DATE. 2025.06.16

4조

데이터과학입문

프로젝트 발표

202111600 이소흔 | 202217189 이서후 | 202315219 정세연 | 202224543 박서경

CONTENTS

목차

1-1. 구성원 소개 & 역할

구성원	역할
이소흔	보고서 작성, 데이터 전처리, ppt 제작, 발표
이서후	기술 통계 해석, 데이터 수집 및 전처리, 발표
정세연	모평균 및 신뢰구간, 데이터 전처리, 발표
박서경	이표본 검정, 데이터 전처리, 발표

1-2. 데이터 소개

▶ 활용 데이터

: Life Expectancy (WHO)

전 세계 국가들의 기대수명과 건강/경제/사회 지표
_WHO

▶ 데이터 출처 : 캐글(kaggle)

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

▶ 데이터 크기

: 샘플 수 2938개, 변수 22개 (결측치 포함)

▶ 주요 변수 설명

- Country : 국가명. 총 193개 존재
- Status : 국가의 개발 상태 (Developed : 선진국, Developing : 개발도상국)
- Life expectancy : 기대수명
- Adult Mortality : 성인 사망률 (15~60세 인구 1,000명당 연간 사망자 수)
- infant deaths : 영아 사망자 수(1세 미만)
- BMI: 평균 체질량지수(Body Mass Index)
- GDP : 1인당 국내총생산(미국 달러)
- Schooling: 평균 교육 연수(년)

1-3. 프로젝트 소개

▶ 주제

- 기술 및 추론 - 기대 수명은 사회/경제적 격차를 반영하는 핵심 지표로, 기술 및 추론 통계로 그 차이를 탐색하고자 함

▶ 소주제

1. 기대수명 분포의 기술통계량 및 특성 분석
2. 인구 집단별 기대수명 평균 및 신뢰구간 비교
3. 집단 간 기대수명 차이 - 이표본 검정

▶ 진행 과정

1. 주제 선정
2. 데이터 수집 및 소주제 설정
3. 데이터 전처리
4. 결과 분석
5. ppt 제작 및 발표 준비

2-1. 결측치 처리

- 결측치가 포함된 행이 전체 행의 절반 이상 → 행 삭제 시 데이터 개수가 과도하게 줄어듦
- 데이터의 맥락을 반영하고 왜곡을 최소화 할 수 있는 결측치 처리가 필요함

▶ 결측치 처리 방법

1. 변수 별 결측치 포함 비율 파악

```
missing_values = df.isnull().sum()
missing_percent = (missing_values / len(df)) * 100

missing_summary = pd.DataFrame({
    '결측치' : missing_values,
    '결측치 비율(%)': missing_percent
}).sort_values(by='결측치', ascending=False)

missing_summary
```

	결측치	결측치 비율(%)
Population	652	22.191967
Hepatitis B	553	18.822328
GDP	448	15.248468

[비율 15% 이상]

Total expenditure	226	7.692308
Alcohol	194	6.603131
Income composition of resources	167	5.684139
Schooling	163	5.547992

[비율 5% 이상 15% 미만]]

2-1. 결측치 처리

2. 비율 별 결측치 처리 방법을 다르게 설정

- 비율이 15% 이상인 변수 → 1. 국가별 평균으로 대체 / 2. 대체되지 않은 결측값은 전체 평균으로 대체

```
#15% 이상의 결측률인 변수를 국가별 평균으로 대체
#여전히 남으면 전체 평균
for col in high_missing_cols:
    df_cleaned[col] = df_cleaned.groupby("Country")[col].transform(lambda x: x.fillna(x.mean()))
    df_cleaned[col] = df_cleaned[col].fillna(df_cleaned[col].mean())
```

- 비율이 15% 미만이면서 5% 이상인 변수 → 전체 중앙값

```
[9]  #결측치가 남아 있는 경우 전체 중앙값으로 대체
     for col in mid_missing_cols:
         df_cleaned[col] = df_cleaned[col].fillna(df_cleaned[col].median())
```

- 비율이 5% 미만인 변수 → 전체 평균

```
[10] #결측치가 남아 있는 경우 전체 평균으로 대체
     for col in low_missing_cols:
         df_cleaned[col] = df_cleaned[col].fillna(df_cleaned[col].mean())
```

2-2. 기초통계량

▶ 평균 및 분산 - 연속형 변수

- 기대수명(Life expectancy)
  - 평균 : 69.22
  - 분산 : 90.40
- 영아 사망자 수(infant deaths)
  - 평균 : 30.30
  - 분산 : 13906.66
- GDP
  - 평균 : 7378.4045
  - 분산 : 173151574.59726
- 성인 사망률 (Adult Mortality)
  - 평균 : 164.80
  - 분산 : 15395.92
- BMI 지수
  - 평균 : 38.32
  - 분산 : 397.11
- Alcohol
  - 평균 : 4.55
  - 분산 : 15.38
- B형 간염 접종률(Hepatitis B)
  - 평균 : 78.65
  - 분산 : 603.18
- 1인당 건강 관련 지출 GDP 대비 비율 (percentage expenditure)
  - 평균 : 738.25
  - 분산 : 3951805.48

2-2. 기초통계량

▶ 평균 및 분산 - 연속형 변수

- 홍역 보고 건수(Measles)
  - 평균 : 2419.59
  - 분산 : 131498338.34
- 전체 인구(Population)
  - 평균 : 12734717.56
  - 분산 : 2896259126605159.5
- 소아마비 백신 접종률(Polio)
  - 평균 : 82.55
  - 분산 : 545.32
- 5~9세 저체중 비율(퍼센트)
  - 평균 : 4.87
  - 분산 : 20.09
- 1~19세 저체중 비율(퍼센트)
  - 평균 : 4.84
  - 분산 : 19.31
- GDP 대비 건강 비용 지출 비율 (Total expenditure)
  - 평균 : 5.92
  - 분산 : 5.76
- 디프테리아 백신 접종률(Diphtheria)
  - 평균 : 82.32
  - 분산 : 558.85
- HIV/AIDS로 인한 사망률(HIV/AIDS)
  - 평균 : 4.55
  - 분산 : 15.38

2-2. 기초통계량

▶ 평균 및 분산 - 연속형 변수

- 소득 자원 구성 지수(0~1)
  - 평균 : 0.63
  - 분산 : 0.04

- 5세 미만 사망자 수(under-five deaths)
  - 평균 : 82.55
  - 분산 : 545.32

분산이 압도적으로 **크다**
→ 개발도상국과 선진국 간의 **격차가 클 것**

▶ 빈도 - 범주형 변수

- 상태(Status)

	Status	count
개발도상국	Developing	2426
선진국	Developed	512

개발도상국이 현저히 많음 (약 83%)

