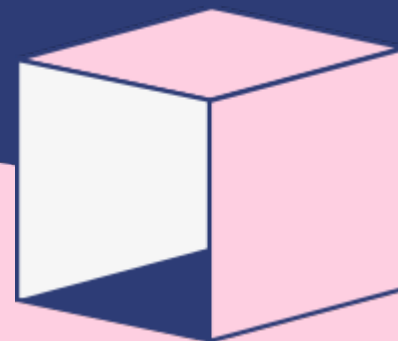
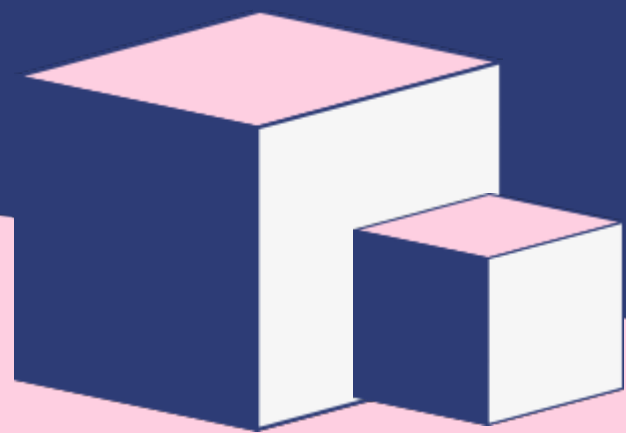


# 기대수명 예측



AIB 10기 박민경

# 목차

1. 주제 및 데이터 설명
2. 가설 설정
3. 데이터 전처리 및 EDA
4. 가설 검증
5. 모델 학습 및 결과
6. 결론 및 회고

# 주제 및 데이터 설명

▶ 주제 선정: 전세계적으로 의학/기술의 발전으로 인해 고령화 현상(=기대수명의 증가)이 나타남

-> 기대수명에 어떤 요소들이 영향을 주는지 분석

▶ 데이터 설명: <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?datasetId=12603&sortBy=voteCount>

총 22개의 컬럼: 국가, 연도, 선진국/개발도상국 여부, 기대 수명, 평균 체지방 지수, 성인 사망률, 유아 사망률, GDP 대비 건강 관련 지출, 알코올 소비량, 1세 인구 대비 b형 간염 예방 접종자 수, 홍역 환자 수, 5세 미만 아동 사망률, 1세 인구 대비 소아마비 예방 접종자 수, 정부 지출 대비 보건 관련 지출, 1세 인구 대비 급성전염병 예방 접종자 수, 5세 미만 아동 hiv/aids 사망건수, GDP, 국가 총인구, 5~9세 및 10~19세 아동인구 대비 저체중 아동의 비율, HDI(각국의 인간 발전 정도와 선진화 정도를 평가한 지수), 교육수준

# 가설 설정

1. 아동 관련 보건 지표가 향상될수록 기대수명은 증가할 것이다.
2. 국가의 선진화 정도가 높아질수록 기대수명이 증가할 것이다.
3. 건강을 위한 지출이 증가할수록 기대수명은 늘어날 것이다.

