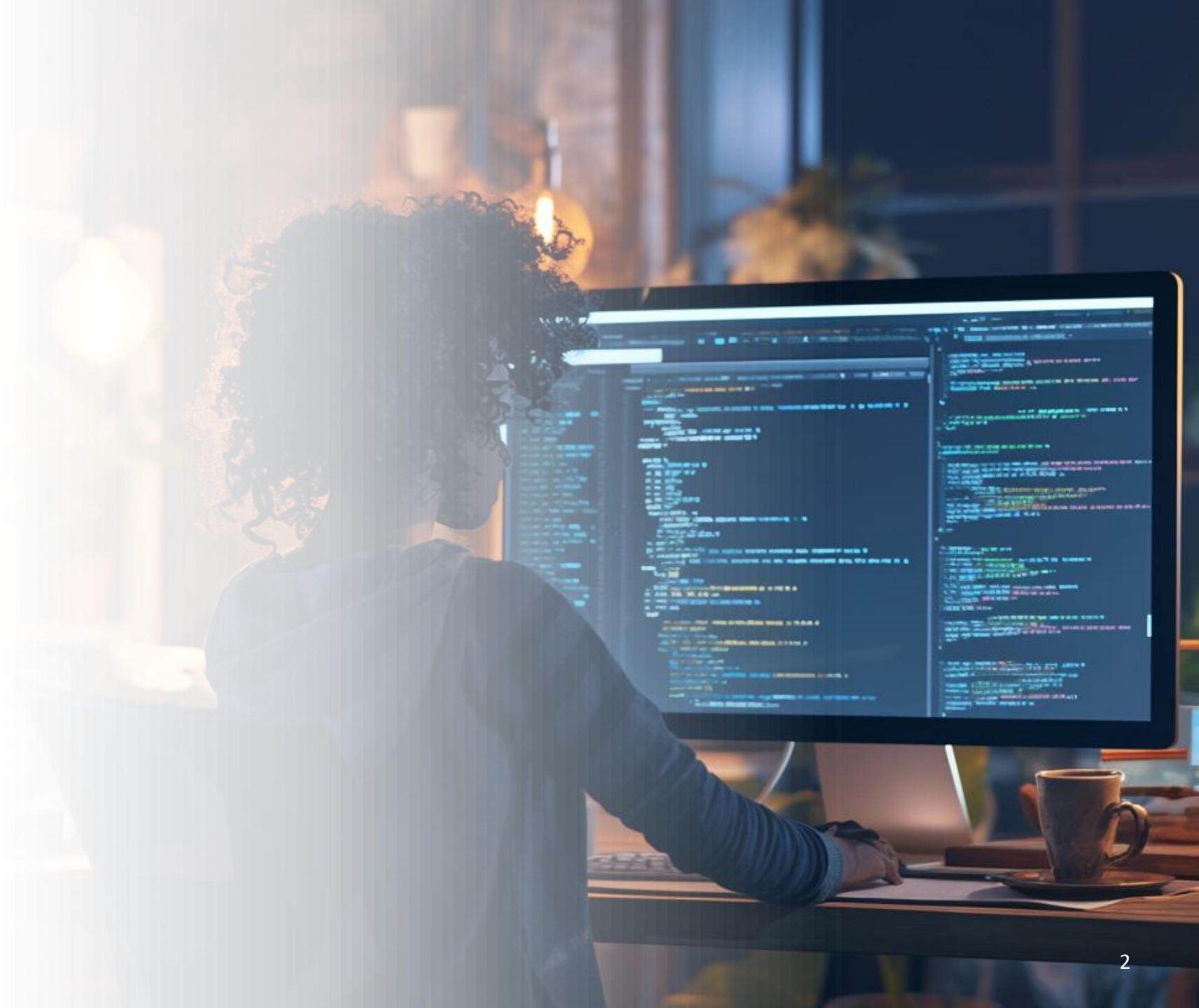
통계응용

2주차

담당 : 안상선 (logix44@hanmail.net)

주요 내용

- 표본과 통계량
- 자료의 정리
- 분포의 특성
- 데이터(자료)의 시각화



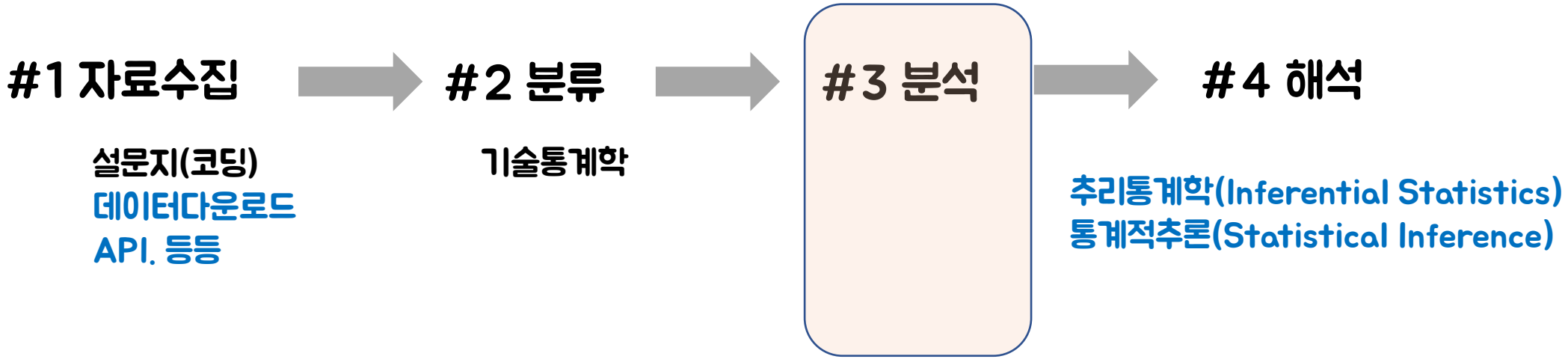
1. 표본과 통계량

1. 표본과 통계량

1) 통계학의 개념 및 체계

- 불확실한 상황에서 현명한 의사결정을 하기위한 이론과 방법의 체계
  - 예를 들어 전수조사는 막대한 비용과 많은 시간이 소요되기 때문에 대부분의 경우 조사계획을 철저히 수립해서 **표본조사**를 사용한다.

