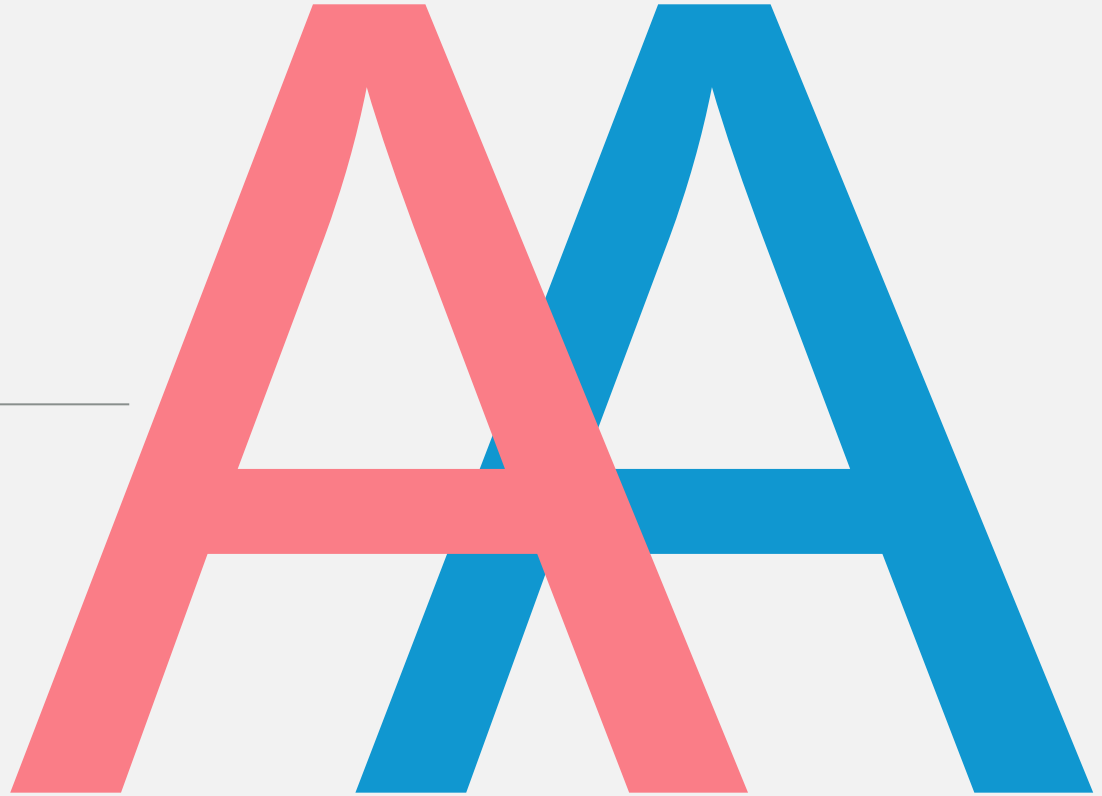
	2020	소비자물가지수
2020	1.000000	0.188674
소비자물가지수	0.188674	1.000000

대륙별 출산율과 여러 변수들의 관계



003

데이터 처리



데이터 처리 과정

출산율에 영향을 끼치는 여러요소를 찾아보자

요소들 : 각 나라별 1인당 소득, 교육기간, 여자의 경제활동 참여율, 도시화율, 지역정보

자료를 수집하고 전처리한다.

요소들과 출산율을 비교할 수 있는 모델을 생성한다.

