

기초 통계 분석

1. 기초통계량

평균 (Mean)

- **정의:** 평균은 데이터 집합의 모든 값의 합을 데이터의 개수로 나눈 값이다. 이는 데이터의 중심 위치를 나타내며, 데이터 집합의 중앙값을 설명한다.

- **계산 방법**

예를 들어, 데이터 집합이 $\{x_1, x_2, \dots, x_n\}$ 라면 평균 \bar{x} 는 다음과 같이 계산된다.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

여기서 n 은 데이터의 개수, x_i 는 각 데이터 값이다.

- **특징**

평균은 데이터의 모든 값을 고려하여 계산되므로, 극단값(이상치)의 영향을 많이 받는다.

중앙값 (Median)

- **정의:** 데이터 집합을 크기 순서대로 정렬했을 때 중앙에 위치한 값
데이터가 홀수 개일 때는 정확히 중앙에 있는 값이고, 짝수 개일 때는 중앙에 있는 두 값의 평균이다.
- **계산 방법:**
데이터를 크기 순서대로 정렬한 후, 데이터 개수가 n 인 경우:
 - 홀수 개일 때: 중앙값 = 정렬된 데이터의 $(n+1)/2$ 번째 값
 - 짝수 개일 때: 중앙값 = 정렬된 데이터의 $n/2$ 번째 값과 $n/2 + 1$ 번째 값의 평균
- **특징:**
중앙값은 극단값의 영향을 받지 않으므로, 데이터에 이상치가 포함된 경우에도 데이터의 중심을 잘 나타낸다.

최빈값 (Mode)

- **정의:** 최빈값은 데이터 집합에서 가장 자주 나타나는 값이다. 데이터에 따라 하나 이상의 최빈값이 존재할 수 있다.
- **계산 방법:**
각 값의 빈도를 계산하여, 빈도가 가장 높은 값을 최빈값으로 선택합니다.
- **특징:** 최빈값은 데이터의 분포를 이해하는 데 유용하며, 특히 명목형 데이터(순서를 매길 수 없는 데이터, 예를 들면, 성별, 종교, 색깔 등, <https://wikidocs.net/161040#nominal>)에서 자주 사용된다.

분산 (Variance)

- **정의:** 분산은 데이터가 평균에서 얼마나 떨어져 있는지를 나타내는 값으로, 데이터의 흩어짐 정도를 나타낸다. 이는 평균에서 각 데이터 값의 편차의 제곱의 평균으로 계산된다.

- **계산 방법:**

- 데이터 집합 $\{x_1, x_2, \dots, x_n\}$ 에 대해 분산 σ^2 는 다음과 같이 계산된다.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- **특징:** 분산은 단위가 원래 데이터의 단위의 제곱이므로, 데이터의 변동성을 직접적으로 비교하기에는 어려움이 있다.

표준편차 (Standard Deviation)

- **정의:** 표준편차는 분산의 제곱근으로, 데이터의 산포도를 나타내는 값이다. 이는 데이터가 평균에서 얼마나 떨어져 있는지를 직관적으로 이해할 수 있도록 해준다.

- **계산 방법:**

분산 σ^2 의 제곱근으로 계산된다.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- **특징:** 표준편차는 원래 데이터와 같은 단위를 가지므로, 데이터의 변동성을 이해하고 비교하는 데 유용하다.

1. 기초통계량

이와 같이 기초통계량은 데이터의 중심 경향과 변동성을 이해하는 데 중요한 지표들을 제공한다.

각 통계량의 계산 방법과 특징을 잘 이해하면, 데이터 분석의 기본기를 탄탄히 다질 수 있다.

2. 상관 관계 분석

2. 상관 관계 분석

상관관계란?

상관관계는 두 변수 간의 관계를 나타내는 개념으로, 한 변수의 변화가 다른 변수의 변화와 어떻게 관련되어 있는지를 보여줌. 상관관계는 양의 상관관계와 음의 상관관계로 구분되며, 이를 통해 변수 간의 관계를 정량적으로 파악할 수 있다.

•**양의 상관관계:** 한 변수가 증가할 때 다른 변수도 증가하는 경향을 보이는 관계를 의미함. 예를 들어, 공부 시간과 성적 사이에는 양의 상관관계가 존재할 수 있다.