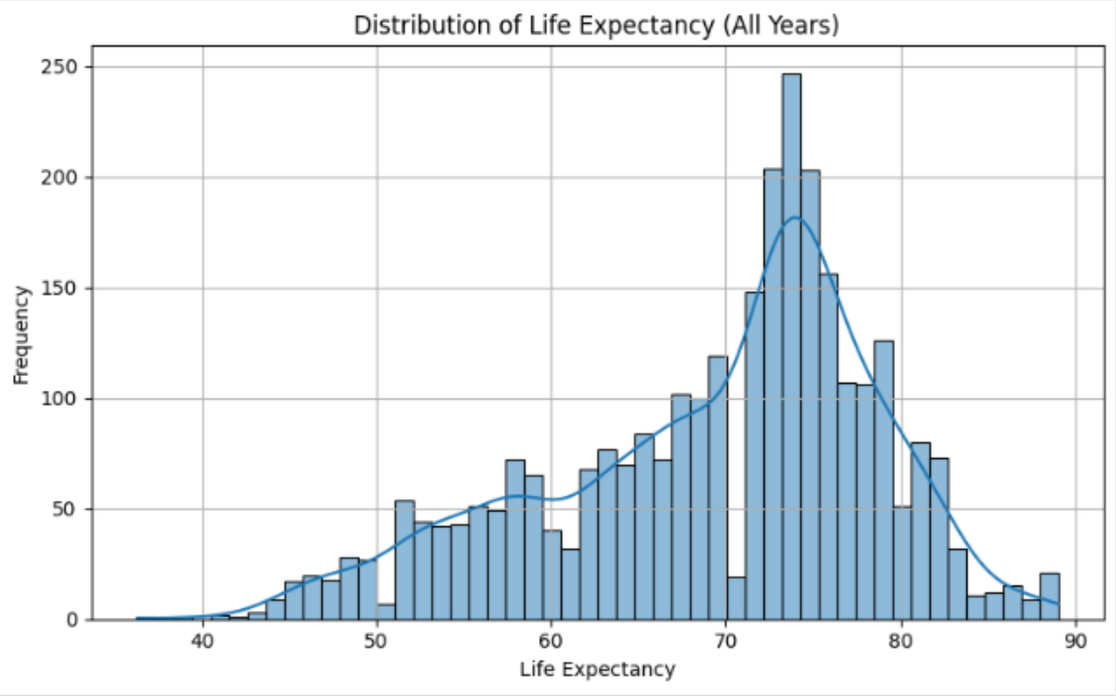
- 국가(Country)

10개의 국가들만 데이터가 1개 존재
나머지 국가의 데이터는 16개씩 존재

- 년도(Year)

Year	count
2013	193
2015	183
2014	183
2012	183
2011	183
2010	183
2009	183
2008	183
2007	183
2006	183
2005	183
2004	183
2003	183
2002	183
2001	183
2000	183

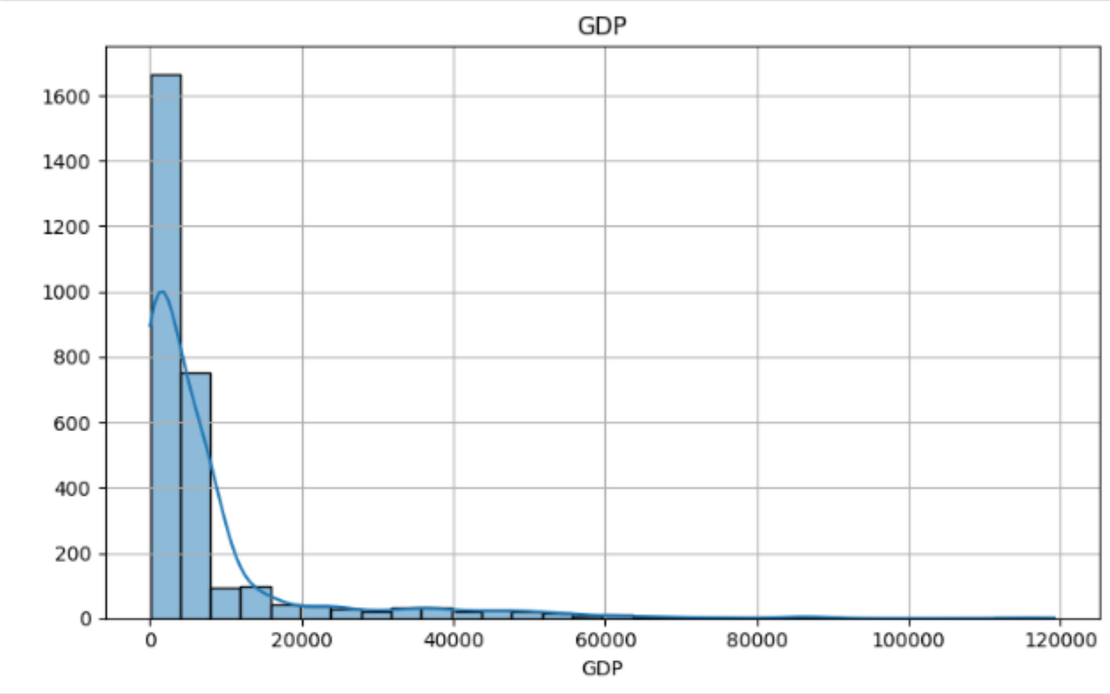
3-1. Life expectancy(기대수명) 변수



- 평균 <중앙값 으로, 오른쪽으로 치우친 분포
- 왜도<0 : 왼쪽으로 꼬리가 긴 분포
- 첨도<0 : 정규분포보다 정점이 낮음

- 평균 (Mean): 69.22
- 중앙값 (Median): 72.00
- 표준편차 (Std): 9.51
- 최빈값 (Mode): 73.00
- 최소값 (Min): 36.30
- 최댓값 (Max): 89.00
- 왜도 (Skewness): -0.64
- 첨도 (Kurtosis): -0.23