-> OECD 분석에 따르면 기대수명의 증가의 주요 결정요인이 **'의료비 지출증가'**

# 데이터 전처리 및 EDA

📎 데이터셋 컬럼명 및 국가명 표기 재설정

📎 case 수로 표기된 컬럼들을 퍼센트로 단위를 변경

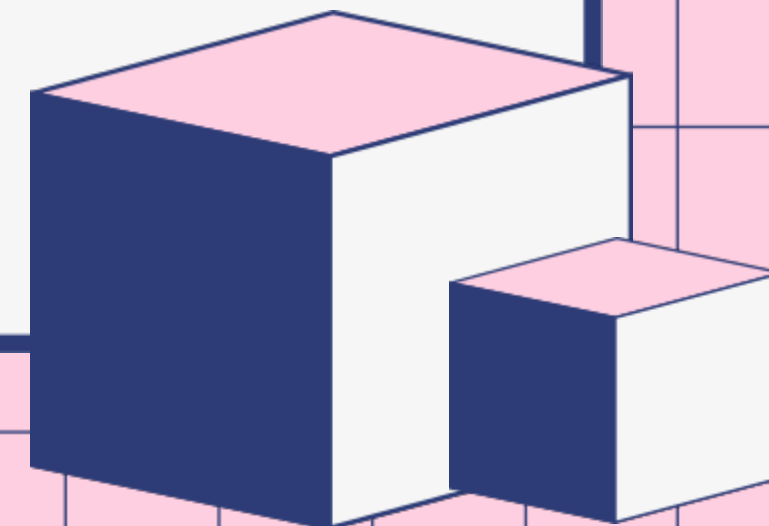
📎 결측치 처리: 세 가지 방법 시도

1. 국가별로 결측치를 보간하여 처리

2. 연도별로 결측치를 보간하여 처리

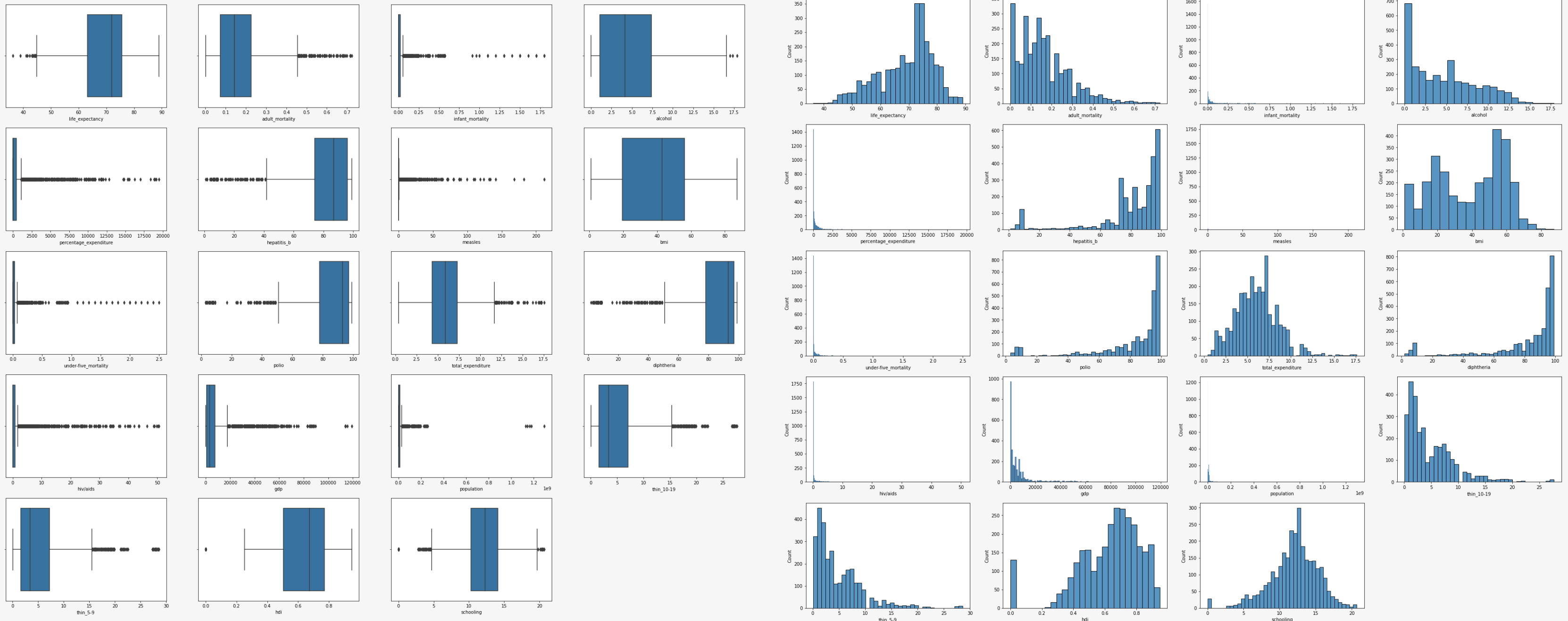
3. 결측치를 제거하여 처리

-> 시계열 데이터인 점을 고려 + 데이터 손실이 가장 적은 방법 선택



# 데이터 전처리 및 EDA

 boxplot과 히스토그램을 통해 이상치를 감지