- 통계학의 체계



1. 표본과 통계량

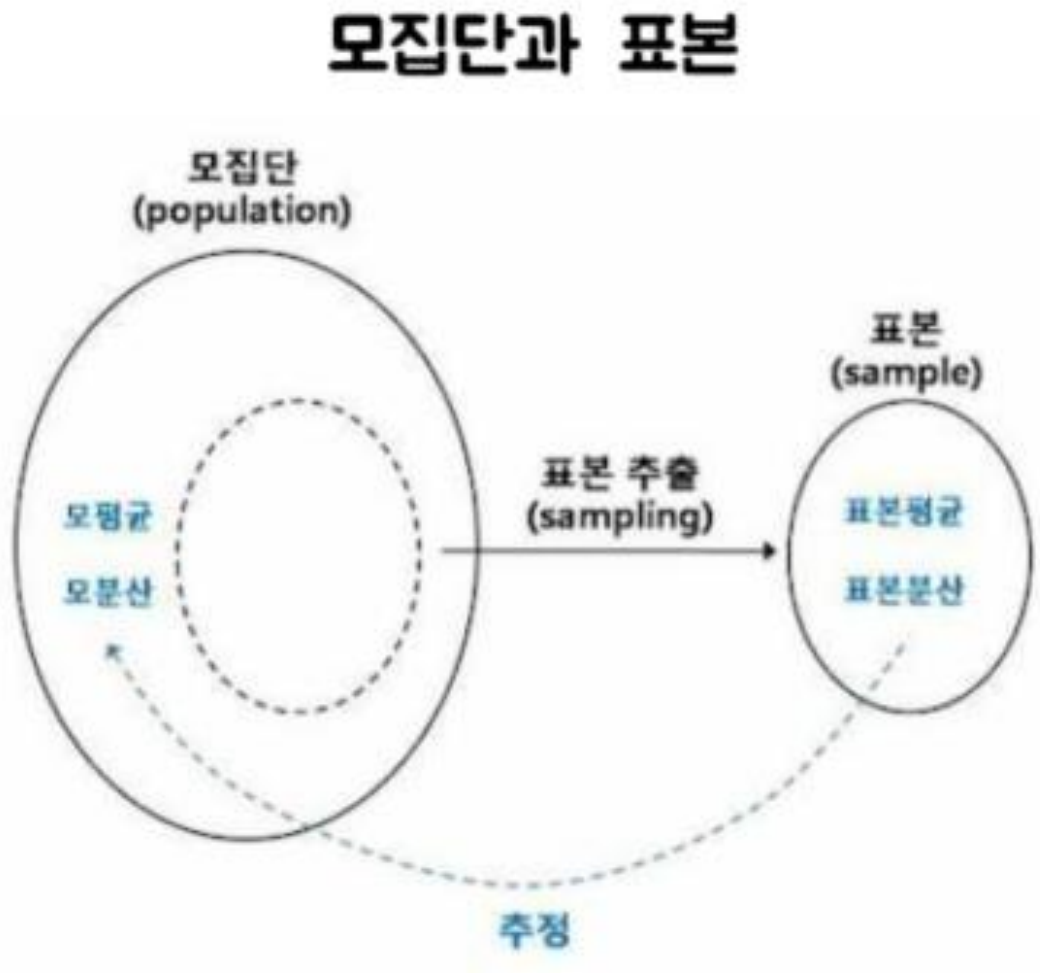
2) 통계학의 용어 : 모수와 통계량

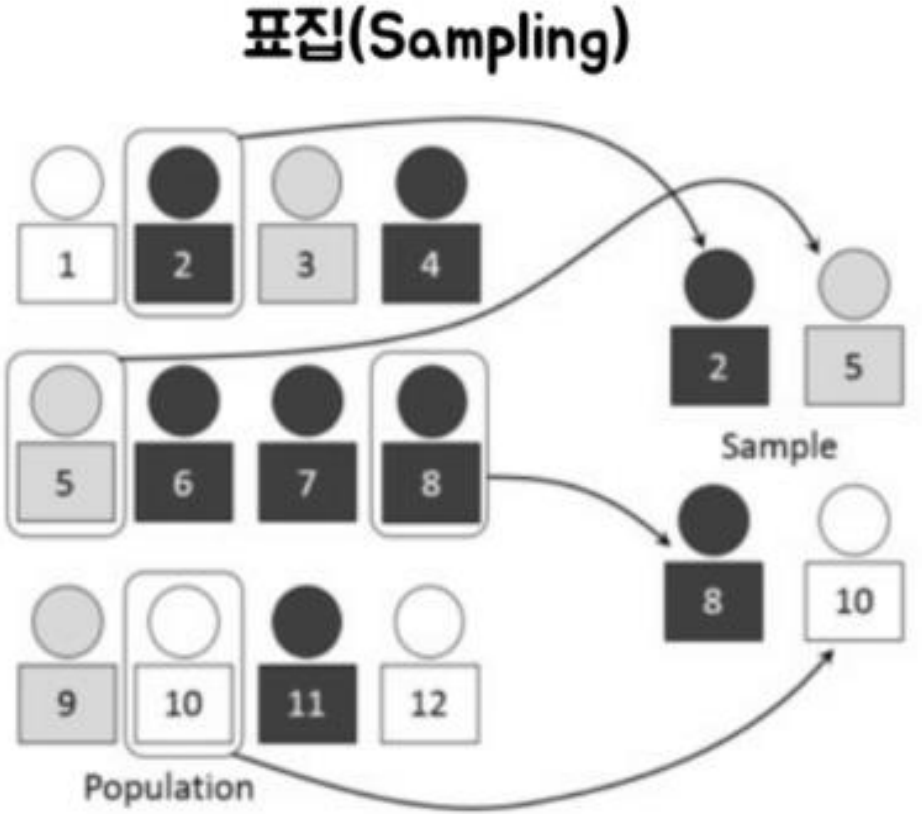
- 모수(Parameter)
  - 모집단의 특성을 수치로 나타낸 것

- 통계량(Statistics)
  - 표본의 특성을 수치로 나타낸 것으로 표본의 특성을 나타내는 확률변수

1. 표본과 통계량

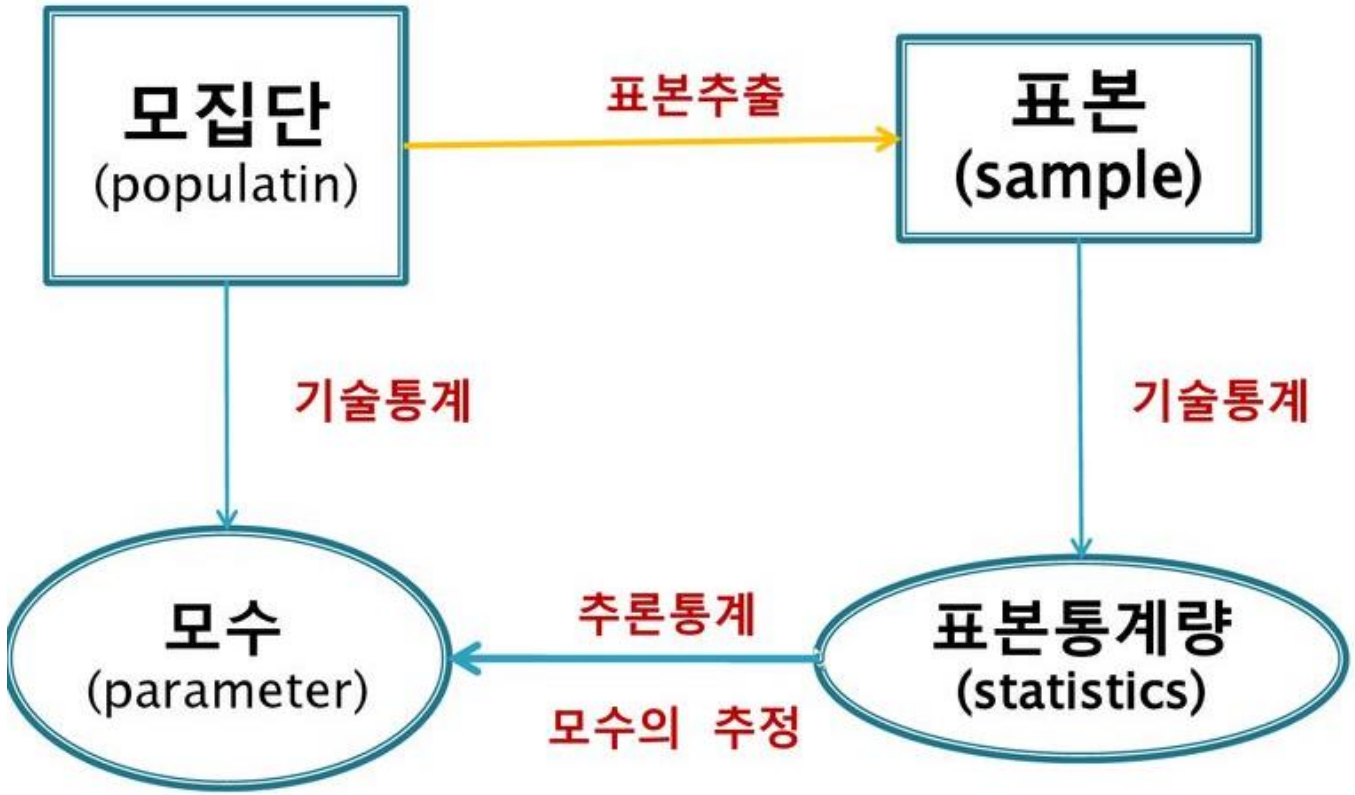
2) 통계학의 용어 : 모수와 통계량





3. 강의 페이지 안내

2) 통계학의 용어 : 모수와 통계량



출처 : https://slidesplayer.org/slide/14377701/

1. 표본과 통계량

2) 통계학의 용어 : 모수와 통계량

- 표본 통계량을 가지고 **모수를 추정**한다.

  · 결국 우리가 하려는 것은 표본 통계량과 모수가 일치하는지 여부임

통계적 특성	모수	통계량
개 체 수	N	n
평 균	μ	$\bar{x}$
분 산	σ^2	s^2
표 준 편 차	σ	s
표 준 오 차	$\sigma_{\bar{x}}$	$s_{\bar{x}}$
상 관 계 수	ρ	r
회 귀 계 수	β	B
※ 모수와 통계량을 나타내는 기호들		

1. 표본과 통계량

2) 통계학의 용어 : 모수와 통계량

- 표본 통계량을 가지고 **모수를 추정**한다.

  · 결국 우리가 하려는 것은 표본 통계량과 모수가 일치하는지 여부임 : **모평균 가설검정**

통계적 특성	모수	통계량
개 체 수	N	n
평 균	μ	$\bar{x}$
분 산	σ^2	s^2
표 준 편 차	σ	s
표 준 오 차	$\sigma_{\bar{x}}$	$s_{\bar{x}}$
상 관 계 수	ρ	r
회 귀 계 수	β	B
※ 모수와 통계량을 나타내는 기호들		

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

1. 표본과 통계량

2) 통계학의 용어 : 모수와 통계량

유한모집단
(Finite population)

모집단을 구성하는 전체 관측치의 수가 제한되어 있으므로 모집단의 전체를 쉽게 파악할 수 있는 모집단

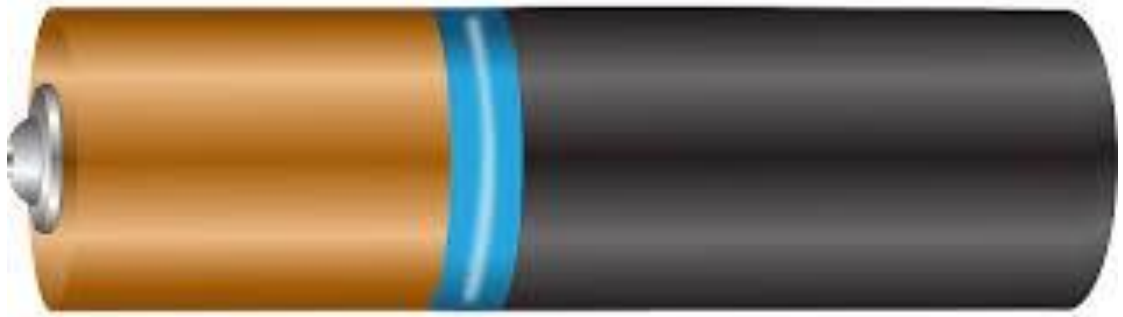
무한모집단
(Infinite population)

모집단을 구성하는 전체 관측치의 수가 무한하거나 그 수가 제한적이라 하더라도 실제적으로 전체 관측치의 조사가 불가능한 모집단

1. 표본과 통계량

3) 추론 통계학

- 어느 한 건전지의 평균 수명은 300시간으로 알려져 있는데, 일부에서는 300시간이 아니라는 의견이 나오고 있다. 그래서 건전지 25개를 무작위로 뽑아서 표본을 만들고 이를 조사했더니, 평균수명이 310시간이 나왔고, 표준편차는 30시간이라고 한다. 평균수명이 300시간인지 가설 검정하시오.



1. 표본과 통계량

4) 추론 통계학과 확률변수

- 왜 통계량(Statistics)가 **확률변수**일까?

  * 일정한 확률을 갖고 발생하는 임의의 사건에 수치를 부여하는 변수

어느 날 L교수는 자신의 건강이 걱정됐다.

작은 병　　큰 병　　이상 없음

1. 모수와 통계량

4) 추론 통계학과 확률변수

- 일정한 확률을 갖고 발생하는 임의의 사건에 수치를 부여하는 변수

어느 날 L교수는 자신의 건강이 걱정됐다.

이상 없음 (S1)	작은 병 (S2)	큰 병 (S3)
30%	40%	30%

1. 모수와 통계량

4) 추론 통계학과 확률변수

- 왜 통계량(Statistics)가 **확률변수**일까?

  (사례)

  모집단 : 1부터 10까지의 숫자로 모집단의 평균은 5.5

  표본 : 모집단에서 3개를 뽑은 수치의 평균

  표본의 평균 값이 나온다면, 이 값의 특성은 어떠할까?

  : 표본의 평균이 모집단의 평균과 같은 확률은 얼마나 될까?

  즉, 확률의 문제라고 할 수 있음

2. 자료의 정리

2. 자료의 정리

1) 변수와 자료

- 변수(Variable) : 연구자의 관심대상이 되는 성격 or 속성
- 자료(Data) : 이러한 변수를 관찰하여 기록한 결과

한 집단의 특성을 쉽게 알아보고 분석하기 위해서는 **수집된 자료를 의미 있는 모양**으로 분류· 정리하는 것이 중요하다.

2. 자료의 정리

2) 자료의 종류

- 범주형 자료는 질적 특성을 나타내는 자료로 개체를 분류하는 데 사용되는 반면, 수치형 자료는 양적 특성을 나타내며 측정이나 계산이 가능한 값을 갖는다. 범주형 자료는 통계적 분석에 제한이 있지만, 수치형 자료는 다양한 수학적 연산과 통계 분석이 가능하다
- 명목자료는 단순히 범주를 구분하는 데 사용되며 범주 간 순서나 크기 비교가 불가능한 반면, 순서변수는 범주 간 순서를 갖지만 간격의 크기는 정확히 측정할 수 없다. 명목자료는 성별이나 혈액형 등이 해당되고, 순서변수는 학력 수준이나 만족도 등이 해당된다.