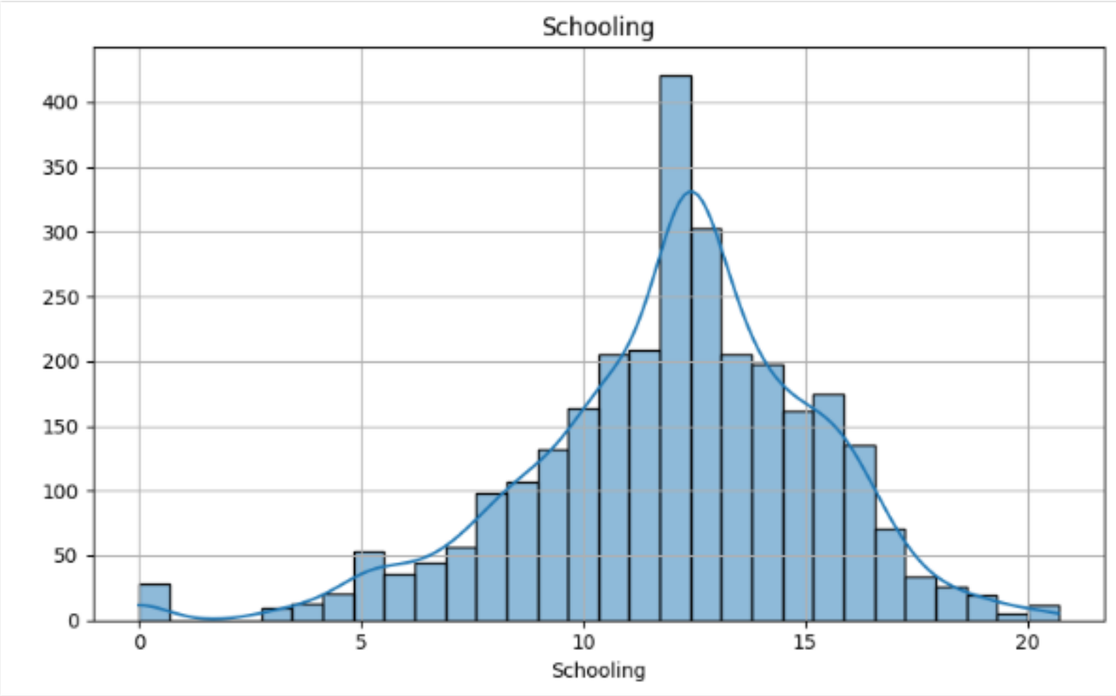
3-2. GDP 변수



- 평균 (Mean): 7378.40
- 중앙값 (Median): 2834.76
- 표준편차 (Std): 13158.71
- 최빈값 (Mode): 7378.40
- 최소값 (Min): 1.68
- 최댓값 (Max): 119172.74
- 왜도 (Skewness): 3.49
- 첨도 (Kurtosis): 15.05

- 표준편차가 매우 큼 - 대다수 국가의 GDP는 낮고, 일부 국가만 매우 높은 GDP를 가짐.
- 왜도 : 매우 큼 → 오른쪽 꼬리가 길게 늘어진 분포
- 첨도 : 매우 큼 → 뾰족하고, 정규분포보다 정점이 높음 (극단값이 분포에 강한 영향)

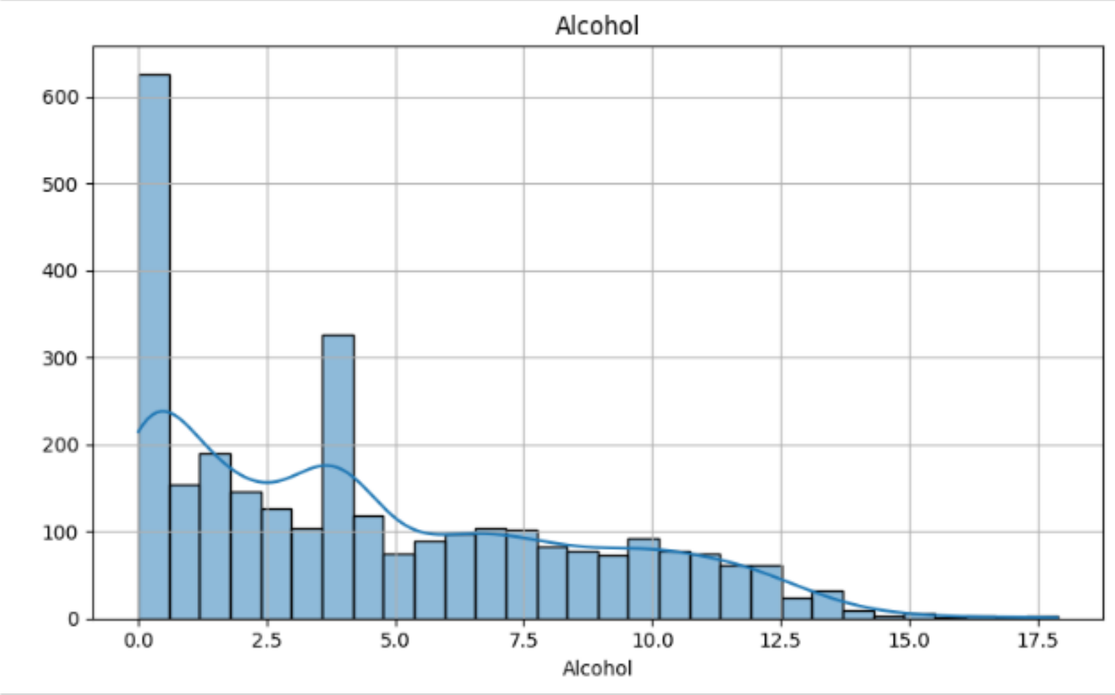
3-3. Schooling(평균 교육 년수) 변수



- 대칭적인 구조, 약간의 왼쪽 치우침
- 왜도 : 왼쪽 꼬리가 긴 분포 (극단적으로 교육 년수가 낮은 값 존재)
- 첨도 : 중심에 과도하게 집중되지 않고 전체적으로 퍼져 있는 형태

- 평균 (Mean): 12.01
- 중앙값 (Median): 12.30
- 표준편차 (Std): 3.27
- 최빈값 (Mode): 12.30
- 최소값 (Min): 0.00
- 최댓값 (Max): 20.70
- 왜도 (Skewness): -0.63
- 첨도 (Kurtosis): 1.12

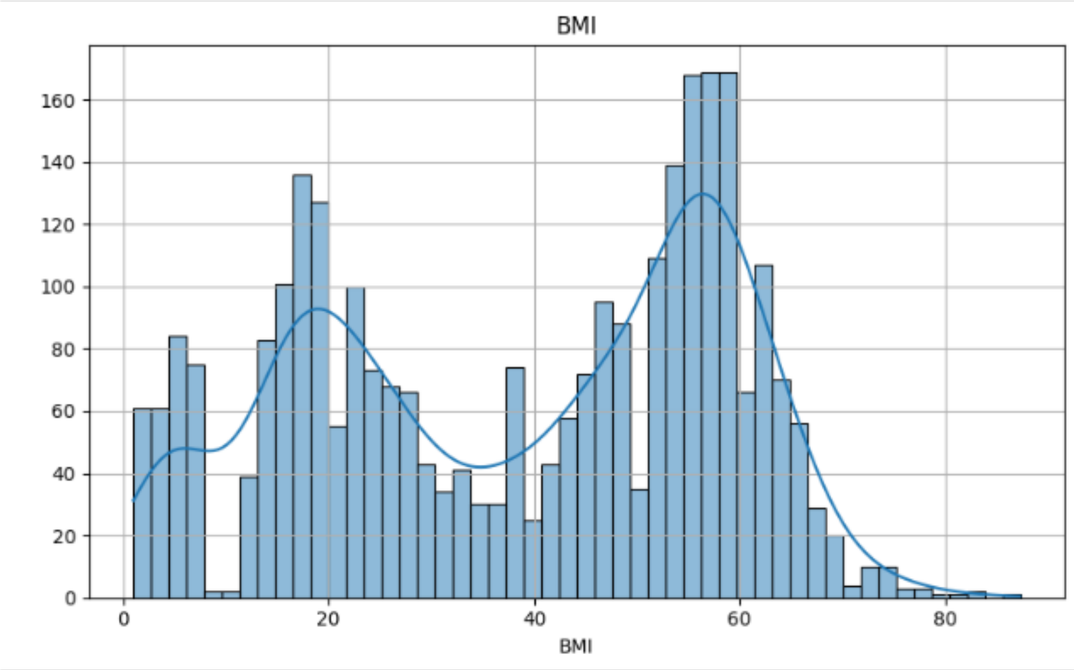
3-4. Alcohol(알코올 소비량) 변수



- 왜도 : 오른쪽 꼬리가 긴 분포
- 첨도 : 뾰족하지 않은 평평한 분포 (중심부에 값이 몰리지 않음)