(지도와 그래프)

출산율과 요소들의 자료 전처리

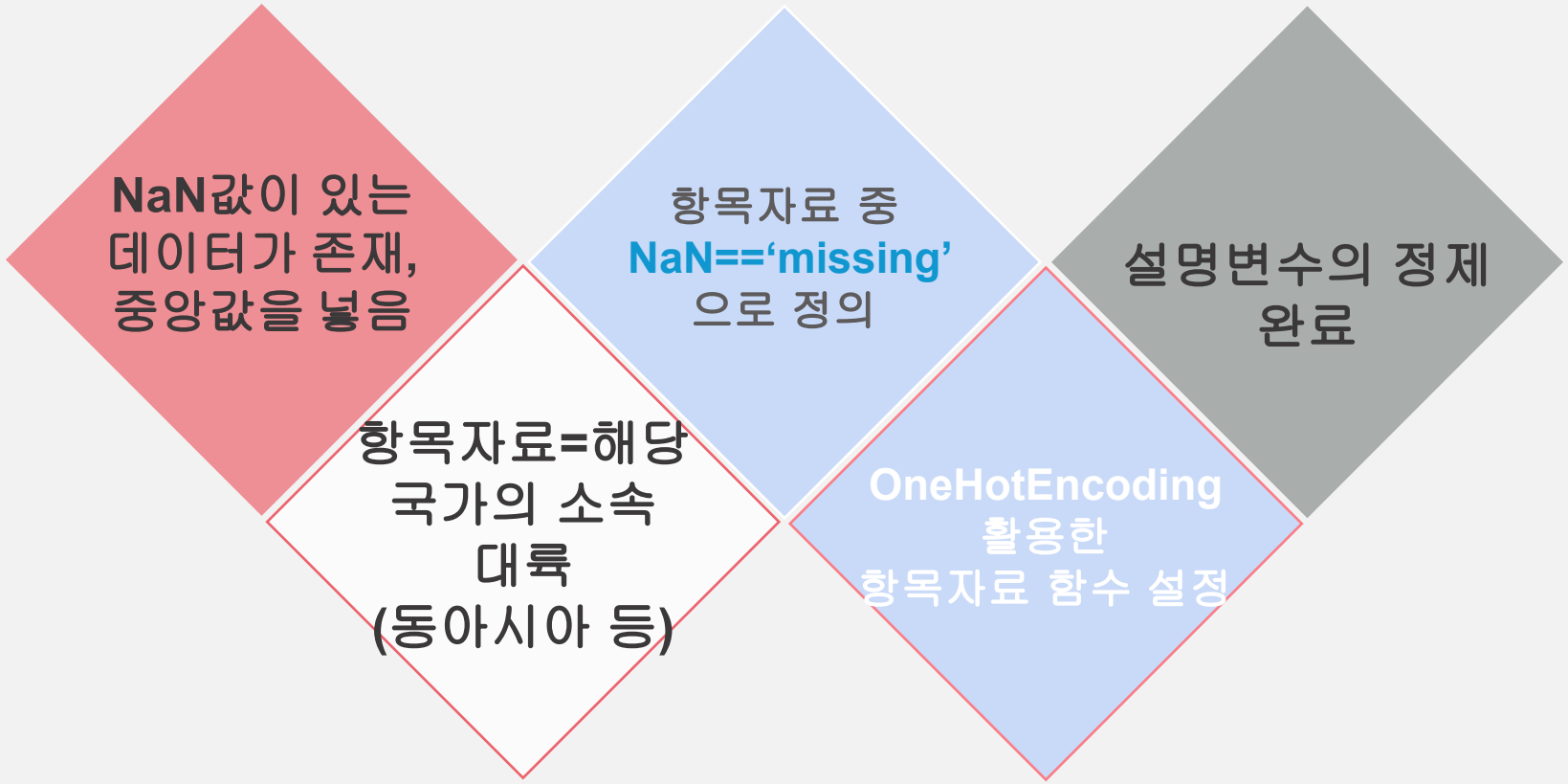
자료 전처리

- 여러개의 각 나라별 데이터를 하나의 테이블로 모아 출산율 중심으로 데이터값으로 연결해준다.
- 각 데이터마다 다른 언어로 저장된 국가 이름을 한글로 변경하였다.
- **merge**할 때 한글 나라 이름이 통일되지 않은 경우(오스트레일리아-호주, 한국-대한민국-남한 등) 하나로 통일
- 데이터 중 '-' 과 같은 알수없는 값은 **NaN**값으로 변경 및 정규표현식 사용(수치화 과정)
- 출산율과 설명변수를 수치화(**float**)하여 지도나 그래프 표현시 사용할 수 있도록 한다.
- 년도 데이터 유무에 따라 **2020**년 혹은 **2015**년 자료끼리 비교
- 교육과정 진학률의 경우 **index**에 있던 성별자료를 **column**으로 만들어서 합쳐주었다.