•**음의 상관관계:** 한 변수가 증가할 때 다른 변수는 감소하는 경향을 보이는 관계를 의미함. 예를 들어, 스트레스 수준과 수면 시간 사이에는 음의 상관관계가 존재할 수 있다.

2. 상관 관계 분석

상관계수 (Correlation Coefficient)

상관계수는 두 변수 간의 상관관계의 정도를 나타내는 값으로, -1에서 1 사이의 값을 가짐. 상관계수가 1에 가까울수록 강한 양의 상관관계를, -1에 가까울수록 강한 음의 상관관계를 나타냄. 0에 가까울수록 상관관계가 없음을 의미한다.

• **피어슨 상관계수 (Pearson Correlation Coefficient):** 두 변수 간의 선형 관계를 측정함. 연속형 변수 간의 상관관계를 분석할 때 주로 사용하며, 다음과 같이 계산한다

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

• 여기서 r 값이 1에 가까울수록 강한 양의 상관관계, -1에 가까울수록 강한 음의 상관관계를 의미한다.

2. 상관 관계 분석

스피어만 상관계수 (Spearman Correlation Coefficient):

두 변수 간의 비선형 관계를 측정함. 순위형 변수 또는 비정규 분포를 가진 변수 간의 상관관계를 분석할 때 사용하며, 다음과 같이 계산한다

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

여기서 d_i 는 각 관찰치의 순위 차이, n 은 데이터의 개수임. 스피어만 상관계수는 데이터가 순위형일 때나 비선형 관계가 있을 때 유용하다

상관관계 분석의 목적

상관관계 분석은 변수 간의 관계를 이해하고 예측 모델을 개발하는 데 중요한 역할을 한다.

예를 들어, 마케팅에서 광고비와 판매량 간의 상관관계를 분석하여 광고비 증가가 판매량에 미치는 영향을 파악할 수 있다.

또한, 금융 분야에서는 주식 간의 상관관계를 분석하여 포트폴리오를 최적화할 수 있다. 이를 통해 변수 간의 관계를 이해하고, 더 나은 의사결정을 내릴 수 있다.

상관관계와 인과관계의 차이

상관관계는 두 변수 간의 관계를 나타내지만, 인과관계를 증명하지는 못한다. 상관관계가 있다고 해서 한 변수가 다른 변수의 원인이라고 단정할 수 없다.

예를 들어, 아이스크림 판매량과 익사 사고 건수 간에 상관관계가 있을 수 있지만, 이는 여름철이라는 공통 요인이 원인일 가능성이 크다. 따라서, 상관관계를 인과관계로 해석할 때는 주의가 필요하다.

인과관계를 확인하려면 추가적인 실험이나 연구가 필요하다.

상관관계 분석의 한계

상관관계 분석은 변수 간의 관계를 파악하는 데 유용하지만, 몇 가지 한계가 있다.

첫째, 상관관계는 선형 관계만을 나타내므로 비선형 관계를 포착하지 못할 수 있다.

둘째, 이상치의 영향을 받을 수 있다.

마지막으로, 상관관계는 인과관계를 나타내지 않으므로, 변수 간의 원인과 결과를 분석하려면 추가적인 연구가 필요하다.

이러한 한계를 인지하고 분석을 진행해야 신뢰성 있는 결과를 도출할 수 있다.

2. 상관 관계 분석

상관관계 분석의 예시

1. 광고비와 판매량

광고비와 판매량 간의 상관관계를 분석하여 광고비가 판매량에 미치는 영향을 파악할 수 있다. 이를 통해 마케팅 전략을 수립할 때 유용한 정보를 얻을 수 있다.

2. 키와 체중

사람들의 키와 체중 간의 상관관계를 분석하여 키가 클수록 체중도 증가하는 경향이 있는지 확인할 수 있다. 이는 건강 및 체력 관리에 중요한 정보를 제공할 수 있다.

3. 학습 시간과 성적

학생들의 학습 시간과 시험 성적 간의 상관관계를 분석하여 학습 시간이 길수록 성적이 높아지는지 확인할 수 있다. 이를 통해 효과적인 학습 전략을 개발할 수 있다.