- 평균 (Mean): 4.55
- 중앙값 (Median): 3.75
- 표준편차 (Std): 3.92
- 최빈값 (Mode): 0.01
- 최소값 (Min): 0.01
- 최댓값 (Max): 17.87
- 왜도 (Skewness): 0.65
- 첨도 (Kurtosis): -0.63

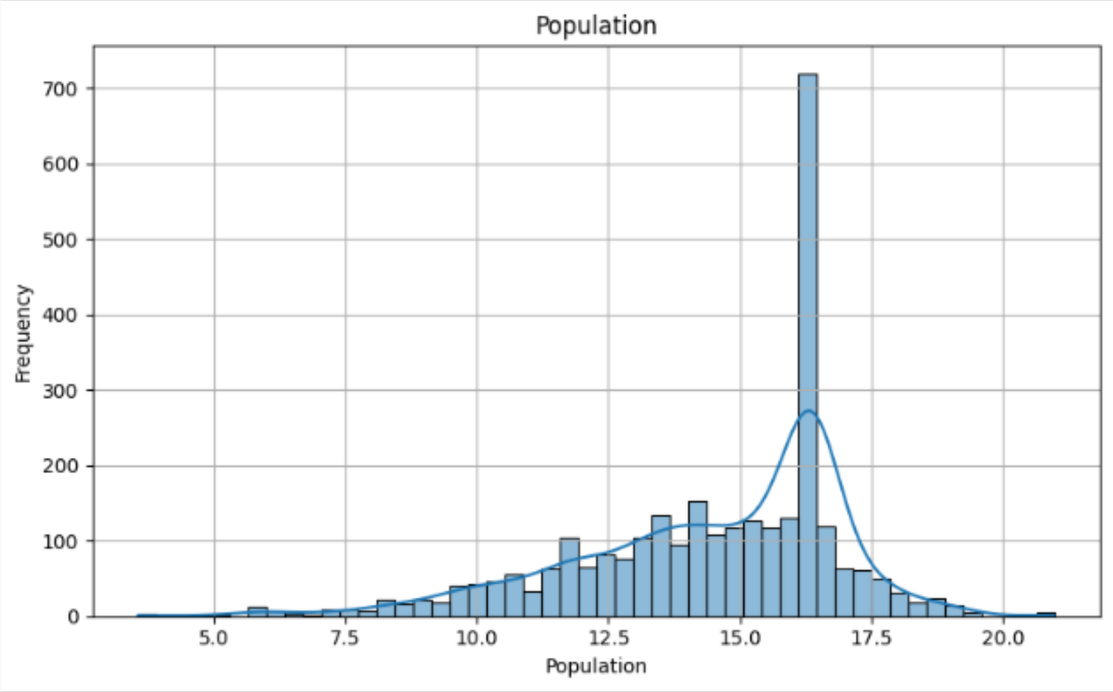
3-5. BMI 변수



- 다봉분포 형태를 띔
- 왜도 <0 : 왼쪽 꼬리가 약간 길게 늘어남
- 첨도 <0 : 중심값에 덜 집중되어있음

- 평균 (Mean): 38.32
- 중앙값 (Median): 43.00
- 표준편차 (Std): 19.93
- 최빈값 (Mode): 38.32
- 최소값 (Min): 1.00
- 최댓값 (Max): 87.30
- 왜도 (Skewness): -0.22
- 첨도 (Kurtosis): -1.27

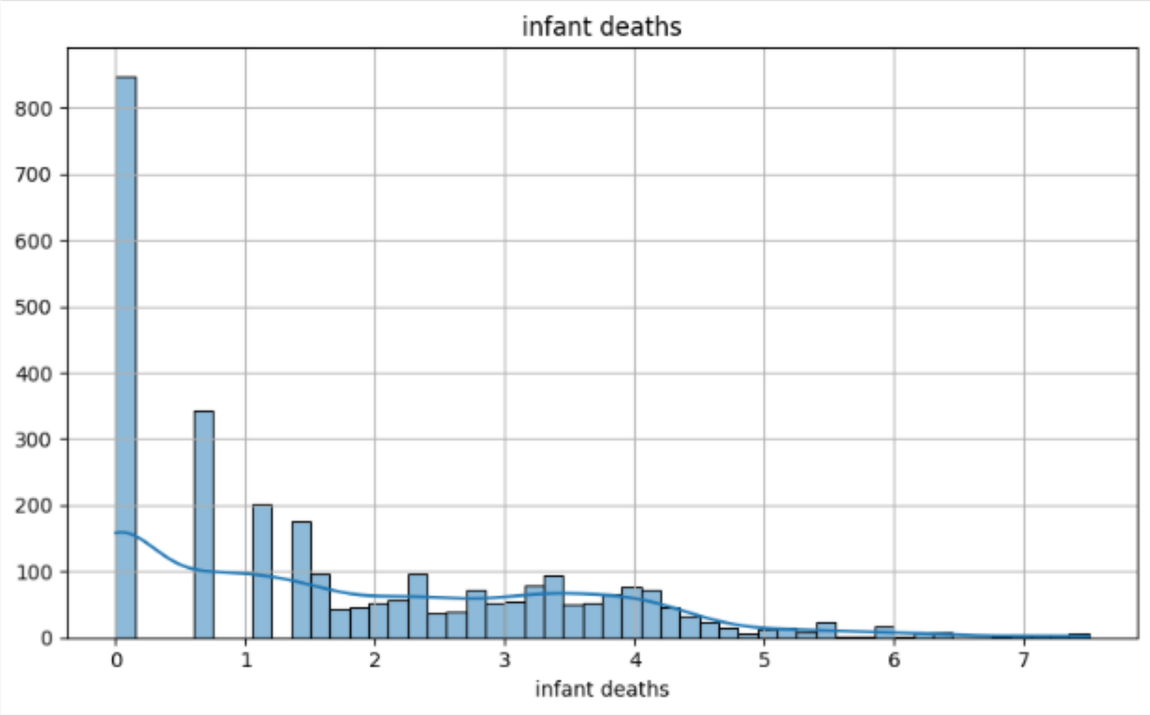
3-6. Population(인구) 변수



- 평균 (Mean): 12734717.56
- 중앙값 (Median): 3625717.50
- 표준편차 (Std): 53816903.73
- 최빈값 (Mode): 12734717.56
- 최소값 (Min): 34.00
- 최댓값 (Max): 1293859294.00
- 왜도 (Skewness): 18.03
- 첨도 (Kurtosis): 383.01