[데이터 분석에서 사용하는 변수의 유형(Type)]

- **범주형 자료(질적, 비계량)**
  - **명목변수(Nominal Variable)** : 개체나 사람의 특성만 (성별, 종교 등)
  - **순서(서열)변수(Ordinal Variable)** : 측정대상 간 선호를 부여 (선호도, 만족도)
- **양적인 자료(양적, 계량)**
  - **등간변수(Interval Variable)** : 측정대상 간의 순서 + 값 사이의 간격이 일정 (IQ,온도)
    절대 영점이 없음(100도는 50도의 2배 뜨겁다?)
  - **비율변수(Ratio Variable)** : 측정대상 간에 비율 계산이 가능함(연령, 무게, 거리, 시간)
  - **이산형 변수(Discrete Variable)** : 점수, 빈도수
  - **연속형 변수(Continuous Variable)** : 실수, 키, 몸무게

2. 자료의 종류

3) 자료의 특성

- 관찰된 자료의 분포를 나타내는 특성

· 집중화경향(Central Tendency)

: 관찰된 자료가 어디에 집중돼 있는지 나타낸 것

· 분산도(Dispersion)

: 관찰된 자료의 흩어진 정도 (범위, 분산, 표준편차, 변동계수)

· 비대칭도(Skewness)

: 관찰된 값이 어디에 치우쳐져 있는가? (비대칭도, 왜도, 첨도)

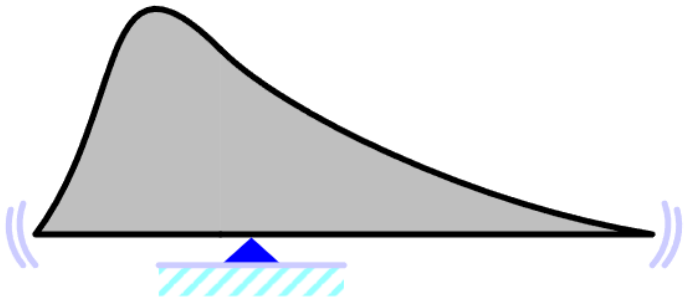
3. 분포의 특성

3. 분포의 특성

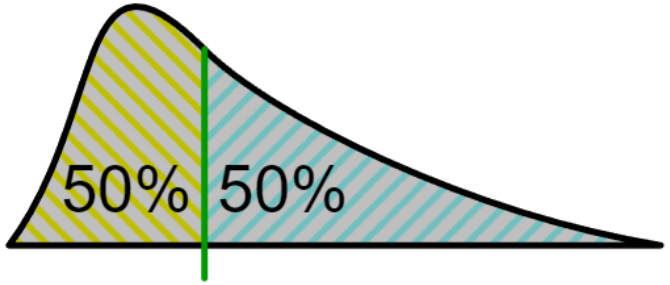
1) 데이터의 집중화 경향

- 데이터의 집중화 경향은 데이터 값들이 중심점 또는 특정 범위에 얼마나 밀집되어 있는지를 나타냄. 예를 들어, 학생들의 시험 점수가 70-80점 사이에 많이 몰려 있다면, 이 범위에 데이터가 집중되어 있다고 볼 수 있음. 이러한 집중화 경향을 분석함으로써, 데이터의 대표 값이 무엇인지, 대부분의 값들이 어느 범위에 속하는지를 파악할 수 있음.
- 집중화 경향을 측정하는 주요 통계지표로는 평균, 중앙값, 최빈값이 있음. 평균(Mean)은 모든 데이터 값의 합을 데이터 수로 나눈 값으로, 전체 데이터의 산술적 중심을 나타냄. 중앙값(Median)은 데이터를 크기순으로 나열했을 때 가운데 위치한 값으로, 극단 값의 영향을 덜 받음. 최빈값(Mode)는 데이터에서 가장 자주 나타나는 값으로, 범주형 데이터에서 유용하게 쓰임.

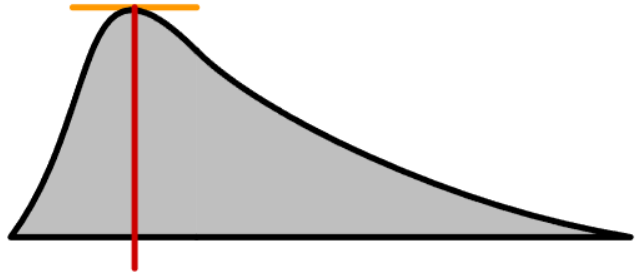
데이터의 집중화 경향 통계 지표



평균(Mean)



중앙값(Median)



최빈값(Mode)

3. 분포의 특성

2) 데이터의 집중화 경향 지표

- 평균(Mean)은 모든 데이터 값의 합을 데이터의 개수로 나눈 것으로 이는 데이터의 전체적인 경향을 잘 나타내며, 모든 데이터 포인트를 고려한다는 장점이 있음. 다만 극단 값(outlier)에 매우 민감하여, 소수의 매우 크거나 작은 값이 전체 평균을 크게 왜곡할 수 있음.
- 중앙값(Median)은 데이터를 크기 순으로 정렬했을 때 정확히 중앙에 위치한 값으로, 데이터의 개수가 짝수일 경우, 중앙에 있는 두 값의 평균을 사용함. 중앙값은 극단 값의 영향을 거의 받지 않아 치우친 분포에서도 안정적인 중심 경향을 나타냄,. 다만, 데이터의 전체적인 분포를 완전히 반영하지는 못한다는 단점이 있음
- 최빈값(Mode)은 데이터에서 가장 빈번하게 나타나는 값으로, 가장 잘 팔리는 상품의 품목이나 색상의 파악과 같은 범주형 데이터나 이산형 데이터를 분석 할 때 유용함. 다만, 데이터의 전반적인 분포나 극단 값에 대한 정보를 제공하지 않는다는 한계가 있음

집중화 경향 지표 계산식

평균 : (μ) = Σx / n

(여기서, Σx는 모든 값의 합, n은 데이터의 개수)

중앙값

- 데이터 개수가 홀수일 때: (n+1)/2 번째 값
- 데이터 개수가 짝수일 때: n/2 번째 값과 (n/2)+1 번째 값의 평균

최빈값

-가장 높은 빈도로 나타나는 값(별도의 수식 없이 빈도 계산)

집중화 경향 지표 관련 사례

평균(Mean)의 오류

- 소득 데이터에서 극소수의 고소득자가 있다면 평균 소득이 실제 대부분의 사람들의 소득보다 훨씬 높게 나타날 수 있음.

중앙값(Median)의 활용 사례

- 소득이나 주택 가격 등과 같이 극단 값이 존재할 수 있는 데이터에서 자주 사용됨.

최빈값(Mode)의 한계점

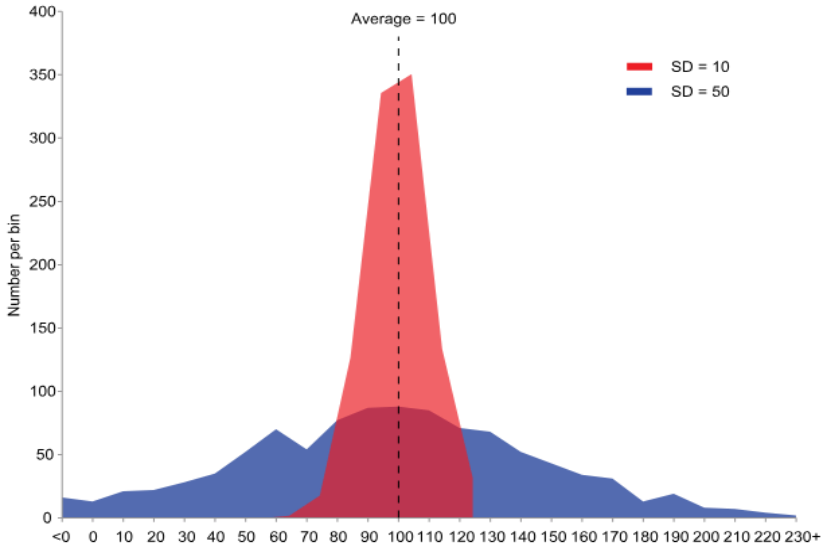
- 주식 종가(Close Price)와 동일한 값이 나오기 어려운 연속형 데이터에서는 사용하기 어려움

3. 분포의 특성

3) 데이터의 분산도

- 분산도는 데이터가 중심에서 얼마나 흩어져 있는지를 나타내는 지표로, 데이터의 변동성 및 안정성을 평가하는데 쓰임. 예를 들어, 어느 회사 고객들의 신용도의 평균 점수가 80점이라고 할 때, 대부분의 고객들이 75-85점 사이에 몰려 있다면 분산도가 낮은 것이고, 30-100점까지 넓게 퍼져 있다면 분산도가 높다고 평가함
- 이러한 분산도를 수치화한 지표로 분산과 표준편차가 제일 많이 쓰이는 데, 분산은 각 점수가 평균에서 얼마나 떨어져 있는지를 제곱하여 평균 낸 값임. 한편 표준편차는 분산의 제곱근으로, 원래 데이터와 같은 단위를 가지며, 변동계수는 표준편차를 평균으로 나눈 값으로, 서로 다른 데이터 세트를 비교할 때 유용함

데이터의 분산도 사례 : 평균은 같지만 분산도가 다른 경우



- 좌측 그림은 평균값이 동일하지만 분산도가 다른 두 집단의 데이터 분포를 보여주고 있음. 두 집단 모두 평균이 100이지만, 붉은색 분포의 표준편차는 10, 파란색 분포의 표준편차는 50임

- 여기서 표준편차는 표준편차가 데이터의 분산 정도를 어떻게 나타내는지 시각적으로 보여주고 있음. 표준편차가 작을수록(SD=10) 데이터가 평균에 가깝게 모여 있기 때문에 평균 값으로 전체 데이터를 예측할 가능성이 높음(변동성이 작음), 반대로 표준편차가 클수록(SD=50) 데이터가 평균에서 멀리 퍼져 있어 예측 가능성이 낮음(변동성이 큼)