요약

상관관계 분석은 다양한 분야에서 변수 간의 관계를 이해하고 예측하는데 중요한 도구로 사용된다. 상관계수를 통해 변수 간의 관계의 강도와 방향을 정량적으로 파악할 수 있으며, 이를 기반으로 다양한 의사결정에 활용할 수 있다.

3. 탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA - Exploratory Data Analysis)란?

탐색적 데이터 분석(EDA)은 데이터셋의 주요 특성을 요약하고 시각화하는 초기 분석 단계임. 이를 통해 데이터의 구조를 이해하고, 패턴을 발견하며, 가설을 수립하는 데 도움을 줄 수 있다. EDA는 데이터 분석 과정에서 중요한 첫 단계로, 데이터의 질을 평가하고 분석 방향을 설정하는 데 유용하다.

데이터 시각화

데이터 시각화는 EDA의 핵심 요소로, 데이터를 시각적으로 표현하여 패턴, 이상치, 관계 등을 쉽게 파악할 수 있게 해준다. 주요 시각화 기법은 다음과 같다.

- **히스토그램 (Histogram)**: 데이터의 분포를 시각화하는 데 사용됨. 연속형 변수의 분포와 빈도수를 파악할 수 있다.
- **상자 그림 (Box Plot)**: 데이터의 중앙값, 사분위수, 이상치를 시각화하는 데 사용됨. 데이터의 분포와 변동성을 한눈에 확인할 수 있다.
- **산점도 (Scatter Plot)**: 두 변수 간의 관계를 시각화하는 데 사용됨. 변수 간의 상관관계를 파악할 수 있다.
- **막대 그래프 (Bar Chart)**: 범주형 데이터의 빈도를 시각화하는 데 사용됨. 각 범주의 빈도나 비율을 비교할 수 있다.
- **선 그래프 (Line Chart)**: 시간에 따른 데이터의 변화를 시각화하는 데 사용됨. 시간 시계열 데이터를 분석할 때 유용하다.

3. 탐색적 데이터 분석

데이터 요약

데이터 요약은 데이터의 주요 특성을 요약하는 과정으로, 주요 통계량을 계산하여 데이터를 이해하는 데 도움을 줍니다. 주요 요약 통계량은 다음과 같다.

- **평균 (Mean):** 데이터의 중심 경향을 나타내는 값으로, 데이터의 모든 값을 더한 후 개수로 나눈 값이다.
- **중앙값 (Median):** 데이터의 중앙에 위치한 값으로, 데이터의 분포를 이해하는 데 유용하다.
- **최빈값 (Mode):** 데이터에서 가장 자주 나타나는 값이다.
- **분산 (Variance):** 데이터가 평균에서 얼마나 떨어져 있는지를 나타내는 값으로, 데이터의 흩어짐 정도를 파악할 수 있다.
- **표준편차 (Standard Deviation):** 분산의 제곱근으로, 데이터의 변동성을 나타내는 값이다.

데이터 클렌징(Data Cleansing)

데이터 클렌징은 분석을 위해 데이터를 **정제**하는 과정으로, 결측값 처리, 이상치 탐지, 중복 데이터 제거 등을 포함함. 데이터 클렌징은 데이터의 정확성과 일관성을 높이는 데 중요하다.

- **결측값 처리**: 결측값을 제거하거나 대체하여 분석의 정확성을 높일 수 있다. 결측값을 대체하는 방법으로는 평균, 중앙값, 최빈값 등을 사용할 수 있다.
- **이상치 탐지**: 데이터에서 정상 범위를 벗어난 값을 식별하고 처리하여 분석의 정확성을 높일 수 있다. 이상치를 제거하거나 다른 값으로 대체할 수 있다.
- **중복 데이터 제거**: 동일한 데이터가 중복되어 있는 경우 이를 제거하여 데이터의 일관성을 유지할 수 있다.

데이터 변환

데이터 변환은 분석을 위해 데이터를 변형하는 과정으로, 스케일링, 로그 변환, 원-핫 인코딩 등이 포함됨. 이를 통해 데이터의 특성을 변환하여 분석을 용이하게 할 수 있다.