- 오른쪽으로 매우 치우쳐져 있는 분포
- 왜도 : 대부분의 국가가 왼 쪽에 몰려있고, 오른쪽에 극소수 많은 인구의 국가 존재
- 첨도: 거의 모든 데이터가 한 지점에 몰려있음

3-7. Infant deaths(영아 사망자 수) 변수



- 평균 (Mean): 30.30
- 중앙값 (Median): 3.00
- 표준편차 (Std): 117.93
- 최빈값 (Mode): 0.00
- 최소값 (Min): 0.00
- 최댓값 (Max): 1800.00
- 왜도 (Skewness): 9.78
- 첨도 (Kurtosis): 115.84

- 평균이 중앙값보다 10배 이상 큼 , 최빈값은 0 → 대부분 국가는 유아 사망 수가 매우 낮음
- 왜도 : 극단적인 왜도, 오른쪽 꼬리가 길게 늘어진 분포
- 첨도: 첨도가 매우 높음, 극단치 영향력이 매우 큼

4-1. GDP

- GDP Group별 GDP 평균 및 기대수명, 그룹별 95% 신뢰구간

Group_ GDP	n	Average GDP	mean_life	sd_life	se	ci_lower	ci_upper
Low GDP	980	377.7	62.8	9.3	0.3	62.2	63.4
Mid GDP	979	3079.0	71.0	7.7	0.2	70.6	71.4
High GDP	979	18685.7	73.9	7.7	0.2	73.5	74.3

1. Low 그룹, 다른 그룹에 비해 데이터의 분산↑

2. 표준오차 작고 신뢰구간은 겹치지 않음

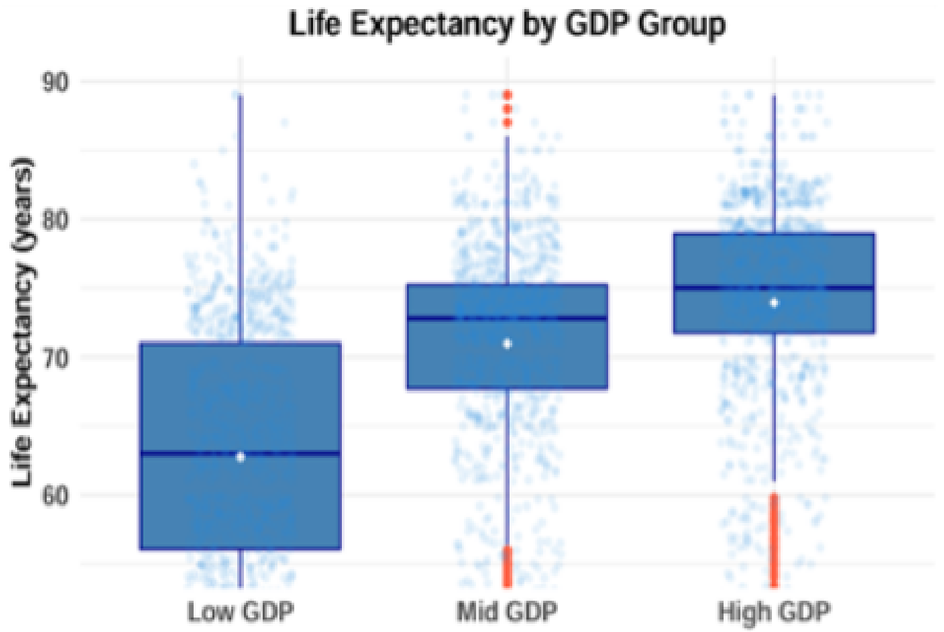
: 평균값의 신뢰도 ↑, 통계적으로 유의미한 차이 보임.

3. Mid ↔ High Group 차이가 크지 않음.

: **어느정도 GDP 궤도에 오른 후**에는 추가적인 경제적 자원이 기대수명을 **크게 개선시키지 못함**

4-1. GDP

- GDP Group별 GDP 평균 및 기대수명, 그룹별 95% 신뢰구간



4-2. Schooling

- 교육연수 그룹별 기대수명 및 교육연수 평균, 그룹별 95% 신뢰구간

Group_ Schooling	n	Average Schooling	mean_life	sd_life	se	ci_lower	ci_upper
Low Edu	980	8.4	60.5	7.8	0.2	60.1	60.9
Mid Edu	979	12.3	70.3	7.2	0.2	69.9	70.7
High Edu	979	15.3	76.9	4.7	0.2	76.5	77.3

1. 교육수준이 높을 수록 기대수명이 뚜렷하게 증가

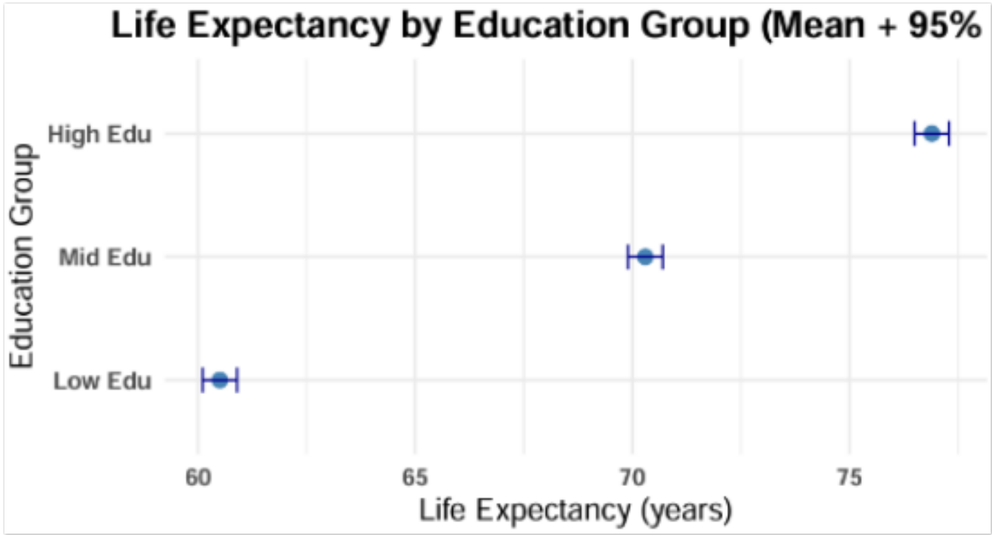
Low Edu: 60.5세 → Mid Edu: 70.3세 → High Edu: 76.9세

2. 신뢰구간이 겹치지 않음 → 통계적으로 유의미한 차이

3. 표준편차(sd)와 표준오차(se)가 작다 → 평균이 신뢰성 있음

4-2. Schooling

- 교육연수 그룹별 기대수명 및 교육연수 평균, 그룹별 95% 신뢰구간



4-3. Status

- 개발수준 그룹별 기대수명 평균 및 95% 신뢰구간(범주형 변수)