3. 분포의 특성

4) 데이터의 분산도 지표 #1

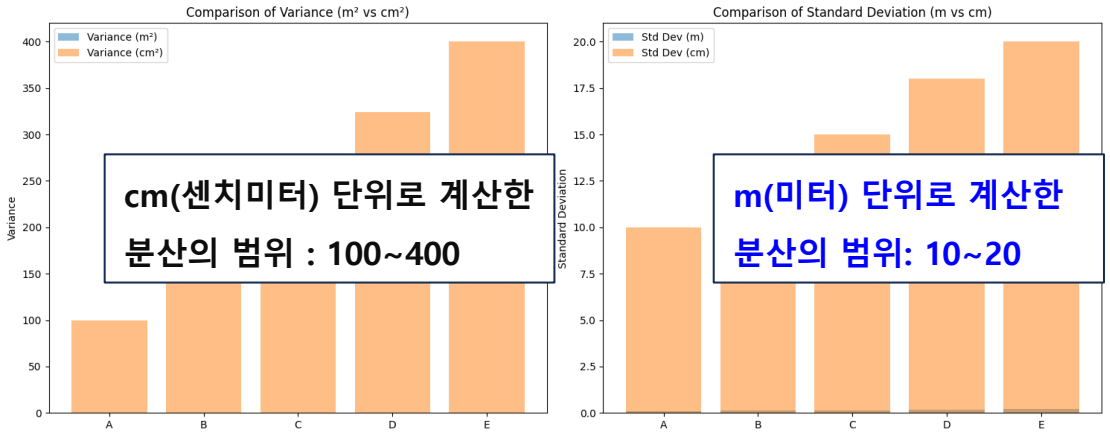
- **분산(Variance)** : 각 데이터 포인트와 평균 간의 차이를 제곱한 값들의 평균으로, 데이터의 퍼짐 정도를 나타냄. 모든 데이터 포인트를 고려하여 변동성을 측정하는 장점이 있으나, 단위가 원래 데이터의 제곱이라 단위가 없으며, 이 때문에 특정 단위의 영향을 크게 받음(예 : cm 단위로 측정한 분산이 m 단위 측정한 분산보다 더 큼)

  (분산 계산식) $\sigma^2 = \Sigma(x - \mu)^2 / N$

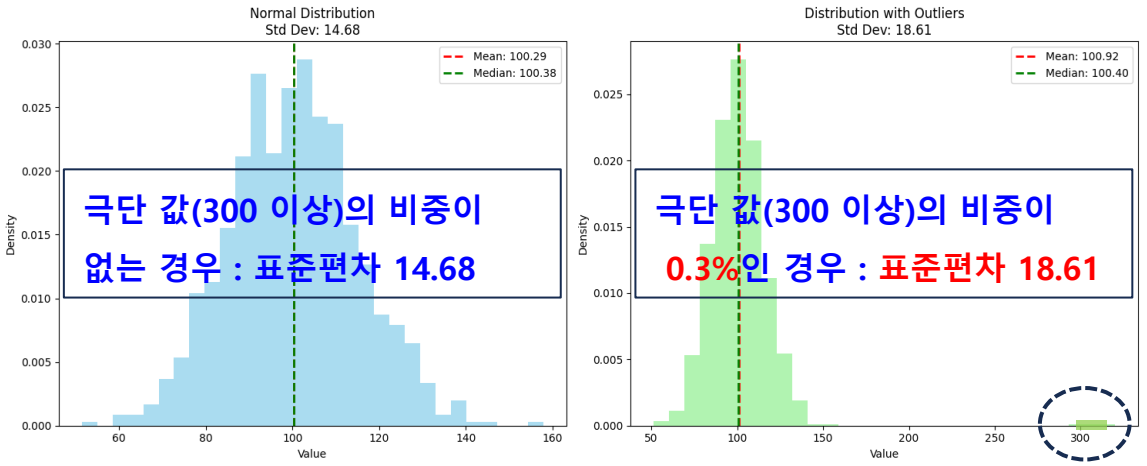
- **표준편차(Standard Deviation)** : 분산의 제곱근으로, 원래 데이터와 같은 단위를 가져 해석이 직관적임. 데이터의 분산 정도를 나타내는 가장 일반적인 지표로 사용되며, 정규분포에서 특히 유용함. 다만, 극단값에 민감하여 왜곡될 수 있음.

  (표준편차 계산식) $\sigma = \sqrt{\Sigma(x - \mu)^2 / N}$

측정단위 영향이 큰 분산의 사례



극단 값에 민감한 표준편차 사례



3. 분포의 특성

4) 데이터의 분산도 지표 #2

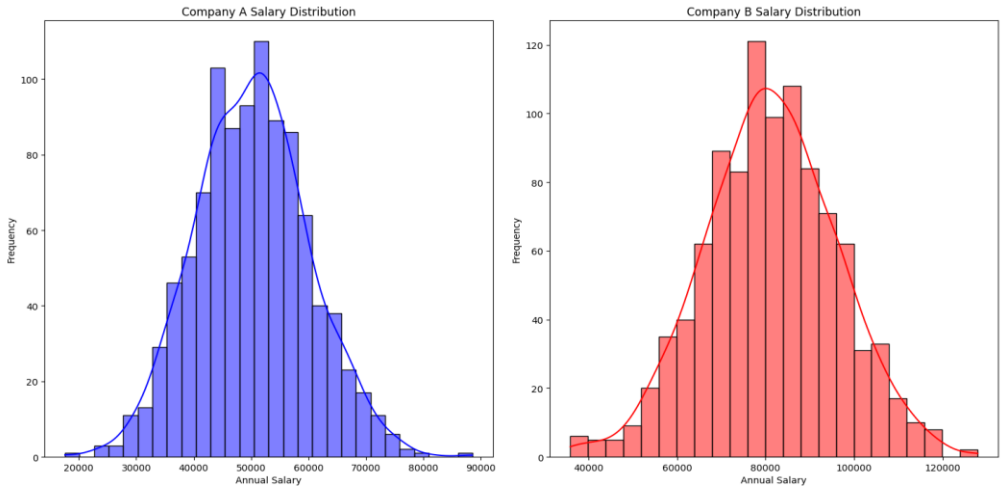
- **변동계수(Coefficient of Variation)** : 표준편차를 평균으로 나눈 값으로, 단위가 다른 데이터 세트 간의 변동성을 비교할 때 유용함. 평균이 0에 가까울 때 사용이 제한적이며, 음수 값을 가진 데이터에는 적용하기 어려움

  (변동계수 계산식) $CV = (\sigma / \mu) * 100\%$
- **최대값(Maximum), 최소값(Minimum)** : 각각 데이터 세트에서 가장 큰 값과 가장 작은 값을 나타냄. 데이터의 전체적인 범위를 빠르게 파악할 수 있게 해주지만, 극단값에 매우 민감하여 전체 분포를 대표하지 못할 수 있음.

  (최대값 계산식) $\max(x_1, x_2, \ldots, x_n)$, (최소값 계산식) : $\min(x_1, x_2, \ldots, x_n)$
- **범위(Range)** : 최대값과 최소값의 차이로, 데이터의 전체적인 퍼짐을 간단히 나타냄. 계산이 쉽고 이해하기 쉬운 장점이 있으나, 극단 값에 매우 민감하고 데이터의 중간 분포 정보를 제공하지 않는 단점이 있음.

  (범위 계산식) $R = \max(x) - \min(x)$

A, B 두 회사의 연봉 비교



- 좌측 그림은 A. B회사의 연봉 분포를 나타낸 것으로, A회사의 평균 연봉은 50,000달러, 표준편차는 10,000달러, B회사는 평균 80,000달러, 표준편차는 15,000달러임
- A와 B 회사의 평균과 표준편차가 서로 다르기 때문에, 단순히 표준편차만으로는 두 회사의 연봉 변동성(연봉의 분포가 평균에서 얼마나 떨어져 있는지)을 직접 비교하기 어려움

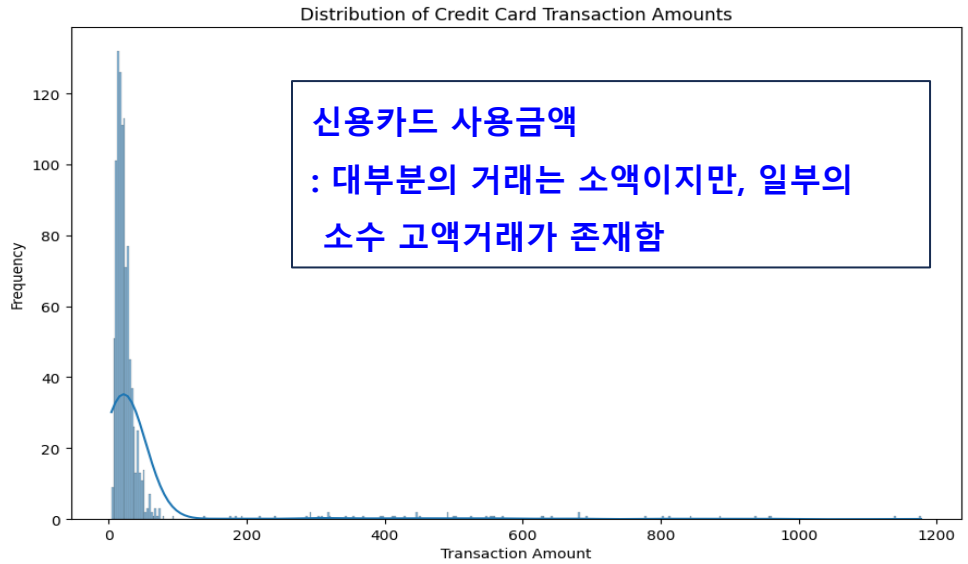
- **변동계수는 표준편차를 평균으로 나눈 값으로, 평균 값을 1로 표준화한 상대적인 표준편차**라고 볼 수 있음. 따라서 변동계수는 평균이 다른 두 집단의 변동성을 비교할 때 유용함
  : A회사의 변동계수는 0.20, B회사는 0.18로 A회사가 변동성이 더 큼

3. 분포의 특성

5) 데이터의 편향성

- 데이터의 *편향성(Skewness)은 데이터 분포의 대칭성(Symmetry) 또는 비대칭성(Asymmetry)을 나타내는데, 데이터의 극단값과 비정상적인 패턴을 식별하는 데 중요한 역할을 함. 편향성을 통해 데이터의 전반적인 형태와 특성을 파악할 수 있으며, 이는 핀테크 데이터 분석 및 모델링에 있어 핵심적인 정보를 제공함
- 데이터의 편향성 관찰 방법
  - 데이터의 **분포 형태**를 통해 **중심을 기준으로 한쪽으로 치우친 정도**를 파악할 수 있음
  - **극단적으로 크거나 작은 값의 존재 여부**를 확인함으로써 편향성을 판단할 수 있음.
  - **평균, 중앙값, 최빈값과 같은 대표값들 간의 관계를 분석**하는 것도 편향성을 확인할 수 있음
  - **일반적인 예상과 다른 특이한 패턴을 식별**함으로써 **데이터의 비정상적인 특성**을 파악가능