- **스케일링 (Scaling)**: 데이터의 범위를 조정하여 모델의 성능을 향상시킬 수 있다. 주로 Min-Max 스케일링과 표준화(정규화)를 사용한다.
- **로그 변환 (Log Transformation)**: 데이터의 분포를 정규 분포에 가깝게 만들기 위해 로그 변환을 사용할 수 있다. 이는 특히 분포가 한쪽으로 치우쳐 있을 때 유용하다.
- **원-핫 인코딩 (One-Hot Encoding)**: 범주형 데이터를 이진 변수로 변환하여 분석을 용이하게 할 수 있다. 이는 주로 머신러닝 모델에 범주형 데이터를 입력할 때 사용한다.

데이터 분석의 목적

EDA의 목적은 데이터의 특성을 이해하고, 데이터의 구조와 관계를 파악하며, 추가 분석을 위한 기초를 마련하는 데 있음. 이를 통해 데이터의 질을 평가하고, 이상치를 식별하며, 가설을 수립할 수 있다. EDA는 데이터 분석의 첫 단계로, 데이터에 대한 깊은 이해를 바탕으로 효과적인 분석을 수행할 수 있다.

EDA의 중요성

EDA는 데이터 분석 과정에서 매우 중요한 단계로, 데이터의 질을 평가하고 분석 방향을 설정하는 데 도움을 줌. 이를 통해 데이터의 특성을 이해하고, 추가 분석을 위한 기초를 마련할 수 있다. EDA를 통해 데이터의 패턴을 발견하고, 가설을 수립하며, 데이터의 구조를 파악할 수 있다.

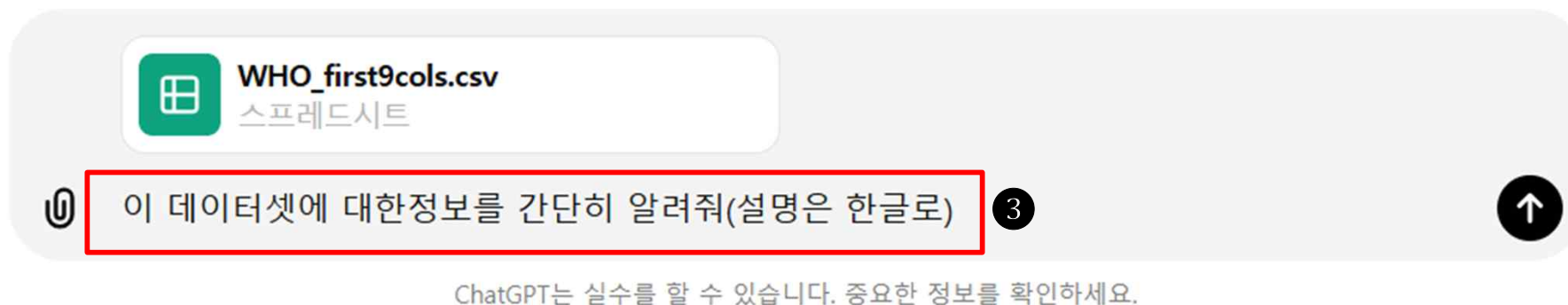
이와 같이 탐색적 데이터 분석(EDA)은 데이터의 특성을 이해하고 분석 방향을 설정하는 데 중요한 역할을 함. 다양한 시각화 기법과 요약 통계량을 통해 데이터를 이해하고, 데이터 클렌징과 변환을 통해 분석의 정확성을 높일 수 있다. 이를 통해 효과적인 데이터 분석을 수행할 수 있다.

4. 로 데이터 탐색 실습

4. ChatGPT로 데이터 탐색 실습

WHO 국가 복지 데이터셋 사용

- 데이터셋 로드 및 확인: 데이터셋을 불러오고 기본 정보를 확인.



4. ChatGPT로 데이터 탐색 실습

- 데이터셋 소개: WHO 국가 복지 데이터셋의 내용과 구조.