3

003 데이터 처리

NaN값을 예측하고 대륙별로 나누기 위한 회귀분석 모델 전처리과정



NaN값을 구하기 위한 회귀분석 모델

문제 설정

explanatory_variables(요소들) =
['1인당GNI(구매력환산기준2011)(달러);2018',
'평균교육기간(년);2018',
'경제활동참여율;여자',
'도시화율;2020','sub-region']

X=
country_data_with_answer[**explanatory_variables**]
y=
country_data_with_answer['합계출산;2015']

X와 Y를 계산해야하는데 NaN값이 있는 부분이 존재한다.

전처리 과정

```
# 각 지역을 하나의 설명변수로 설정

(대륙별로 전처리)
preprocess_cat = Pipeline(steps = [
    ('imputer', SimpleImputer(
# 어느 곳에 속한 지 모르면 missing으로 지역이름 설정
     strategy = 'constant', fill_value = 'missing')),
    ('encoder', OneHotEncoder(
     (encode)
 handle_unknown = 'ignore')),
# when encountering unknown, encode all zero
    ])
(요소별 전처리)
preprocess_num = Pipeline(steps = [
          ('imputer', SimpleImputer(strategy = 'median')), # NaN를 중앙값으로 변경
    ])
(위 두가지를 합친 column 정의)
preprocess = ColumnTransformer(
    transformers = [('numerizer',
                                    preprocess_num,
explanatory_variables[:-1]),
                    ('categorizer', preprocess_cat,
explanatory_variables[-1:]),
        ])
```

모델 생성

```
(합친 column을 모델에 적용, fit 메소드 사용)
reg =
   Pipeline(steps = [('preprocessor', preprocess),
                     ('regressor', LinearRegression()),
                     ])
reg.fit(X, y) # 선형 회귀분석
print('RMSE:              '+str(mean_squared_error(y,
reg.predict(X)))) # 평균오차 산정
print(
    pd.Series(
        reg.named_steps['regressor'].coef_,
        index = explanatory_variables[:-1] + \

list(reg.named_steps['preprocessor'].named_transformers_['categorizer'].\
            named_steps['encoder'].categories_[0]),
        name = '각 설명변수의 회귀계수')
    )
```

회귀식

(사용한 식)
y = a1x1+a2x2+a3x3+b 꼴
y는 출산율, x는 각 설명변수, a는 x가 y에 미치는 영향(회귀계수), b는 모든 x가 0일 때 y값

(결과)
(오차값)
RMSE: 0.41

```
1인당 GNI((달러)   -0.000006
평균교육기간(년)              -0.204814
경제활동참여율;여자(%)        -0.002334
도시화율(%)              -0.003232
Eastern Asia                -0.758318
Northern America             0.235157
...
Sub-Saharan Africa           1.135146
Western Europe              -0.049007
missing                     -0.603728
Name: 각 설명변수의 회귀계수
```