Status	n	mean_life	sd_life	se	ci_lower	ci_upper
Developing	2426	67.1	9	0.2	66.7	67.5
Developed	512	79.2	3.9	0.2	78.8	79.6

1. 개발 수준에 따른 기대수명 차이가 매우 큼

개발도상국(67.1세) – 선진국(79.2세) : 12.1세 차이

2. 신뢰구간이 겹치지 않음

→ 통계적으로 유의미한 차이

3. 표준편차(sd)와 표준오차(se)

- 개발도상국: **표준편차 9.0**

→ 국가 간 기대수명 편차가 큼 (불균형)

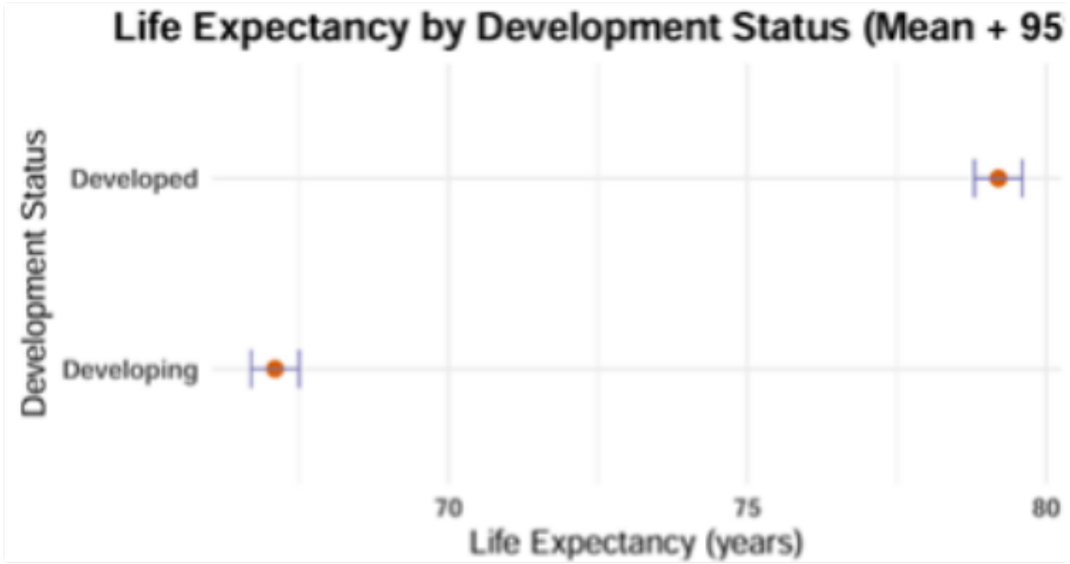
- 선진국: **표준편차 3.9**

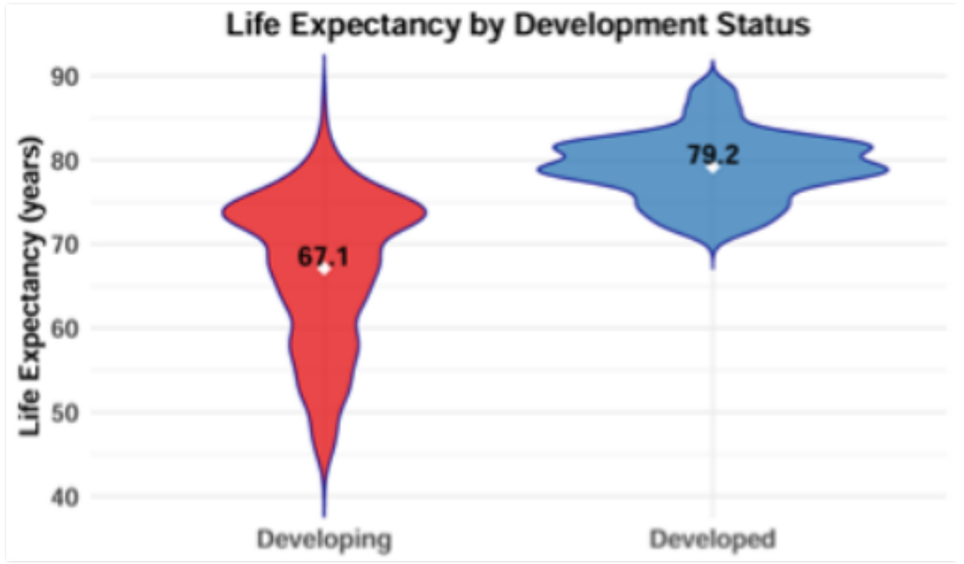
→ 기대수명이 상대적으로 일정하고 안정적

- 표준오차는 둘 다 0.2로 작아, **평균의 신뢰성 높음**

4-3. Status

- 개발수준 그룹별 기대수명 평균 및 95% 신뢰구간(범주형 변수)





4-4. Infant deaths

- 영아 사망 수준 그룹별 평균 영아 사망 수 및 기대수명 , 그룹별 95% 신뢰구간

Group_ Infant deaths	n	Average Infant deaths	mean_life	sd_life	se	ci_lower	ci_upper
High infant deaths	1370	64.1	63.7	9	0.2	63.3	64.1
Low infant deaths	1568	0.8	74.1	7	0.2	73.7	74.5