# 데이터 전처리 및 EDA

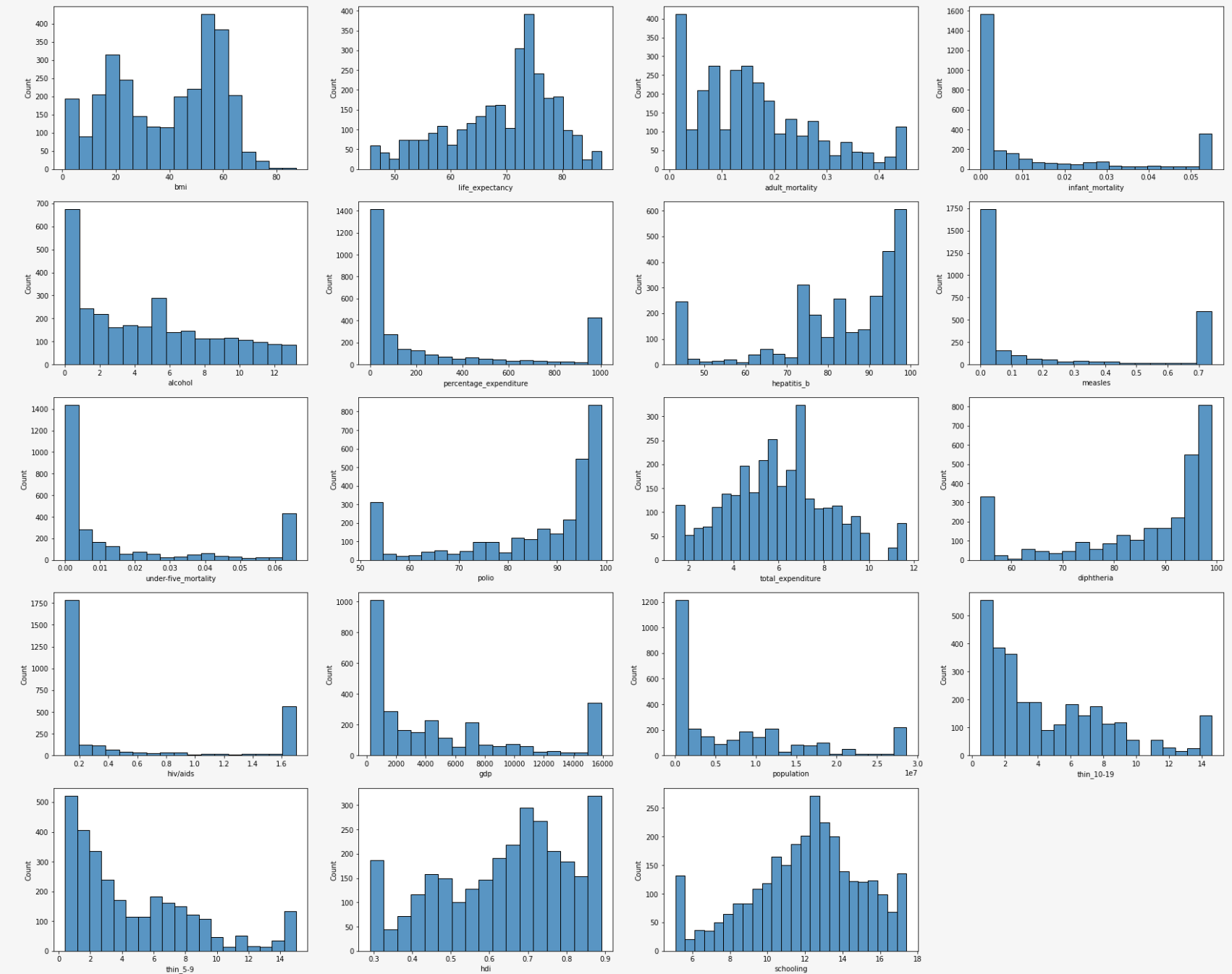
📎 이상치 처리 방법: 1) 이상치 제거 2) winsorization

-> 이상치 제거는 데이터 손실을 발생시키기 때문에 winsorization로 처리

📎 winsorization: 극단 값을 특정 값으로 대체하는 방법

ex) 상위 99%를 넘는 데이터는 상위 99%의 값으로 대체

# ▶ winsorization을 시도한 이후





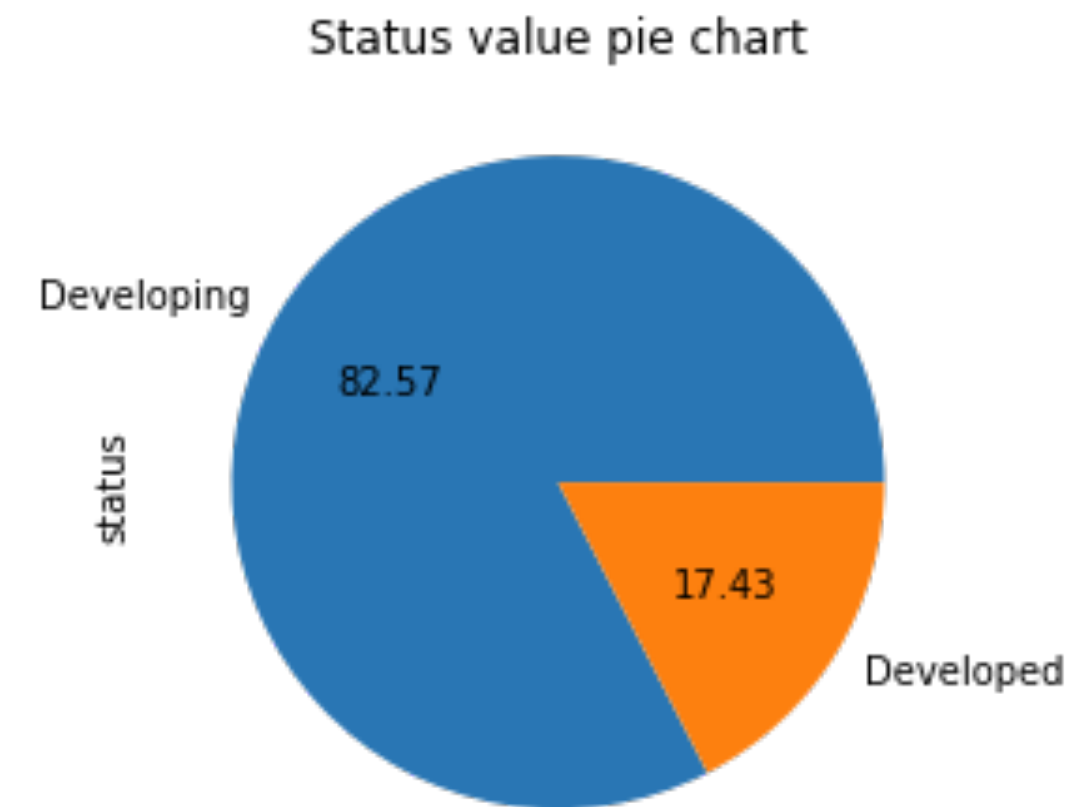
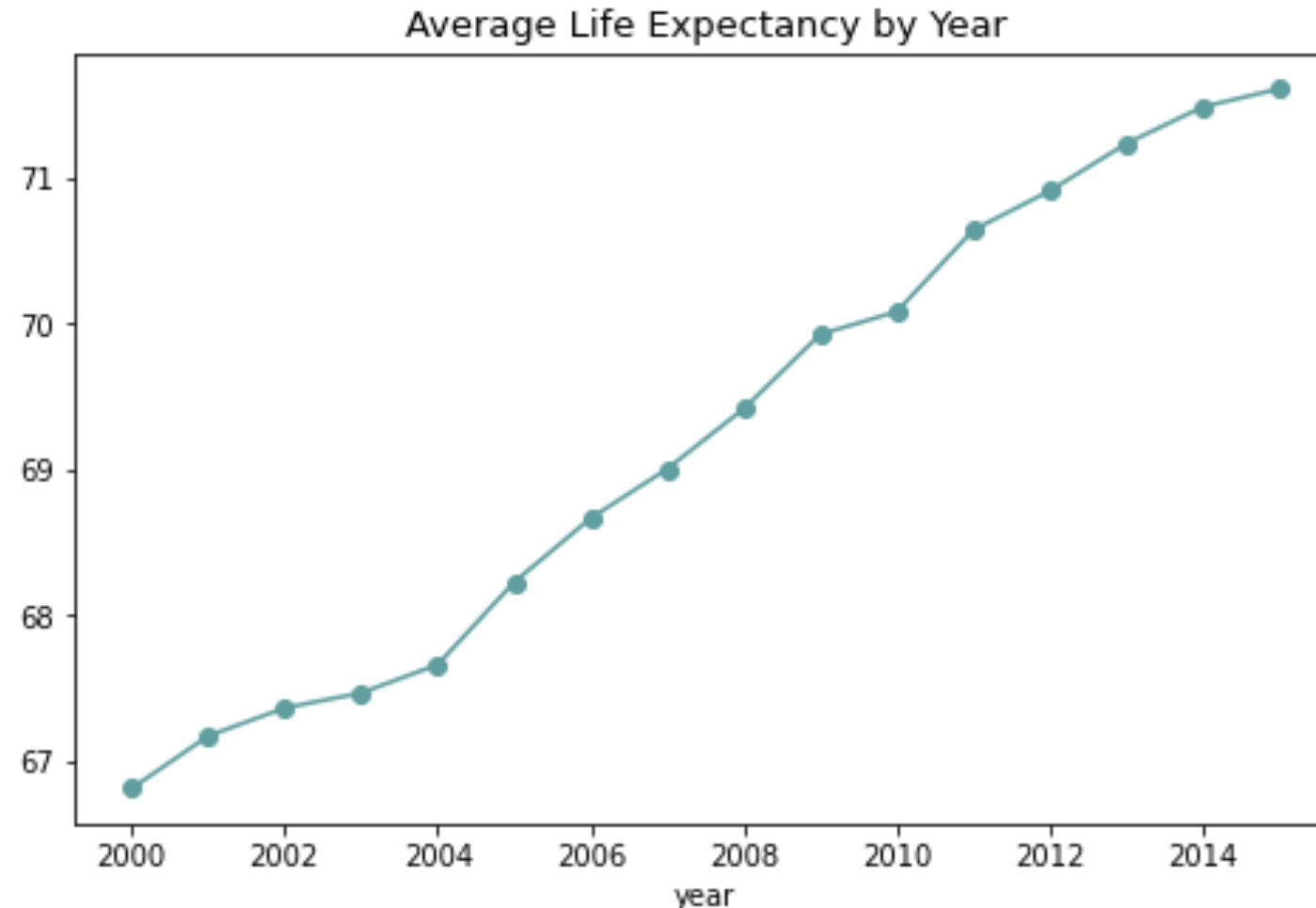
# 데이터 전처리 및 EDA