3

001 소제목

회귀분석을 이용한 평균 결과

1인당 GNI가 **1**만달러 늘 때마다 출산율은 **0.06**씩 감소

교육기간이 **1**년 늘 때마다 출산율이 **0.2**씩 감소

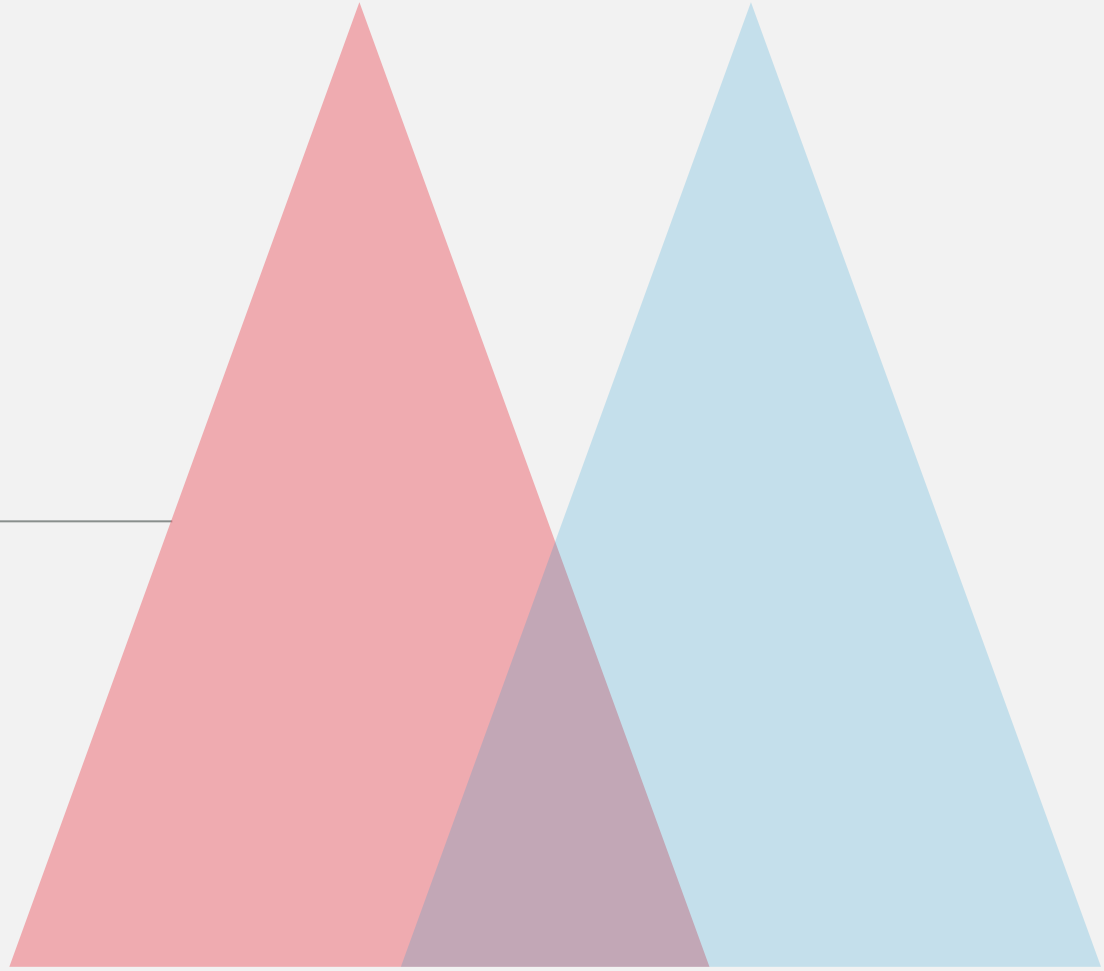
여성의 경제참여율이 **10%** 늘 때마다 출산율은 **0.023**씩 감소

도시화율이 **10%** 늘 때마다 출산율은 **0.032**씩 감소

동아시아에 속하면 출산율이 **0.7**만큼 감소하고, 사하라 사막 이남 아프리카에 속하면 출산율이 **1.1**씩 증가

004

분석결과



분석결과

소비자 물가지수

경제 활동 참여율
(여성)

1인당 GNI

GDP

평생교육기간

도시화율

감사합니다

조장: 손가현
조원: 송현욱
배지용