데이터 편향성 사례 : 신용카드 카드 금액 분포



- 핀테크 분야에서 데이터의 편향성을 보이는 구체적인 사례로, 신용카드 사용 금액 데이터를 들 수 있음
- 일반적으로 신용카드 사용 금액은 양(+)의 편향(오른쪽 꼬리 분포)을 보이는데, 이는 대부분의 거래는 소액이지만, 소수의 고액 거래가 존재하기 때문임.
- 이러한 편향성을 인식하면 이상 거래 탐지, 고객 세그먼테이션, 신용 한도 설정 등에 있어 더 정확한 모델을 구축할 수 있음
- 고객의 등급 분류를 위해서 평균값 대신 중앙값을 사용하거나, 또는 머신러닝 모델을 사용함에 있어 원(Raw) 데이터가 아닌 로그 변환을 통해 데이터를 편향성을 낮추기 위해 정규화하는 등 방법을 적용할 수 있음

*여기서 편향성(Skewness)이란 분포의 비대칭성을 뜻하는 말로, 통계적 추정의 편향인 Bias와는 다른 개념임

3. 분포의 특성

6) 데이터의 편향성 지표

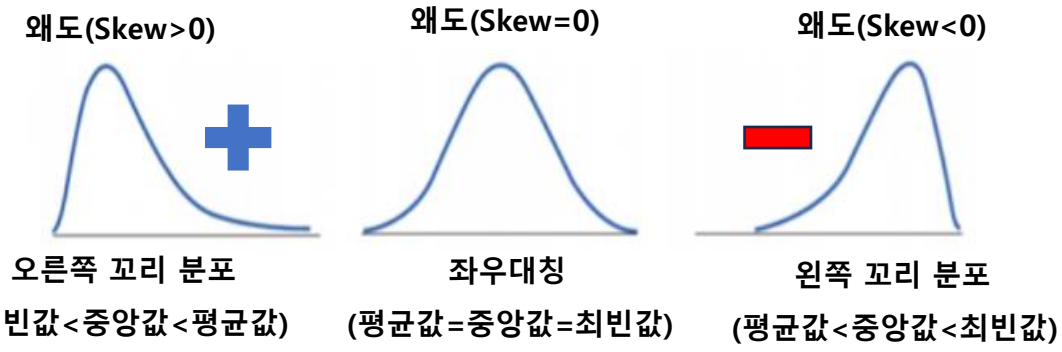
- **왜도(Skewness)** : 분포의 비대칭 정도를 나타내며, 양(+)의 왜도는 오른쪽으로, 음(-)의 왜도는 왼쪽으로 치우친 분포를 의미함. 일반적으로 왜도가 0보다 크면 오른쪽 꼬리분포(하향 평준화 경향), 0보다 작으면 오른쪽 꼬리분포(상향 평준화 경향)을 보인다고 할 수 있음

(왜도 계산식) $\text{Skewness} = \frac{1}{n} \times \frac{\sum[(Xi-\mu)^3]}{\sigma^3}$ (xi는 각 데이터 포인트, μ 는 평균, σ 는 표준편차, n은 데이터 포인트 수)

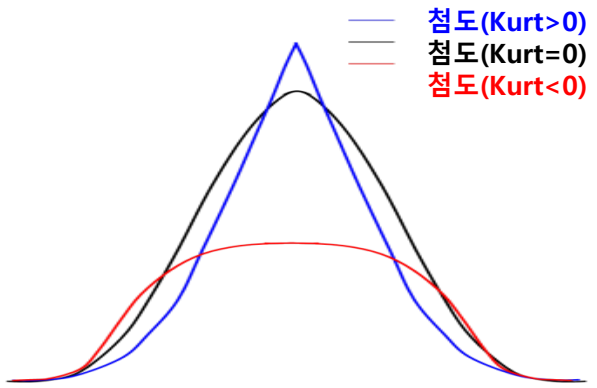
- **첨도(kurtosis)** : 분포의 뾰족한 정도를 나타내며, 정규분포와 비교한 데이터의 집중도를 나타냄. 첨도가 3보다 크면 정규분포보다 뾰족하고, 3보다 작으면 정규분포보다 완만함. 계산의 편의를 위해서 초과첨도(Excess kurtosis)를 사용하는데 "첨도- 3"으로 계산하며, 초과첨도의 절대값이 2 미만이면 정규분포에 가깝고, 2에서 7 사이면 중간 정도의 첨도, 7 이상이면 높은 첨도로 간주함

(첨도 계산식) $\text{Kurtosis} = \frac{1}{n} \times \frac{\sum[(Xi-\mu)^4]}{\sigma^4}$ (xi는 각 데이터 포인트, μ 는 평균, σ 는 표준편차, n은 데이터 포인트 수)

왜도에 따른 분포의 특징



첨도에 따른 분포의 특징



1. 핀테크 데이터의 특성 시각화

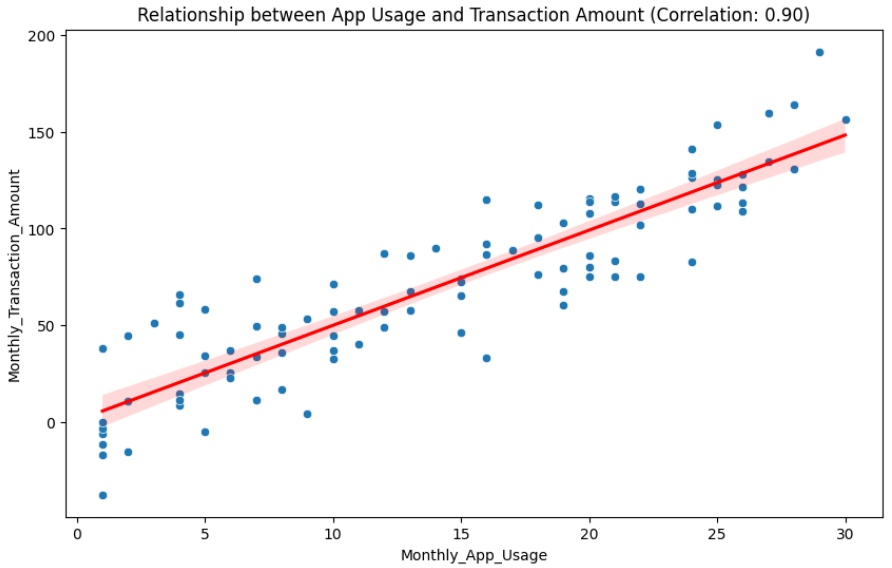
7) 데이터의 상관 관계

- 상관관계(Correlation)는 두 변수 간의 관계를 나타내는 통계적 개념으로, 관계의 존재 여부, 강도, 방향, 형태 그리고 예측 가능성을 포함함. 이러한 상관관계는 변수들 간의 관계를 이해하고, 잠재적인 인과관계를 탐색하며, 예측 모델을 구축하는 데 기초 됨
- 상관관계를 나타내는 주요 통계지표로는 공분산과 피어슨 상관계수가 있으며, 이들은 변수 간 관계의 정도를 수치화함. 이중 공분산은 두 변수의 변동 방향성을, 피어슨 상관계수는 -1에서 1 사이의 값으로 관계의 강도와 방향을 나타냄.
- 아래 [그림]과 같이 핀테크 데이터 분석 시 변수 X(월간 앱 사용빈도)는 다른 변수 Y(월간 거래 금액)과 양(+)관계가 있는 것으로 나타남. 이를 통해 앱 사용을 늘리는 마케팅 캠페인의 예상 결과를 계산할 수 있음

상관 관계의 주요 특징

특성	설명
관계의 존재 여부	• 두 변수 사이에 의미 있는 관계가 있는지를 나타냄 • 관계가 없다면 두 변수는 서로 독립적으로 변화함
관계의 강도	• 두 변수가 얼마나 밀접하게 연관되어 있는지를 나타냄 • 강한 관계는 한 변수의 변화가 다른 변수의 변화와 밀접하게 연관되어 있음을 의미함
관계의 방향	• 두 변수가 같은 방향으로 변화하는지(양의 관계) 또는 반대 방향으로 변화하는지(음의 관계)를 나타냄
관계의 형태	• 대부분의 경우 선형 관계를 가정함 • 비선형 관계도 존재할 수 있음
예측 가능성	• 한 변수의 값을 알면 다른 변수의 값을 어느 정도 예측할 수 있는지를 나타냄

앱 사용 빈도와 거래금액의 관계

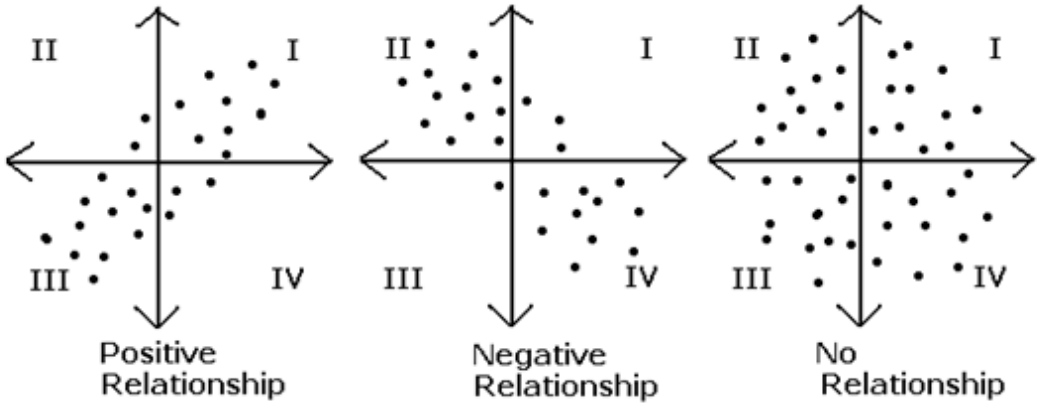


3. 분포의 특성

8) 데이터의 상관관계 지표 #1

- **공분산(Covariance)** : 공분산은 두 변수가 어떻게 함께 변화하는지를 나타내는 지표로, 두 변수의 편차 곱의 평균으로 계산됨. 계산 결과 양(+)의 값이 나오면 한 변수가 증가할 때 다른 변수도 증가하는 경향, 반대로 음(-)의 공분산이 나오면 한 변수가 증가할 때 다른 변수가 감소하는 경향을 나타냄. 하지만 공분산은 관계의 방향성 외에 방향성의 강도를 잘 보여주지 못하며, 동일한 값들이라도 측정 단위에 따라 공분산 크기가 달라짐
- **(공분산 계산식) Cov(x, y)=** $\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{n}$ (xi, yi는 각 데이터 포인트 μ_x, μ_y 는 각각 x, y의 평균, n은 데이터 포인트 수)