📎 카테고리형 변수

▶ 국가: 한국, 미국 등을 포함한 총 193개의 국가로 구성

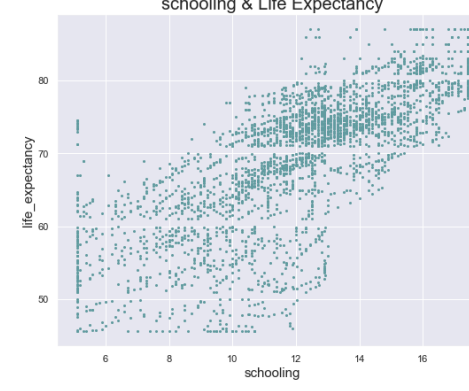
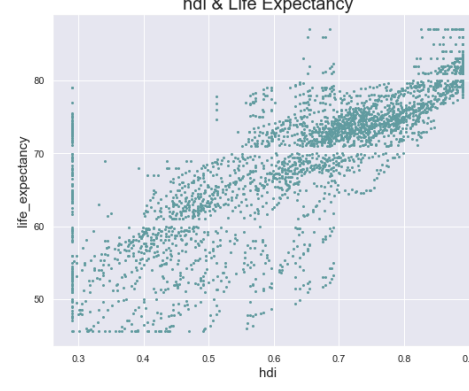
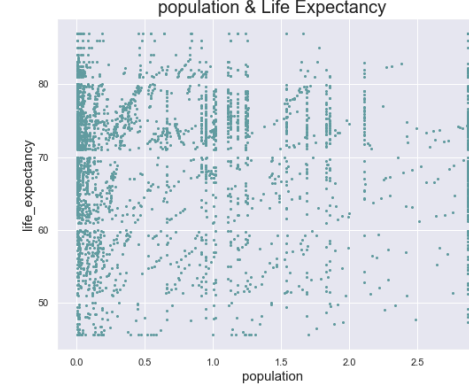
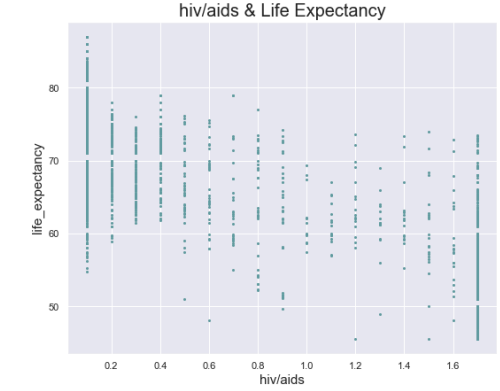
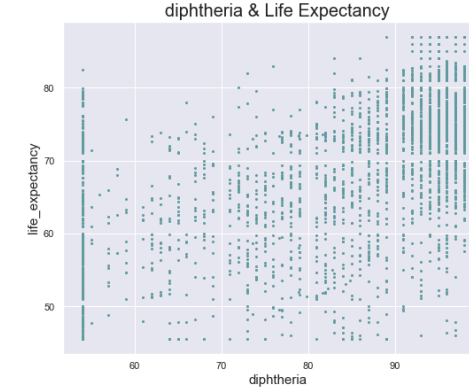
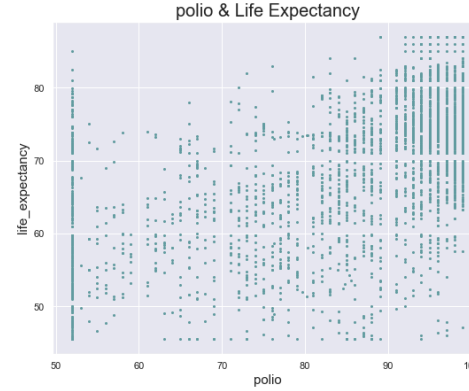
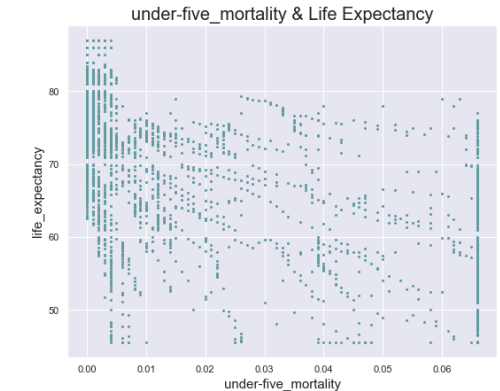
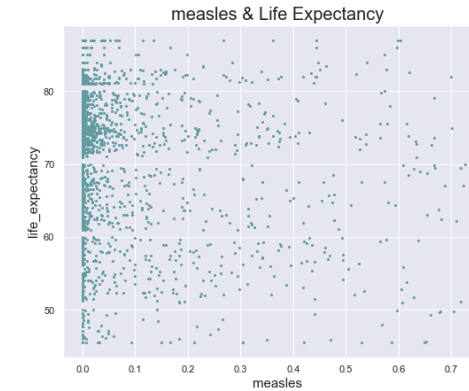
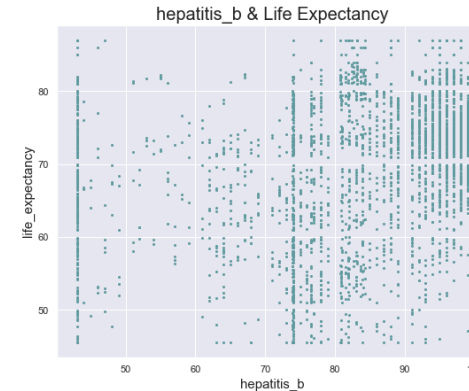
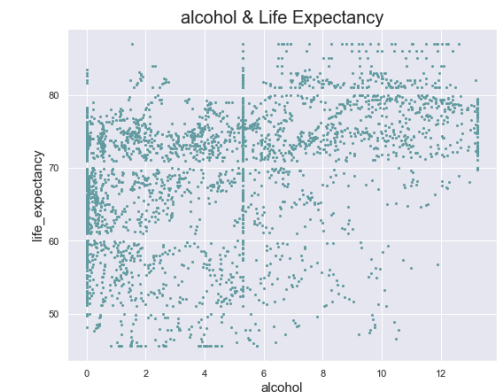
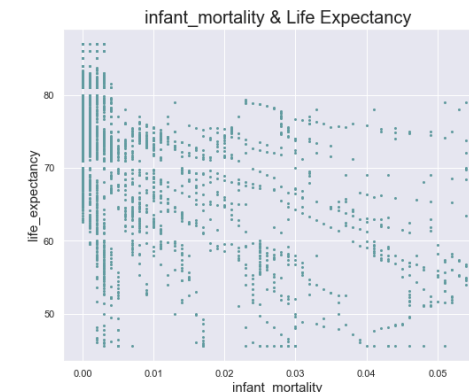
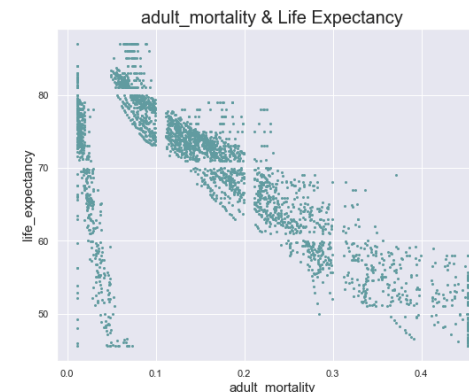
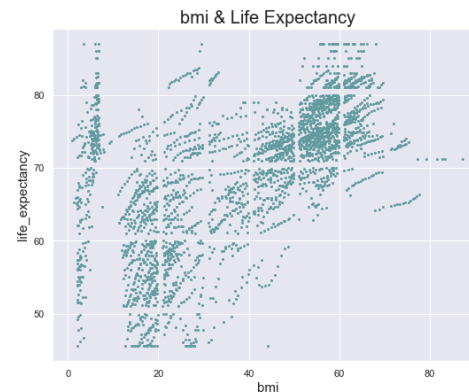
▶ 연도: 2000~2015년으로 구성