WHO First9cols

	Country	CountryID	Continent	Adolescent fertility
1	Afghanistan	1	1	151
2	Albania	2	2	27
3	Algeria	3	3	6
4	Andorra	4	2	
5	Angola	5	3	146
6	Antigua and	6	4	

이 데이터셋에 대한정보를 간단히 알려줘(설명은 한글로)

- 이 데이터셋은 총 202개의 행과 9개의 열로 구성되어 있으며, 일부 열에는 결측치가 존재한다. 주요 지표로는 청소년 출산율, 성인 문해율, 1인당 국민 총소득, 초등학교 입학율, 총 인구 등이 포함된다

• 데이터셋 소개: WHO 국가 복지 데이터셋의 내용과 구조.

- **Country:** 국가 이름
- **CountryID:** 국가 식별자
- **Continent:** 대륙 코드
- **Adolescent fertility rate (%)**: 청소년 출산율 (백분율)
- **Adult literacy rate (%)**: 성인 문해율 (백분율)
- **Gross national income per capita (PPP international \$)**: 1인당 국민 총소득 (PPP 국제 달러 기준)
- **Net primary school enrolment ratio female (%)**: 여자 초등학교 순 입학율 (백분율)
- **Net primary school enrolment ratio male (%)**: 남자 초등학교 순 입학율 (백분율)
- **Population (in thousands) total**: 총 인구 (천 명 단위)

WHO 국가 복지 데이터셋 사용

기초 통계값 산출 실습

기초 통계값을 알려줘

- **평균 계산:** 데이터셋의 특정 변수에 대한 평균값 계산.
- **분산 계산:** 데이터셋의 특정 변수에 대한 분산값 계산.
- **표준편차 계산:** 데이터셋의 특정 변수에 대한 표준편차 계산.

4. ChatGPT로 데이터 탐색 실습

WHO 국가 복지 데이터셋 사용

기초 통계값 산출 실습

기초 통계값을 알려줘

Basic Statistics Of WHO Dataset			
		Country	CountryID
1	count	202	202.0
2	unique	202	
3	top	Afghanistan	
4	freq	1	
5	mean		101.5

4. ChatGPT로 데이터 탐색 실습

데이터셋의 기초 통계값은 다음과 같습니다:

- **Country:** 202개의 고유 국가명
- **CountryID:** 최소 1, 최대 202, 평균 101.5
- **Continent:** 최소 1, 최대 7, 평균 3.579
- **Adolescent fertility rate (%):** 최소 0, 최대 199, 평균 59.46
- **Adult literacy rate (%):** 최소 23.6, 최대 99.8, 평균 78.87
- **Gross national income per capita (PPP international \$):** 최소 260, 최대 60870, 평균 11250.11
- **Net primary school enrolment ratio female (%):** 최소 6, 최대 100, 평균 84.03
- **Net primary school enrolment ratio male (%):** 최소 11, 최대 100, 평균 85.70
- **Population (in thousands) total:** 최소 2, 최대 1,328,474, 평균 34,099.64



4. ChatGPT로 데이터 탐색 실습

통계 요약(표를 다운로드 후 엑셀로 열어본다)

	Country	CountryID	Continent	Adolescer	Adult liter	Gross nati	Net prima	Net prima	Population (in thousands) total		
count	202	202	202	177	131	178	179	179	189		
unique	202										
top	Afghanistan										
freq	1										
mean		101.5	3.5792079	59.457627	78.871755	11250.112	84.033519	85.698324	34099.64021	164021	
std		58.456536	1.8082627	49.10529	20.415760	12586.753	17.788046	15.451211	131837.70867	703325	
min		1	1	0	23.6	260	6	11	2		
25%		51.25	2	19	68.4	2112.5	79	79.5	1328		
50%		101.5	3	46	86.5	6175	90	90	6640		
75%		151.75	5	91	95.300000	14502.5	96	96	20971		
max		202	7	199	99.8	60870	100	100	1328474		

4. ChatGPT로 데이터 탐색 실습

성인 문해율 가장 낮은 나라는?



성인 문해율이 가장 낮은 나라는 부르키나파소(Burkina Faso)이며, 성인 문해율은 23.6%입니다. [↗]



성인 문해율 가장 높은 나라는?



성인 문해율이 가장 높은 나라는 쿠바(Cuba)이며, 성인 문해율은 99.8%입니다. [↗]



감사합니다