공분산 값에 따른 두 변수의 관계



- 좌측 그래프는 두 변수 간의 관계를 나타내는 산점도(scatter plot)와 각 관계의 공분산으로 나타낸 것임
- Positive Relationship (양의 관계) : 한 변수가 증가할 때 다른 변수도 증가하는 경향을 보임. 점들이 왼쪽 아래에서 오른쪽 위로 올라가는 패턴을 형성하며, 공분산은 0보다 큼. 관계가 클 수록 값이 증가
- Negative Relationship (음의 관계) : 한 변수가 증가할 때 다른 변수는 감소하는 경향을 보임. 점들이 왼쪽 위에서 오른쪽 아래로 내려가는 패턴을 형성하며, 공분산은 0보다 작음. 관계가 클 수록 값이 감소.
- No Relationship (무관계) : 두 변수 간에 뚜렷한 관계가 없으며, 점들이 특정 패턴 없이 고르게 분포되어 있음. 공분산이 0에 가까우며, 만약 0이면 전혀 관련성이 없다고 해석함

3. 분포의 특성

9) 데이터의 상관관계 지표 #2

- **상관계수(Correaltion)** : 공분산의 한계를 보완하여, 두 변수 간의 선형 관계의 강도와 방향을 -1에서 1 사이의 값으로 표준화하여 나타낸 지표임. 가장 널리 사용되는 피어슨(Pearson) 상관계수는 공분산을 각 변수의 표준편차의 곱으로 나누어 계산함. 상관계수가 1에 가까울수록 강한 양의 선형 관계를, -1에 가까울수록 강한 음의 선형 관계를, 0에 가까울수록 선형 관계가 약함을 의미함.
- **(상관계수 계산식)** $r(x, y) = \frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum(x_i-\mu_y)^2}\sqrt{\sum(y_i-\mu_y)^2}}$ (xi, yi는 각 데이터 포인트 μ_x, μ_y는 각각 x, y의 평균, n은 데이터 포인트 수)

상관계수 값에 따른 두 변수의 관계

공분산과 상관계수의 관계 비교

공분산(Cov) : 값이 클 수록 음(-) 또는 양(+)의 상관관계가 커짐

상관계수(Corr) : 값이 클 수록 음(-) 또는 양(+)의 비례 관계가 명확해짐

- 좌측 그래프는 두 변수 간의 관계를 나타내는 산점도(scatter plot)와 각 관계의 상관계수로 나타낸 것임
- r = -1: 완벽한 음의 선형 관계. 한 변수가 증가하면 다른 변수는 정확히 반비례하면서 감소
- -1 < r < 0: 약한 음의 선형 관계. 한 변수가 증가할 때 다른 변수는 대체로 감소하지만, 정확한 비례 관계는 나타나지 않음
- r = 0: 선형 관계 없음. 두 변수 사이에 선형적인 관계가 없음
- 0 < r < 1: 약한 양의 선형 관계. 한 변수가 증가할 때 다른 변수도 대체로 증가하지만, 정확한 비례 관계는 나타나지 않음
- r = +1: 완벽한 양의 선형 관계. 한 변수가 증가하면 다른 변수도 정확히 비례하면서 증가

3. 분포의 특성

10) 공분산과 상관계수의 주요 특징 비교 #1

주요 특징	공분산	상관계수
척도(측정단위) 의존성	• 측정 단위에 따라 값이 달라짐(같은 값이라도cm, m 단위에 따라 값이 크게 달라짐) • 서로 다른 단위를 가진 변수들 간에 관계를 비교하기는 어려움 (예 : 흡연량과 건강보험료, BMI와 건강보험료 간에 관계 크기 비교)	• 척도 불변성으로 변수의 측정 단위에는 영향을 받지 않음(공분산과 달리 측정 단위에 영향을 받지 않음)- • 변수를 선형 변환(x2=ax1+b형태로 변환)하는 것에는 영향을 받지 않음 • 변수를 비선형 변환(x2=ax1*x1+b형태로 변환)할 경우에는 상관계수 값이 달라짐
상관관계에 대한 해석	• 공분산 값 자체로 상관 관계의 강도를 직관적으로 이해하기 어려움 (예 : 두 변수와의 관계에서 공분산 200은 어떤 의미인가?)	• 선형 관계로 측정하기 때문에 직관적으로 이해가 가능함 (예 : 두 변수의 양(+)의 선형 관계가 강할 수록 상관계수가 1에 가까워짐)
이상치 및 극단값에 의한 영향 여부	• 평균의존성으로 인해 이상치에 민감함. 공분산은 각 변수의 평균으로 부터 편차를 사용해서 계산 : 계산식에서 (X-Xm), (Y-Ym)과 같이 두 변수의 평균을 이용하기 때문에 극단 값으로 평균에 영향을 받을 경우 공분산도 영향 받음	• 평균의존성으로 인해 이상치에 민감함. 상관계수는 각 변수의 평균으로 부터 편차를 사용해서 계산 : 계산식에서 (X-Xm), (Y-Ym)과 같이 두 변수의 평균을 이용하기 때문에 극단 값으로 평균에 영향을 받을 경우 상관계수도 영향 받음 • 대안으로 평균 값을 이용하는 피어슨 상관 계수 외에 중 간값을 기반으로 한 순위 상관계수(스피어만 상관계수)를 사용할 수 있음

3. 분포의 특성

10) 공분산과 상관계수의 주요 특징 비교 #2

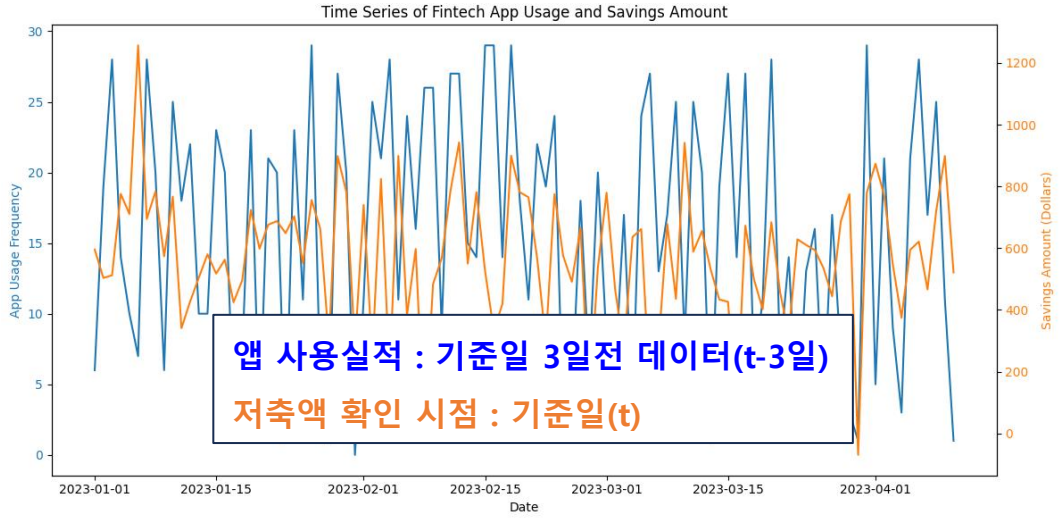
주요 특징	공분산	상관계수
인과관계	• 두 변수 간의 관계의 강도나 방향을 나타내지만 인과 관계를 의미하지 않음 • 높은 상관관계가 있다고 반드시 한 변수가 다른 변수의 원인이 되는 것은 아님	• 공분산과 동일함
변수 간 관계의 복잡성	• 두 변수의 다양한 관계 중 선형관계만을 측정하며, 비선형 관계나 복잡한 상호작용은 포착이 어려움 • 다변량 관계를 두 변수 간의 관계로 단순화 함 : (예) 기업의 주가는 기업의 실적, 투자자의 기대, 경쟁 기업의 실적 등 다양한 요인에 의해서 영향 받지만 공분산은 1:1로 한개의 요인만을 선정함 • 다중회귀분석, 요인분석, 구조방정식 모델 등을 이용한 다변량 분석 방법 등을 통해 한계 극복	• 공분산과 동일함
데이터 포인트의 수(표본의 크기) 영향 여부	• 데이터 포인트의 수(표본의 크기)에 따라 공분산의 값이 크게 달라질 수 있음 ✓ 데이터 포인트 수가 적을 경우에는 극단치 및 이상치에 의해서 공분산 값이 높게 나올 수 있음 ✓ 데이터 포인트 수가 적을 때에는 일반적인 케이스가 다른 상관관계가 나올 가능성이 있음 ✓ 따라서 데이터 포인트 수가 많은(대규모 표본)으로 분석하는 것으로 한계를 보완	• 공분산과 동일함

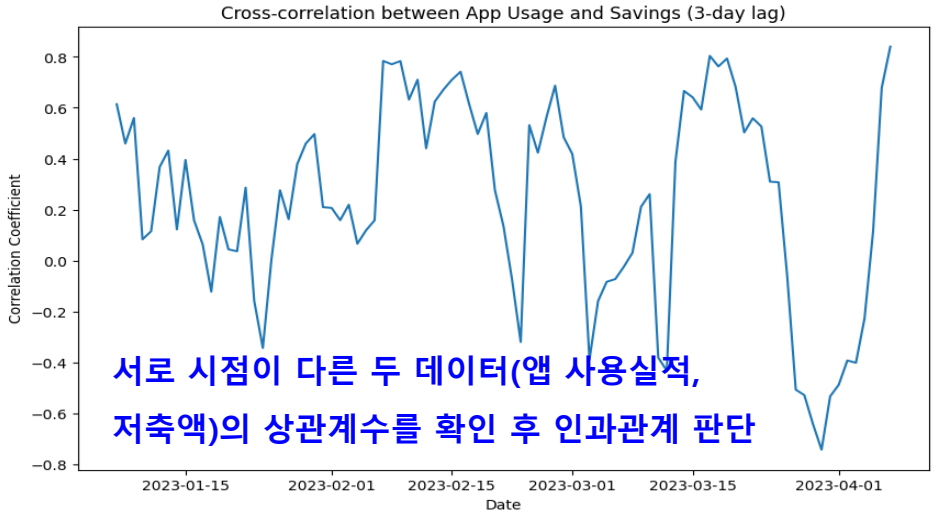
3. 분포의 특성

11) 인과관계에 대한 분석 방법 #1

- 공분산과 상관계수 등의 상관관계 지표들은 두 변수 간의 관계를 수치화해 보여주지만, 인과관계를 정확히 파악하기 어려움. 예를 들어, 핀테크 앱 사용 빈도와 저축액 간의 높은 상관관계가 있다고 해서, 앱 사용이 저축 증가의 직접적인 원인이라고 단정 짓기 어려움
- 인과관계를 파악하기 위해서는 정확한 실험설계가 필요함. 시계열 데이터 분석 방법 중에 하나인 그레인져-인과 관계검증을 하면 다음과 같이 진행할 수 있음
  - ✓ 핀테크앱 사용 시점이 저축액 확인 시점보다 3일 정도 빠르다고 보고, 핀테크앱의 3일 전 사용실적과 저축액의 데이터 사이에 유의미한 상관관계가 확인됐을 경우에는 다른 조건(교육수준, 개인사정 등)을 배제한다는 가정하에 인과 관계 고려 가능