▶ status: 선진국/개발도상국을 'Developed/Developing'으로 구분





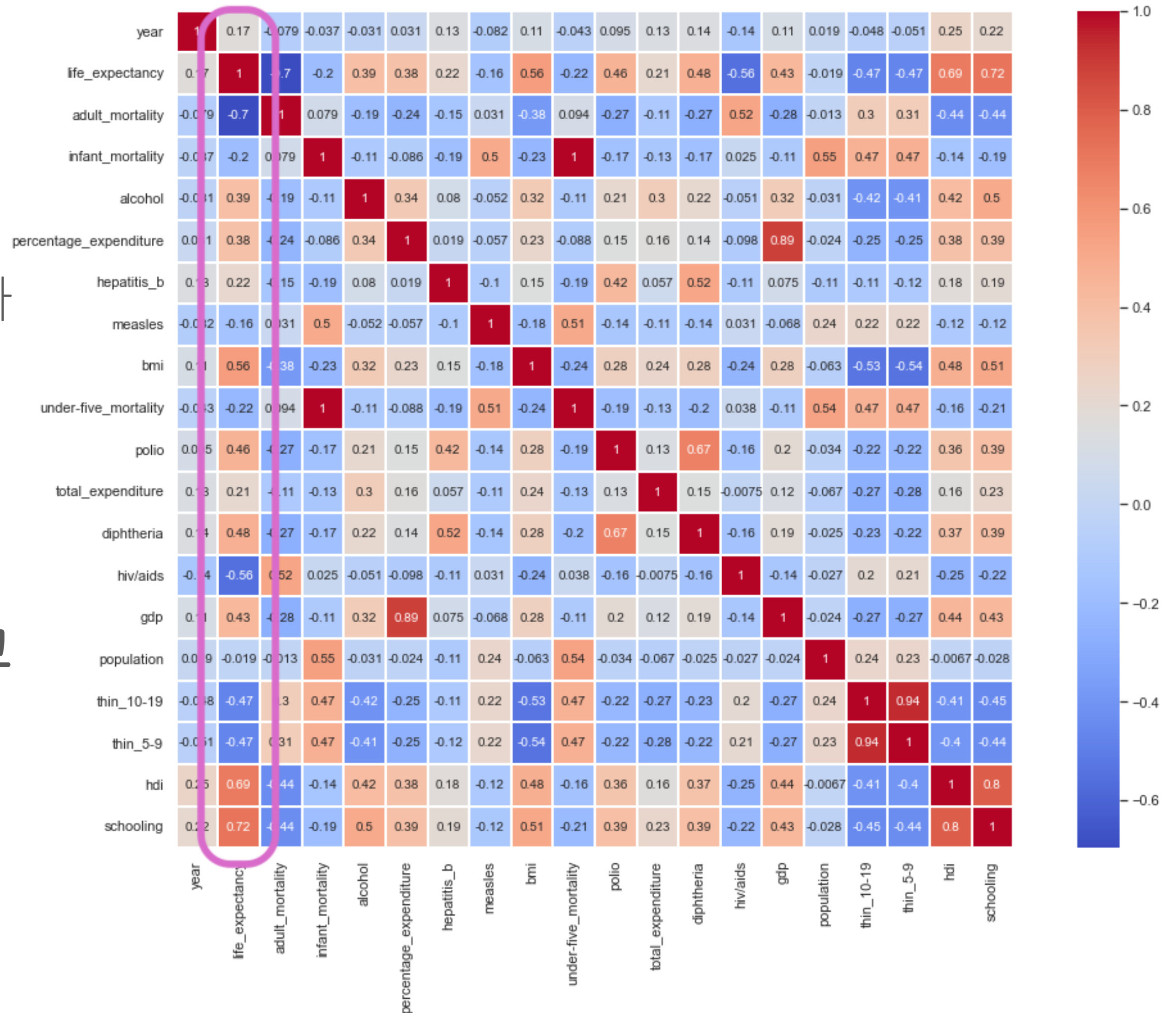
연속형 변수:  
기대수명과 각 변수 간의 분포  
를 나타낸 산점도



📎 연속형 변수:

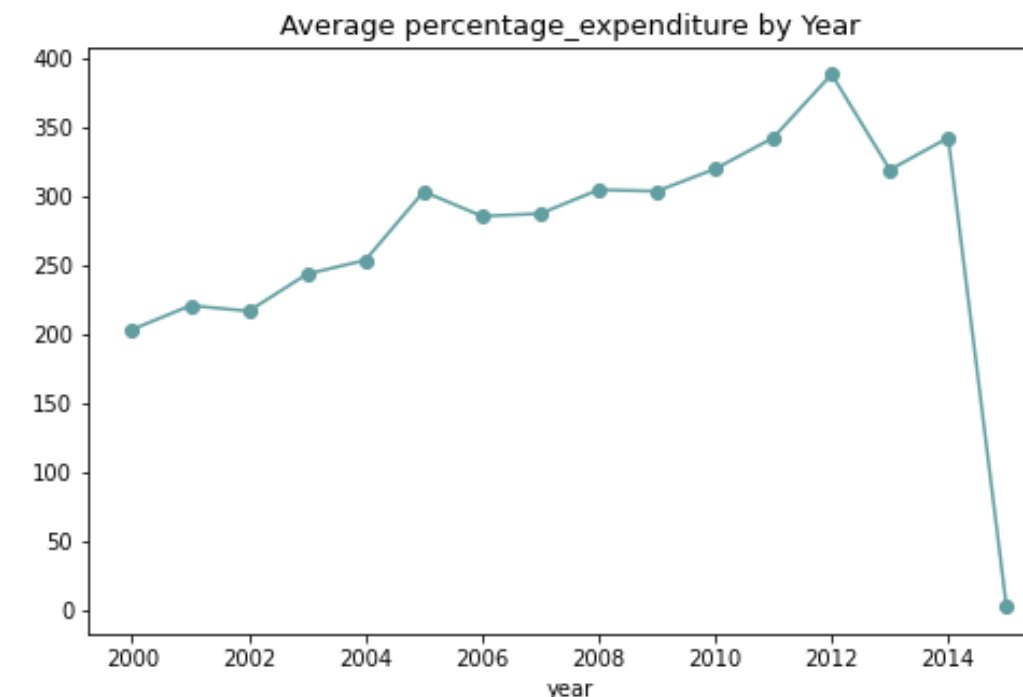
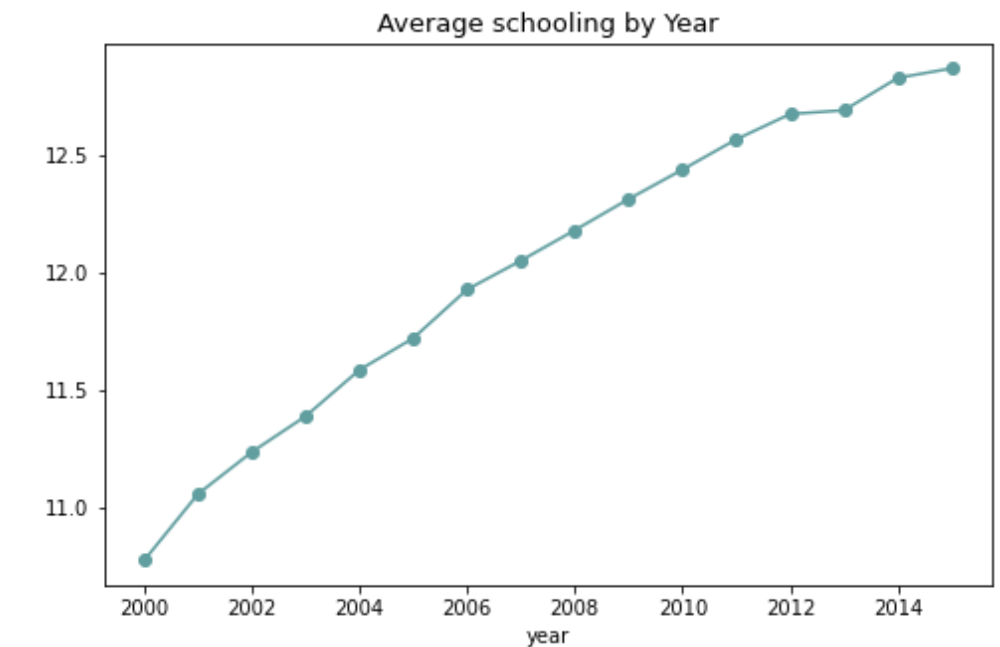
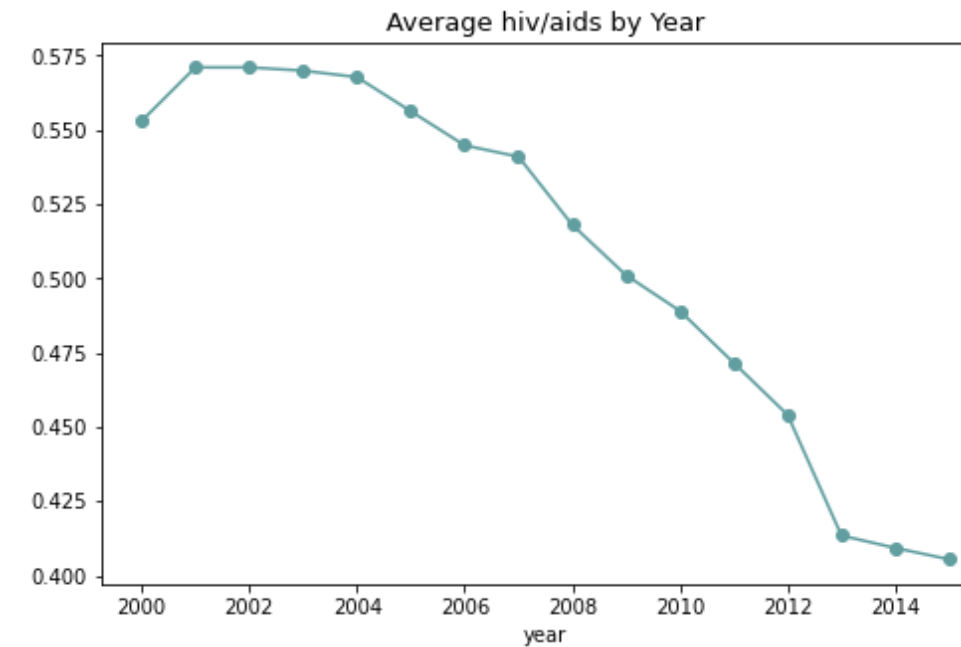
✓ heatmap에서 기대수명과  
인구와 연도의 **상관계수**가  
가장 0에 가까운 값