1. 기대수명 차이

Low Infant Deaths(74.1세) - High Infant Deaths(63.7세) 약 10.4세의 차이

2. 신뢰구간이 겹치지 않음 → 통계적으로 유의미한 차이

3. 분산 차이

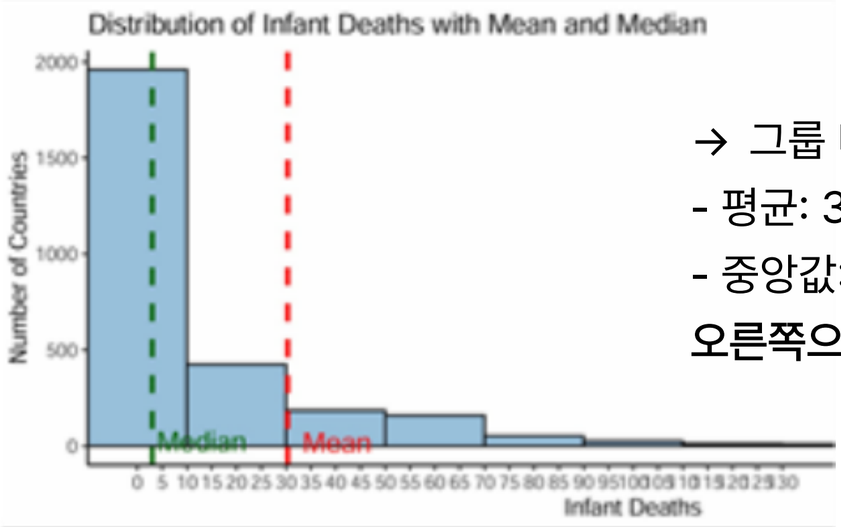
High 그룹 표준편차 9.0, Low 그룹(7.0)보다 큼

High 그룹에서의 기대수명은 Low 그룹에서보다 국가(Country)별로 편차가 큼

4-4. Infant deaths

- 영아 사망 수준 그룹별 평균 영아 사망 수 및 기대수명 , 그룹별 95% 신뢰구간

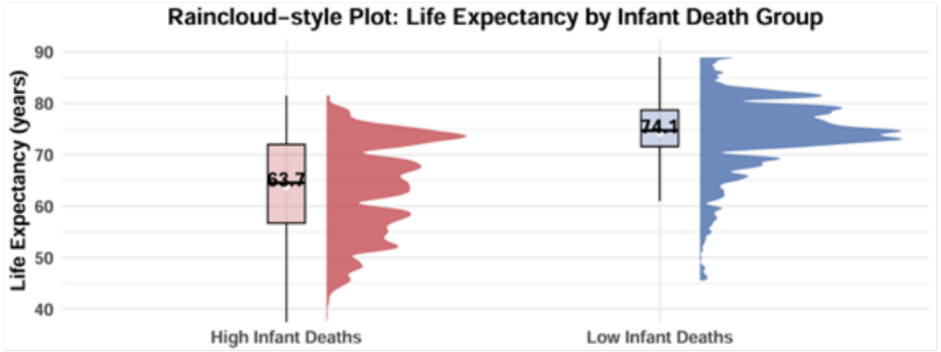
```
# 1. Grouping: Infant deaths, median 기준 Low / High
df <- df %>%
  filter(!is.na(`infant deaths`), !is.na(`Life expectancy`)) %>%
  mutate(Group_InfantDeaths = ifelse(`infant deaths` <= median(`infant deaths`, na.rm = TRUE),
                                     "Low Infant Deaths", "High Infant Deaths"))
```



→ 그룹 나눌때 중앙값 사용,

- 평균: 30.3
- 중앙값: 3.0

오른쪽으로 꼬리가 긴 분포



4-5. Alcohol

- 알코올 소비 그룹별 평균 알코올 소비량, 기대수명 및 그룹별 95% 신뢰구간

Group_ Alcohol	n	Average Alcohol	mean_life	sd_life	se	ci_lower	ci_upper
Low Alcohol	980	0.6	65.8	8.2	0.3	65.2	66.4
Mid Alcohol	979	3.7	67.6	9.7	0.3	67	68.2
High Alcohol	979	9.4	74.3	8.3	0.3	73.7	74.9

1. 기대수명 차이

- High(74.3세) – Mid(67.6세)-Low(65.8세)

  -> 8.5세의 차이

- High Alcohol ↔ Mid/Low 간의 격차 큼

2. 신뢰구간이 겹치지 않음 → 통계적으로 유의미한 차이

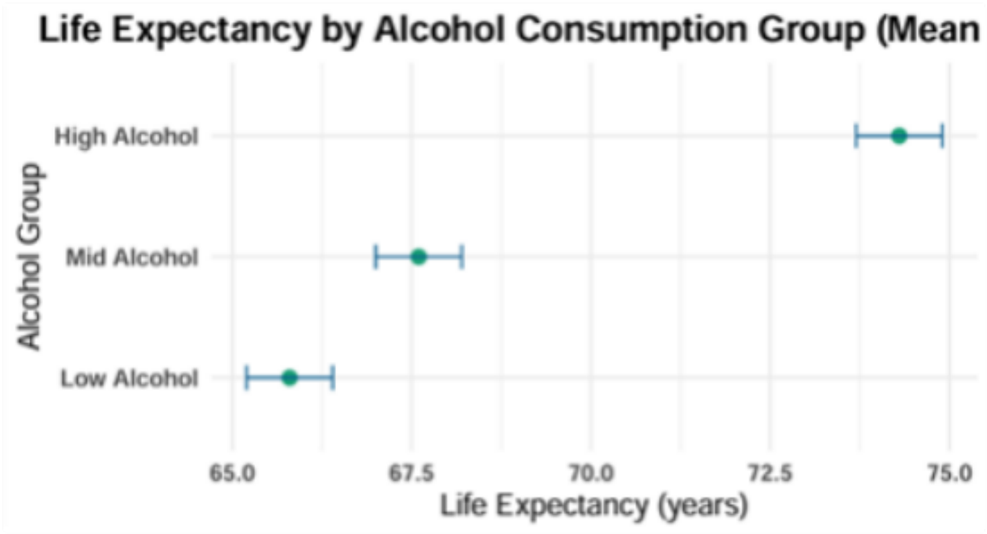
3. 단순히 "알코올 소비량이 많을수록 기대수명이 높다"고 해석 X

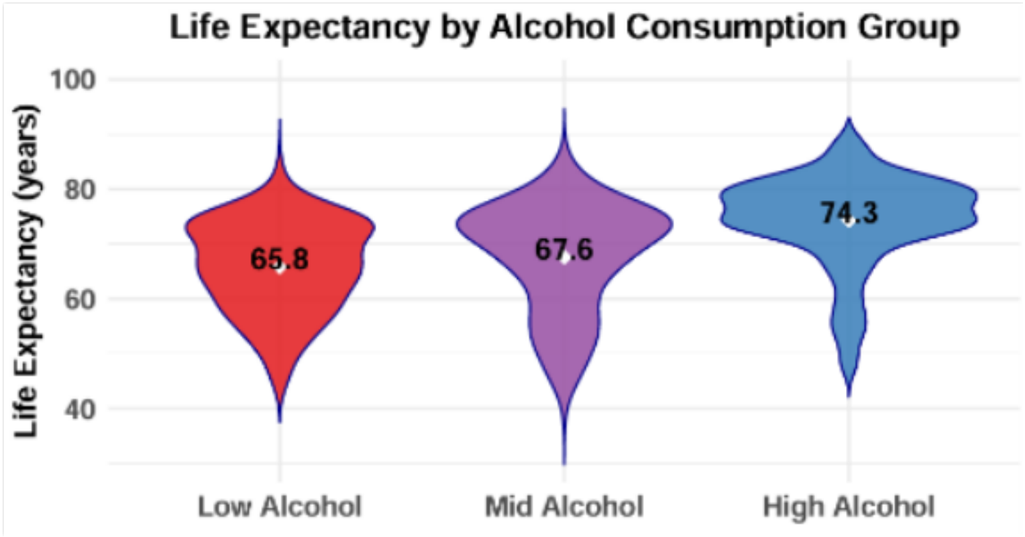
: 고소득 국가들이 의료 시스템이 잘 되어 있고, 음주 소비량도 높아지는 제3의 요인이 개입했을 수 있음.