핀테크앱 사용 빈도와 저축액 간의 인과관계 분석





3. 분포의 특성

12) 인과관계에 대한 분석 방법 #2

분석 방법	주요 특징	한계점
무작위 통제 실험(RCT)	• 참가자를 무작위로 실험군과 대조군으로 나누어 개입 효과를 직접 측정함으로써, 다른 요인들의 영향을 최소화하고 순수한 원인과 결과 관계를 파악하는 실험적 방법	• 윤리적 문제로 인해 수행이 불가능한 경우가 많고, 비용과 시간이 많이 들어 실행하기 어려움. • 실험 환경이 실제 상황과 달라 결과를 일반화하기 어려울 수 있으며, 오랜 기간에 걸친 효과를 측정하기 힘듦.
회귀분석	• 독립변수와 종속변수 간의 관계를 수학적 모델로 표현하여, 변수 간 관계의 강도와 방향을 수치화하고 예측에 활용하는 광범위하게 사용되는 통계적 방법	• 변수들 간의 관계를 보여주지만, 실제로 무엇이 원인이고 결과인지 명확히 알기 어려움. • 중요한 변수를 누락할 수 있고, 원인과 결과가 서로 영향을 주거나 동시에 일어나는 경우 분석이 복잡해짐. 또한, 곡선 형태의 관계를 정확히 파악하기 어려움.
도구변수법	• 관심 변수와는 관련되지만 결과와는 직접 관련이 없는 중간 변수를 활용하여, 변수 간의 복잡한 관계로 인한 문제를 해결하고 원인과 결과 관계를 추정하는 경제학적 방법	• 분석에 필요한 적절한 중간 변수를 찾기가 매우 어려움. 찾은 변수가 약하면 결과가 부정확할 수 있고, 이 방법의 가정들이 맞는지 확인하기 어려움. • 특정 조건에서만 효과를 추정할 수 있어 전체적인 효과를 알기 어려움.
차분법	• 개입 전후의 변화를 영향을 받은 그룹과 받지 않은 그룹 간에 비교함으로써, 시간에 따른 변화를 고려하여 정책 효과 등을 평가하는 유사 실험 방법	• 비교 그룹들이 시간에 따라 비슷한 추세를 보인다는 가정이 필요하지만, 이를 확인하기 어려움. 시간이 지나면서 생기는 다른 영향들을 구분하기 힘듦 • 대부분 단기적인 효과만 볼 수 있으며, 그룹들 간의 차이를 충분히 고려하지 못할 수 있음.
성향점수 매칭	• 개입을 받은 그룹과 유사한 특성을 가진 비교 그룹을 통계적으로 짝지어, 선택 편향을 줄이고 관찰 연구에서 실험적 설계를 모방함으로써 원인과 결과 관계를 추정하는 방법입니다.	• 관찰할 수 없는 요인들은 통제하지 못하고, 비교 그룹을 정확히 짝짓기가 어려울 수 있음. • 이 과정에서 데이터의 양이 줄어들 수 있고, 특정 조건에서만 결과를 해석할 수 있어 전체적인 상황에 적용하기 어려울 수 있음.
구조방정식 모델링(SEM)	• 여러 변수 간의 복잡한 인과관계를 동시에 분석하고 잠재변수를 포함할 수 있어, 직접 및 간접 효과를 종합적으로 평가하는 고급 통계적 기법	• 복잡한 관계를 모델링할 때 오류가 생길 가능성이 높고, 정확한 분석을 위해 많은 양의 데이터가 필요함. • 결과를 해석하기가 복잡하고, 변수들 간의 곡선 형태의 관계를 분석하기 어려움.
그레인저 인과성 검정	• **시계열 데이터에서 한 변수의 과거값이 다른 변수의 현재값 예측에 도움이 되는지를 통계적으로 검증함으로써, 변수 간의 시간적 선행성과 예측력을 바탕으로 인과관계를 추론하는 방법**	• **통계적으로 관계가 있다고 나와도 실제 원인과 결과 관계가 아닐 수 있음. 분석에 포함되지 않은 다른 중요한 요인들의 영향을 고려하기 어렵고, 곡선 형태의 관계를 정확히 파악하기 힘듦.** • **분석에 사용할 시간 간격을 선택할 때 연구자의 주관이 개입될 수 있음.**

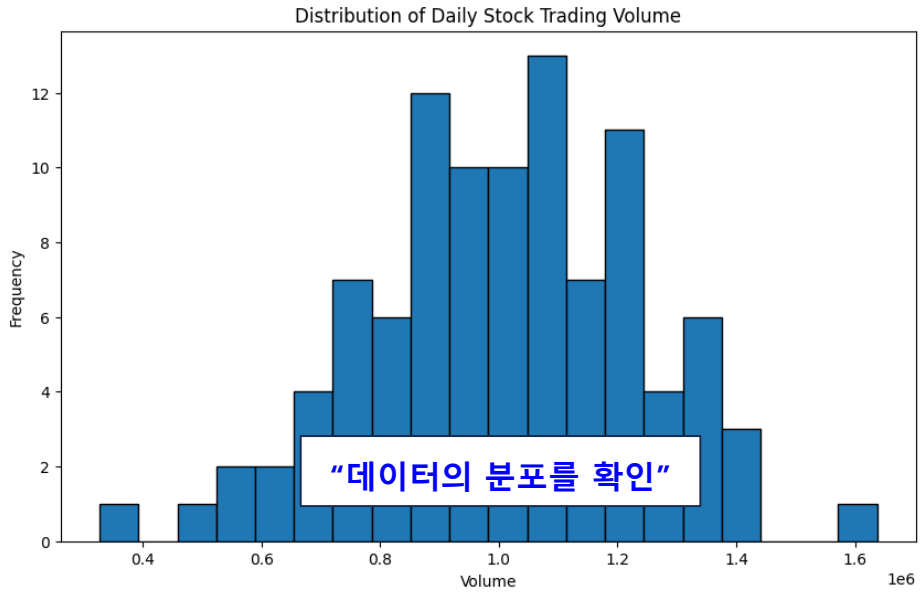
4. 데이터 시각화

4. 데이터 시각화

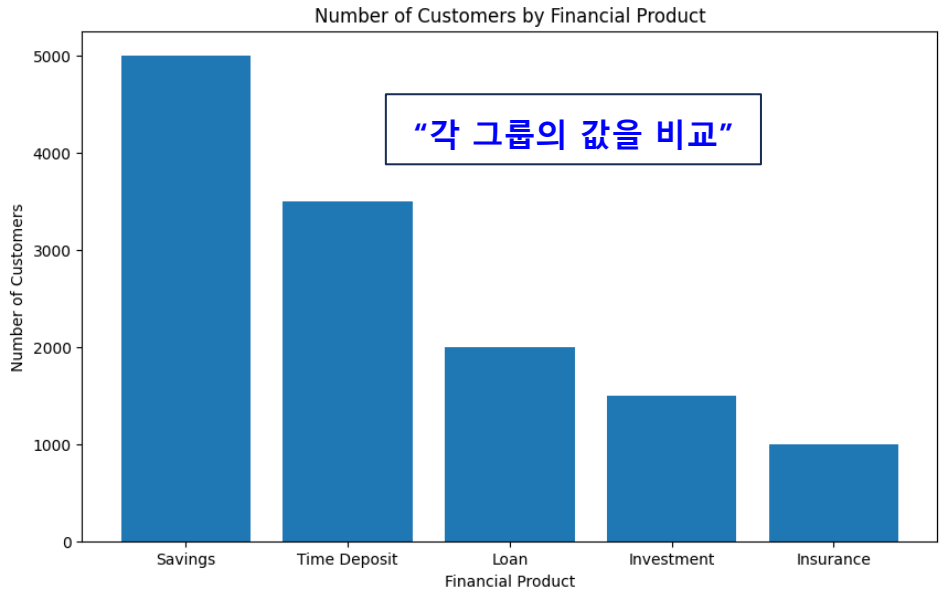
1) 데이터 형태에 따른 차트 #1

- 데이터 유형은 크게 수치형(정량)과 범주형(정성)으로 나뉘며, 수치형 데이터는 연속적인 값을 가지며, 범주형 데이터는 정확히 구분되는 그룹이나 분류를 나타냄. 각 데이터 유형에 따라 자주 쓰이는 차트로는 수치형 데이터의 경우에는 히스토그램이나 산점도, 범주형 데이터의 경우에는 막대 그래프나 파이 차트 등이 있음
- 시계열 데이터를 위해서는 선 그래프나 영역 차트 등 특수한 차트 유형이 사용되는데, 이러한 차트들은 시간에 따른 데이터의 변화를 효과적으로 표현함. 핀테크 데이터를 예로 들면, 주식 가격의 변동은 선 그래프로, 월별 거래량은 막대 그래프로 나타내는 경우가 많음 대전광역시 서구, 문예로16 한가람아파트 2동 1405로

일일 주식거래량 (수치형 자료 : 히스토그램)



금융상품 별 고객 수 (범주형 자료 : 막대그래프)



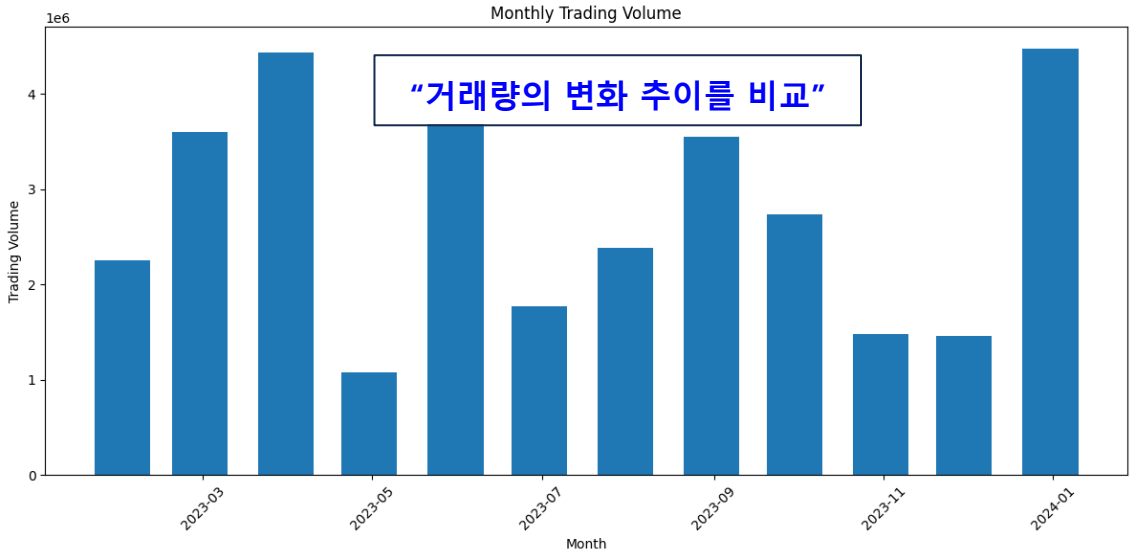
4. 데이터 시각화

1) 데이터 형태에 따른 차트 #2

- 데이터 유형은 크게 수치형(정량)과 범주형(정성)으로 나뉘며, 수치형 데이터는 연속적인 값을 가지며, 범주형 데이터는 정확히 구분되는 그룹이나 분류를 나타냄. 각 데이터 유형에 따라 자주 쓰이는 차트로는 수치형 데이터의 경우에는 히스토그램이나 산점도, 범주형 데이터의 경우에는 막대 그래프나 파이 차트 등이 있음
- 시계열 데이터를 위해서는 선 그래프나 영역 차트 등 특수한 차트 유형이 사용되는데, 이러한 차트들은 시간에 따른 데이터의 변화를 효과적으로 표현함. 핀테크 데이터를 예로 들면, 주식 가격의 변동은 선 그래프로, 월별 거래량은 막대 그래프로 나타내는 경우가 많음.