▶ 따라서 인구&연도는 기대  
수명과 어떤 상관관계를 맺고  
있지 **않다**고 해석 가능



📎 2000~2015년 간: 기대수명은 약 6% 증가

- 어른 사망률 약 24% 감소
- 유아 사망률 약 17% 감소
- 5세 미만 아동 사망률이 약 22.5% 감소
- 5세 미만 아동 hiv/aids 사망건수가 약 27% 감소
- 1세 인구 대비 B형 간염 예방 접종 약 11% 증가
- 홍역 환자 수 약 10% 감소
- 1세 인구 대비 소아마비 예방접종 약 6.7% 증가
- 급성 전염병 예방접종 약 8.8% 증가
- 10~19세 아동인구 대비 저체중 아동의 비율 약 12% 감소
- 5~9세 아동인구 대비 저체중 아동의 비율 약 11% 감소
- HDI 약 23% 증가
- 교육수준 약 33.3% 증가
- GDP 약 51% 증가
- GDP 대비 건강 관련 소비 약 70% 증가
- 정부 지출 대비 보건 관련 지출 비중 약 46% 증가
- 알코올 소비량 약 12.2% 증가






# 가설 검정

1. 아동이 건강할수록 기대 수명은 늘어날 것이다.

- 5세 미만 아동 사망률
- 1세 인구 대비 b형 간염 예방 접종자 수
- 홍역 환자 수
- 1세 인구 대비 소아마비 예방 접종자 수
- 1세 인구 대비 급성전염병 예방 접종자 수
- 5세 미만 아동 hiv/aids 사망건수
- 5~9세 및 10~19세 아동인구 대비 저체중 아동의 비율

해당 변수에 대해 기대수명과 카이제곱검정 실시  
한 결과 

`under-five_mortality & life_expectancy_Chi2 p-value:`  
`9.522686675835971e-201`

`hiv/aids & life_expectancy_Chi2 p-value: 2.5696170719655874e-165`

`hepatitis_b & life_expectancy_Chi2 p-value: 6.180266107590174e-84`

`polio & life_expectancy_Chi2 p-value: 1.0631422486925803e-170`

`diphtheria & life_expectancy_Chi2 p-value: 7.111798349366706e-116`

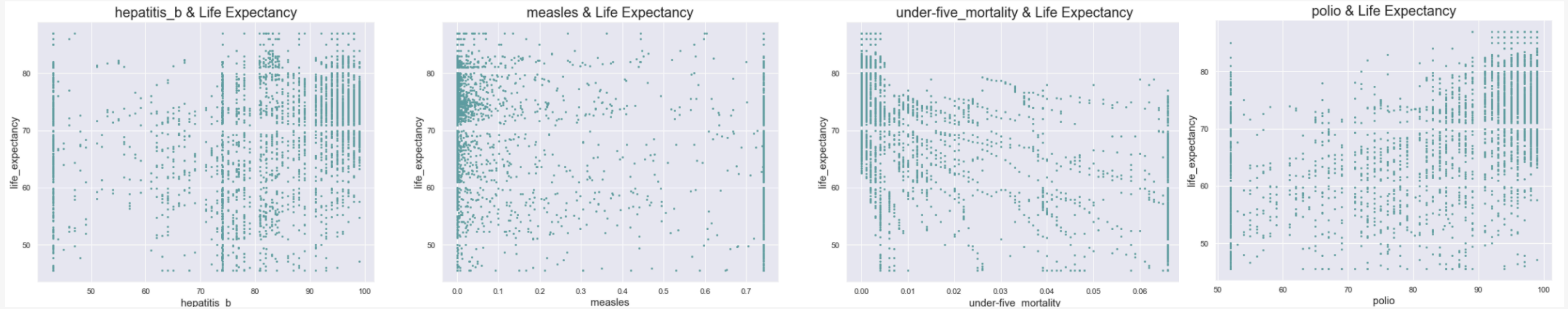
`thin_5-9 & life_expectancy_Chi2 p-value: 0.0`

`thin_10-19 & life_expectancy_Chi2 p-value: 0.0`

 모든 p-value가 0에 수렴 -> 해당 모든 변수는 기대수명과의 연관성 존재

# 가설 검증

1. 아동이 건강할수록 기대 수명은 늘어날 것이다.

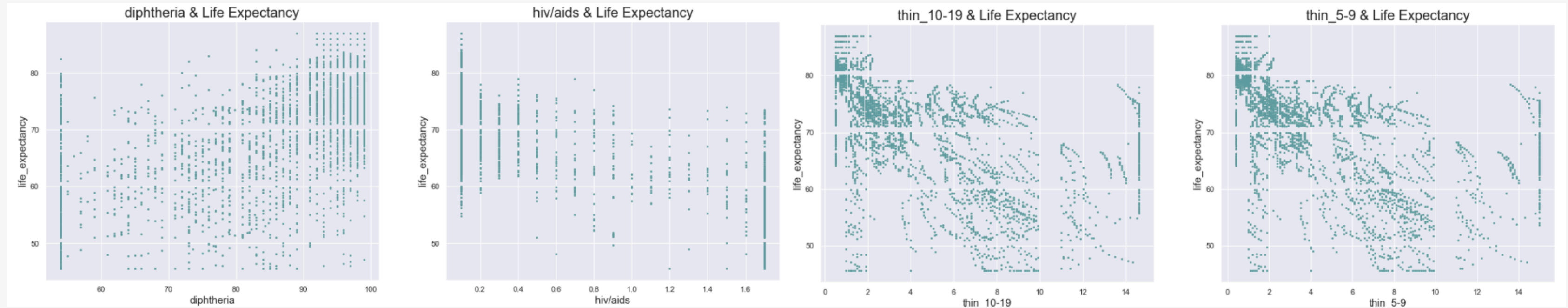


▶ 1세 인구 대비 b형 간염 및 소아마비 예방접종은 증가할수록 기대수명 ↑

▶ 5세 미만 사망률 과 홍역 환자 수는 감소할수록 기대수명 ↑

# 가설 검증

1. 아동이 건강할수록 기대 수명은 늘어날 것이다.



▶ 1세 인구 대비 급성전염병 예방 접종은 늘어날수록 기대수명 ↑

▶ 5세 미만 아동 hiv/aids 사망건수와 5~9세 및 10~19세 인구 대비 마른 아동의 비율은 감소할수록 기대수명 ↑

# 가설 검증

1. 아동이 건강할수록 기대 수명은 늘어날 것이다.

- 🔍 카이제곱검정과 그래프를 통해 가설1이 성립할 수 있음을 확인
- 🔍 어린 아동의 사망률이 증가하면 사망자의 연령이 감소해 기대수명이 단축
- ▶ 아동의 사망률이 감소하고 많은 아동이 건강해지면 사망자 연령 감소가 발생하지 않아 “기대수명의 증가”로 이어지게 됨



# 가설 검정

2. 국가가 선진화될수록 기대수명이 증가할 것이다.