(= 알코올 소비량은 생활 수준의 간접 지표로 작용)

4-5. Alcohol

- 알코올 소비 그룹별 평균 알코올 소비량, 기대수명 및 그룹별 95% 신뢰구간





5-1. 기대수명 데이터의 주요 변수별 이표본 t-검정

▶ 목적

기대수명에 영향을 미칠 수 있는 주요 변수(GDP, 교육수준, 개발수준, 영아 사망수)에 따라 집단을 2개로 나누고, 두 집단 간 기대수명 평균 차이가 통계적으로 유의한지 이표본 t-검정으로 분석

▶ 가설설정

H_0: 두 집단의 기대수명 평균은 같다.

H_1: 두 집단의 기대수명 평균은 다르다.

▶ 분석

1) 집단 구분

2) 평균 및 신뢰구간 산출

3) 이표본 t-검정 실시

```r
df <- df %>%
  mutate(Group_GDP2 = ifelse(GDP <= median(GDP, na.rm=TRUE), "Low GDP", "High GDP"))
t_gdp <- t.test(`Life expectancy` ~ Group_GDP2, data = df)
print(t_gdp)
if (t_gdp$p.value < 0.05) {
  print("결과: 두 집단(GDP)에 따른 기대수명 차이는 통계적으로 유의합니다 (p < 0.05).")
} else {
  print("결과: 두 집단(GDP)에 따른 기대수명 차이는 통계적으로 유의하지 않습니다 (p >= 0.05).")
}
```

[t검정 R코드]

5-1. 기대수명 데이터의 주요 변수별 이표본 t-검정

```
        Welch Two Sample t-test

data:  Life expectancy by Group_GDP2
t = 27.21, df = 2826.5, p-value < 2.2e-16
alternative hypothesis: true difference in means between group High GDP and group Low GDP is not equal to 0
95 percent confidence interval:
 7.917169 9.146819
sample estimates:
mean in group High GDP  mean in group Low GDP
              73.49093                64.95893
```

```
"결과: 두 집단(GDP)에 따른 기대수명 차이는 통계적으로 유의합니다 (p < 0.05)."
```

[출력값]

- t값

두 집단 평균의 차이가 표준오차(집단 내 변동성) 대비 얼마나 큰지를 나타냄.
t값이 0에서 멀수록 두 집단의 차이가 우연이 아닐 가능성이 높아짐.

- 통계적 결론

- p-value < 2.2e-16로, GDP가 높은 국가와 낮은 국가의 기대수명 평균 차이가 통계적으로 매우 유의하다.
- 귀무가설(H_0): "두 집단의 기대수명 평균이 같다"를 기각하고, GDP가 높은 집단이 기대수명이 더 높다고 결론 내릴 수 있음.

5-2. 변수별 결과 요약

변수명	Low 그룹의 평균	High 그룹의 평균	P-value	유의성	기대수명과의 관계 요약
GDP	64.96	73.49	<2.2e-16	유의함	높을수록 기대수명 높음 (양의 상관관계)
Schooling	63.56	75.58	<2.2e-16	유의함	높을수록 기대수명 높음 (양의 상관관계)
Status	67.12	79.20	<2.2e-16	유의함	선진국이 기대수명 높음
Infant deaths	74.07	63.69	<2.2e-16	유의함	높을수록 기대수명 낮음 (음의 상관관계)

5-3. 기대수명과 가장 관련이 깊은 변수는? - 상관계수

- 상관계수: 두 변수의 선형적 관계의 강도를 나타내는 값으로, 1에 가까울수록 강한 양의 관계,
  -1에 가까울수록 강한 음의 관계를 의미

- 분석 방법

1. 기대수명과 주요 변수(GDP, 교육수준, 개발수준, 영아 사망수) 간의 상관계수를 계산
2. 상관계수의 절대값이 가장 큰 변수를 "가장 관련이 깊은 변수"로 선정

- 분석 결과

```
> print(round(cor_vec, 3))
          GDP     Schooling        Status Infant_deaths
        0.437         0.714         0.482        -0.197
```

→ Schooling의 상관계수 절댓값이 제일 크다. 기대수명과 가장 강한 상관관계를 보임

6-1. 최종 결론

1. 기대수명 분포의 기술통계와 시각화
   → 전체 데이터의 대표값, 산포도, 분포의 특성(왜도, 첨도, 이상치 등)을 파악

2. 소득수준, 개발수준 등 주요 집단별로 기대수명 평균과 신뢰구간을 비교
   → 집단 간 격차가 실제로 존재함을 확인

3. 이표본 t-검정 등 가설검정 → 각 변수별 두 집단 간 기대수명 차이가 통계적으로 유의한지를 검증
   대부분의 변수에서 유의한 차이가 있음을 확인

∴ 기대수명은 단순한 개인의 건강 문제가 아니라, 경제적·사회적 요인(예: GDP, 교육수준, 개발수준, 영아 사망수 등)과 밀접하게 연관되어 있음

6-2. 구성원 소감

이소흔

처음에는 단순히 데이터를 다루는 활동일거라 생각했지만 직접 결측치를 제거하고 기술통계량을 계산해보면서 이 활동들이 데이터의 해석에 얼마나 도움을 주는지 체감할 수 있었습니다. 데이터 수집부터 발표 자료 완성까지의 과정을 경험하면서 통계 분석의 흐름을 이해할 수 있었습니다.

이서후

이번 조사를 통해 단순한 평균이나 수치만으로 파악하는 것보다 왜도, 첨도와 같은 지표를 통한 분석의 중요성을 실감할 수 있었다. 특히 GDP나 인구 변수 처럼 심하게 치우친 변수들은 평균만으로는 데이터를 해석할 수 없고 왜도와 첨도같은 지표가 필요함을 느꼈다.

6-2. 구성원 소감

정세연

알코올 소비량과 기대수명 간의 관계를 분석하며 내가 기존에 가지고있던 직관과는 다르게 결과가 나타나 흥미로웠다. 이를 통해 상관관계가 반드시 인과관계를 의미하지 않음을 체감할 수 있었다. 교수님께서 수업 시간에 자주 예로 드셨던 상어 출몰과 아이스크림 판매량의 관계가 떠오르기도 했다. 또한 영아 사망률이 낮은 국가일수록 기대수명이 높고, 분산도 적다는 점을 통해 보건 수준의 중요성도 실감했다.
이번 발표는 통계학 복수전공을 시작한 이후 처음 참여하는 팀 프로젝트이자 발표였다. 솔직히 복수전공을 하면서도 본 전공과의 괴리때문에 "계속 복수전공을 이어가는게 맞을까?" 하는 고민도 있었지만, 이번 기회를 통해 데이터를 다루는 재미를 느낄 수 있었다. 직접 분석하고 시각화하며 통계학에 대한 흥미가 한 층 더 커진 거 같다.

박서경

분석 주제를 어떻게 정하느냐에 따라 전체 연구의 방향성과 결과 해석이 크게 달라지기 때문에 데이터 분석에서 주제 수집과 선정의 중요성을 다시 한 번 느꼈습니다. 이표본 t-검정 등 실제 통계 방법을 직접 적용해보며 결과를 해석하는 과정이 재미있었습니다.

감사합니다