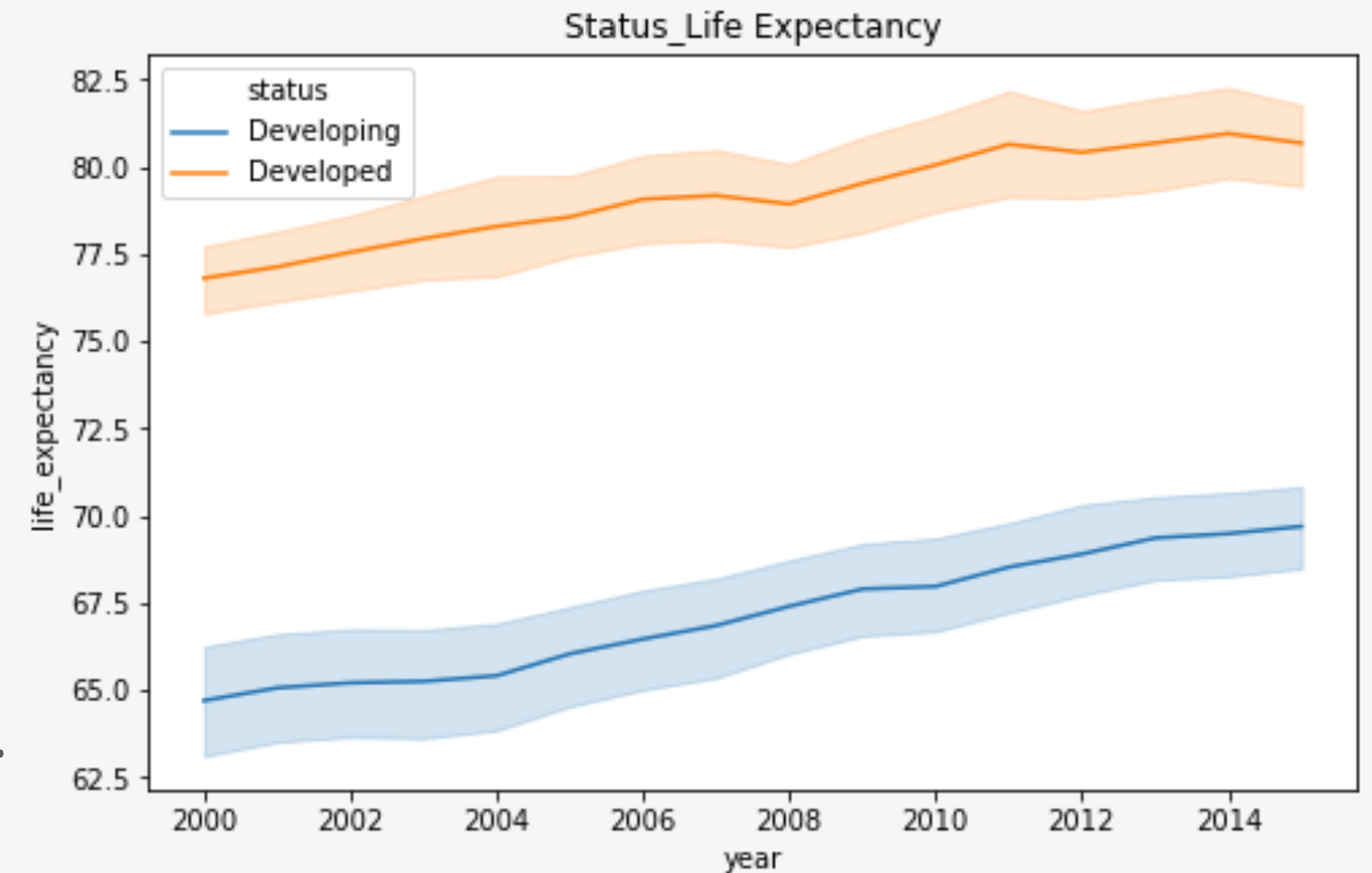
📎 status를 기준으로 두 그룹으로 나눠 **t-test**

시행:

p-value=4.238344691910959e-171

-> 0에 수렴하는 작은 값

▶ 따라서 두 그룹 평균 기대수명에 차이 존재함



# 가설 검정

2. 국가가 선진화될수록 기대수명이 증가할 것이다.

\* HDI: Human Development Index, 각국의 인간 발전 정도와 선진화 정도를 평가한 지수

 교육수준/HDI/GDP을 각각 기대수명과 카이제곱검정 실시한 결과 

- schooling & life\_expectancy\_Chi2 p-value: 5.433720742542671e-255
- hdi & life\_expectancy\_Chi2 p-value: 0.0
- gdp & life\_expectancy\_Chi2 p-value: 1.0

 교육수준, HDI: p-value가 0에 수렴하므로 기대수명과 연관성을 가짐

 GDP: p-value=1 -> 기대수명으로부터 독립적인 변수

# 가설 검정

2. 국가가 선진화될수록 기대수명이 증가할 것이다.

- 🔍 t-test, 카이제곱검정, 그래프를 통해 **나라의 선진화 정도**에 따른 기대 수명 간의 변화가 발생한다는 것을 통해 검증
- 🔍 국가가 선진화될수록 공중보건 체계가 잘 구축되어 있고, 교육수준이 높을수록 위생 및 건강과 관련된 교육이 이루어지기 때문에 발전 정도에 따라 기대수명이 다르게 측정될 수 있음


# 가설 검정


3. 건강을 위한 지출이 증가할수록 기대수명은 늘어날 것이다.

- GDP 대비 건강 관련 지출 비중
- 정부 지출 대비 보건 관련 지출 비중
- 알코올 소비량

 각 특성과 기대수명에 대해 카이제곱검정 실시한 결과 

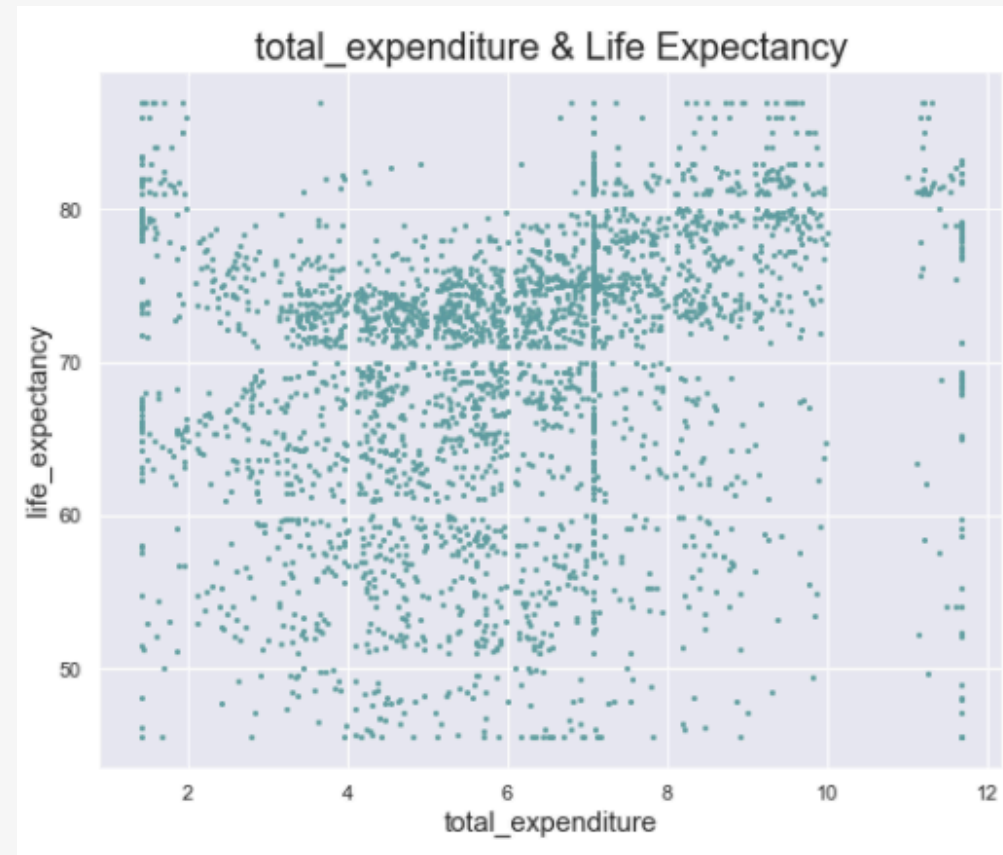
- percentage\_expenditure & life\_expectancy\_Chi2 p-value: 9.766660350236392e-116
- total\_expenditure & life\_expectancy\_Chi2 p-value: 0.7575309470505511
- alcohol & life\_expectancy\_Chi2 p-value: 1.0

 **GDP 대비 건강 관련 지출 비중:** p-value=0에 수렴하므로 기대수명과 연관성을 가진다고 볼 수 있음

 **정부 지출 대비 보건 관련 지출 비중& 알코올 소비량:** p-value=1 혹은 1에 수렴하므로 기대수명과 독립적인 관계일 것

# 가설 검증

3. 건강을 위한 지출이 증가할수록 기대수명은 늘어날 것이다.



- ▶ GDP 대비 건강 관련 지출 비중만 늘어날수록 기대수명이 증가하는 경향이 있음
- ▶ 다른 두 변수의 분포는 비선형적 분포를 띠고 있음 -> 기대수명에 주는 영향이 미미할 것

# 가설 검증

3. 건강을 위한 지출이 증가할수록 기대수명은 늘어날 것이다.

🔍 카이제곱검정, 그래프를 통해 GDP 대비 건강 관련 지출이 증가할 때 가설 3이 성립

↔ 그러나 알코올 소비량 & 정부 지출 대비 보건 관련 지출에 대해서는 가설3이 성립 x

🔍 건강 관련 지출 비중을 늘린다는 것은 그만큼 개인 혹은 사회가 건강에 관심을 가지고 있다는 증거  
ex)

- 개인의 측면: 지속적으로 건강에 관심을 가지면 질병에 걸려도 초기에 발견 및 치료 가능

- 사회적 측면: 의료기술 발전에 원활한 투자가 이루어지면 많은 질병에 대한 다양한 치

# 모델 학습 및 결과

📎 Feature Engineering:

- ✓ '5세 미만 아동 사망률 대비 hiv로 인한 사망률'을 컬럼으로 추가
- ✓ 기대수명과 상관관계가 매우 낮은 '인구' 제거
- ✓ <유아 사망률과 5세 미만 아동 사망률>, <교육수준과 hdi>, <5~9세 및 10~19세 저체중 아동의 비율> 은 높은 상관관계를 맺고 있음  
-> 다중공선성이 발생할 수 있으므로 이들 중 기대수명과 강한 상관관계를 가진 한 변수만 남기고 학습 진행

# 모델 학습 및 결과

📎 Baseline 모델: 기대수명의 평균값  
-> R2 score: 0.0, MAE: 7.741863297777127

📎 모델 학습 진행한 후 검증 데이터에 대한 성능 도출

1. Linear Regression

**R2 score: 0.8559515731374612**

2. Ridge Regression: best\_alpha=0.1

**R2\_score: 0.9933360441826822**

3. Randomforest Regressor

**R2 score: 0.9642667362864851**

4. XGBoost Regressor

**R2\_score: 0.9992083390007229**

👉 테스트 데이터에 대한 성능: 0.9376987037089975



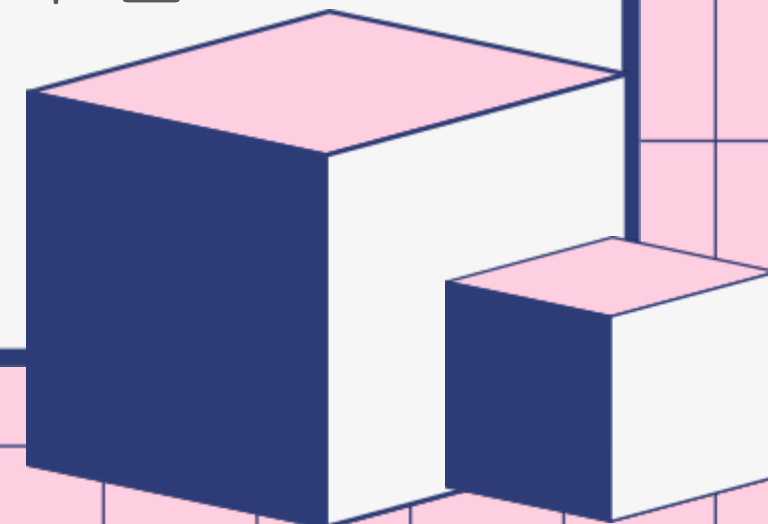
# 결론

📎 여러 변수 중 HDI, 교육수준, 5세 미만 아동 사망률, GDP 대비 건강 관련 지출 비중이 기대수명에 상당한 영향을 미치고 있음

📎 유아기의 사망률 및 예방접종과 관련된 요소들이 예상보다 기대수명에 상당한 영향을 미치고 있음

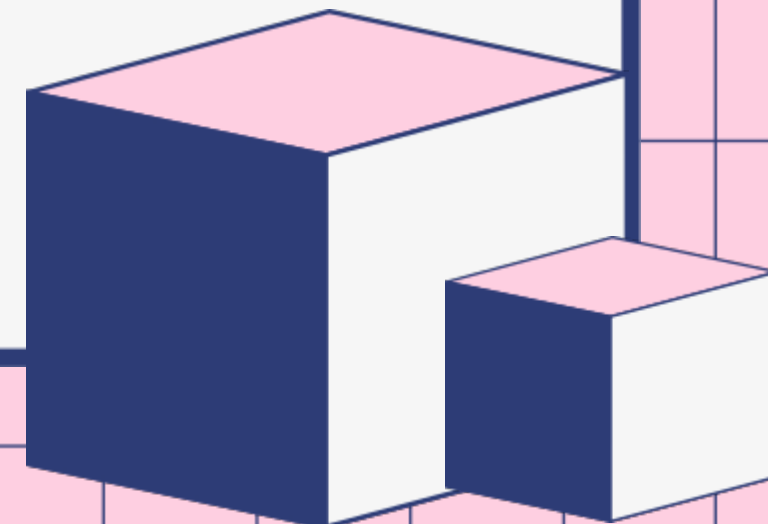
-> 다양한 질병으로부터 선제적 보호를 받지 못하는 아이들을 위한 범국가적 차원의 조치가 필요할 것

📎 더불어 교육수준 및 HDI 등 사회/경제적 요인이 기대수명에 상당한 연관성을 가진다는 점에서 건강한 생활습관 및 위생 등에 대한 교육의 중요성이 대두됨



# 회고

- 📎 EDA와 데이터 전처리에서 생각보다 많은 시간을 소요했지만, 더 다양하게 시각화 및 데이터 분석 진행해보았으면 하는 아쉬움이 남음
- 📎 최근 데이터를 찾고자 했으나 찾지 못해 최근 데이터를 다루지 못한 점이 가장 아쉬움
- 📎 데이터 분석 프로젝트를 진행하면서 데이터 전처리를 다양하게 시도하고, 시각화에서 많은 오류를 경험하면서 배울 수 있었음



감사합니다:)