일일 주식 가격 변동(시계열 데이터 : 선 그래프)

월별 주식 거래량 (시계열 데이터 : 막대 그래프)

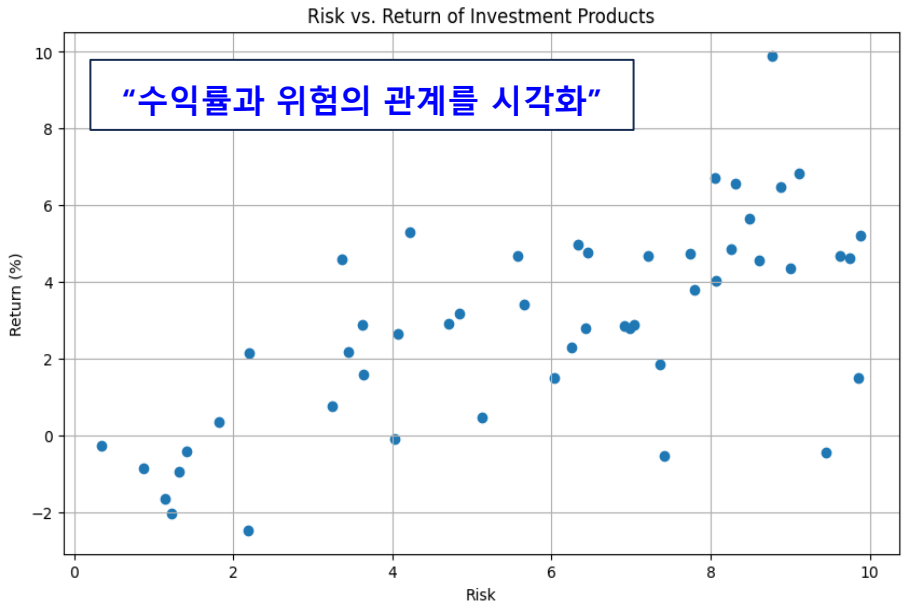


4. 데이터 시각화

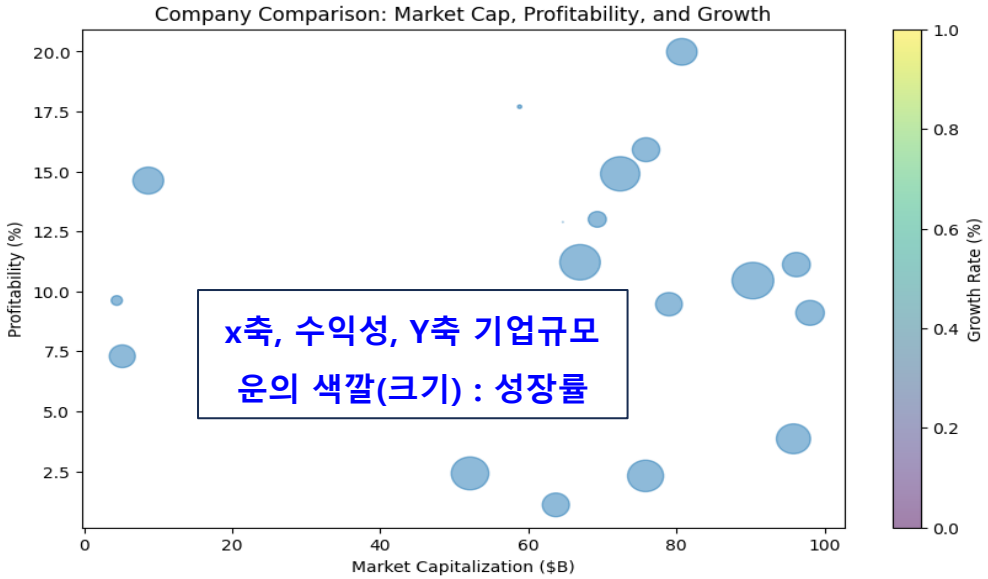
2) 두 변수의 관계 시각화 #1

- 데이터의 유형 뿐 만 아니라, 각 데이터(변수) 간의 관계를 볼 때도 다양한 차트가 활용됨. 산점도(Scatter)는 위험과 수익률 같은 두 변수의 관계를 시각화하는 데 유용하며, 버블 차트는 기업 규모, 수익성, 성장률과 같은 세 변수의 관계를 효과적으로 표현함. 또한 파이 차트는 포트폴리오 자산 구성과 같은 전체에 대한 부분의 비율을 보여주는 데 사용되지만, 비교가 어려운 한계가 있어 트리맵 등의 대안 차트가 활용됨
- 이러한 차트들은 금융 데이터 분석에서 중요한 인사이를 제공함. 예를 들면 산점도를 통해 투자 상품의 위험-수익 프로필을 파악할 수 있고, 버블 차트로 기업의 종합적인 성과를 비교할 수 있음, 또한 파이 차트나 트리맵은 자산 배분 현황을 한눈에 볼 수 있게 해주어 투자 결정에 도움을 줌

산점도(투자자산의 위험과 수익률의 관계)



버블차트 (기업규모, 수익성, 성장률)

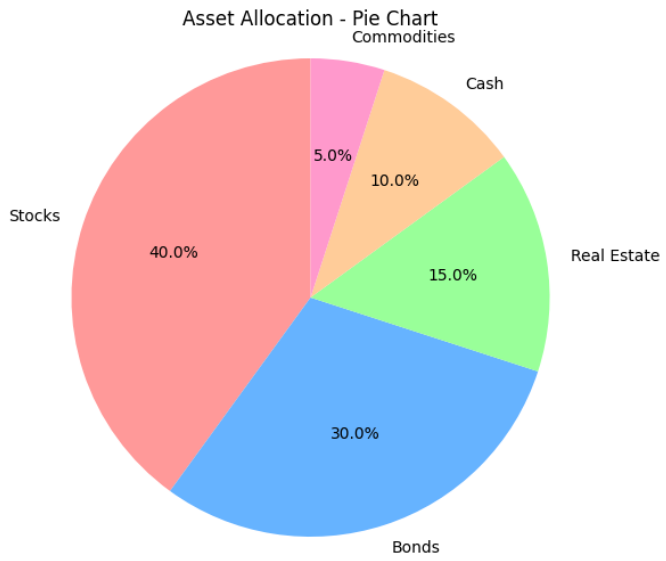


4. 데이터 시각화

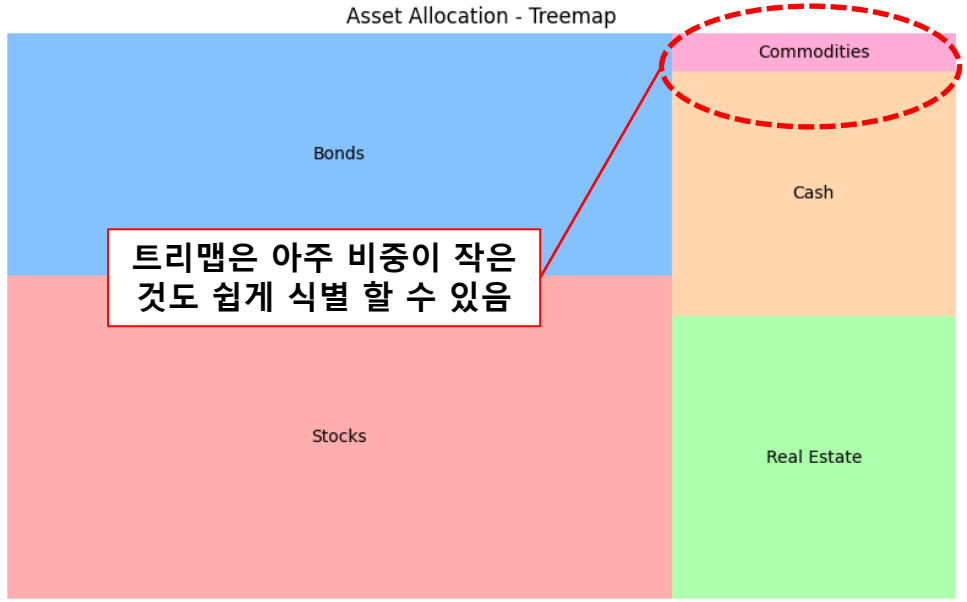
2) 두 변수의 관계 시각화 #2

- 데이터의 유형 뿐 만 아니라, 각 데이터(변수) 간의 관계를 볼 때도 다양한 차트가 활용됨. 산점도(Scatter)는 위험과 수익률 같은 두 변수의 관계를 시각화하는 데 유용하며, 버블 차트는 기업 규모, 수익성, 성장률과 같은 세 변수의 관계를 효과적으로 표현함. 또한 파이 차트는 포트폴리오 자산 구성과 같은 전체에 대한 부분의 비율을 보여주는 데 사용되지만, 비교가 어려운 한계가 있어 트리맵 등의 대안 차트가 활용됨
- 이러한 차트들은 금융 데이터 분석에서 중요한 인사이를 제공함. 예를 들면 산점도를 통해 투자 상품의 위험-수익 프로필을 파악할 수 있고, 버블 차트로 기업의 종합적인 성과를 비교할 수 있음, 또한 파이 차트나 트리맵은 자산 배분 현황을 한눈에 볼 수 있게 해주어 투자 결정에 도움을 줌

파이 차트 (자산의 배분 현황 : 비중을 보여줌)



트리 맵 차트 (자산의 배분 현황 : 비중을 보여줌)

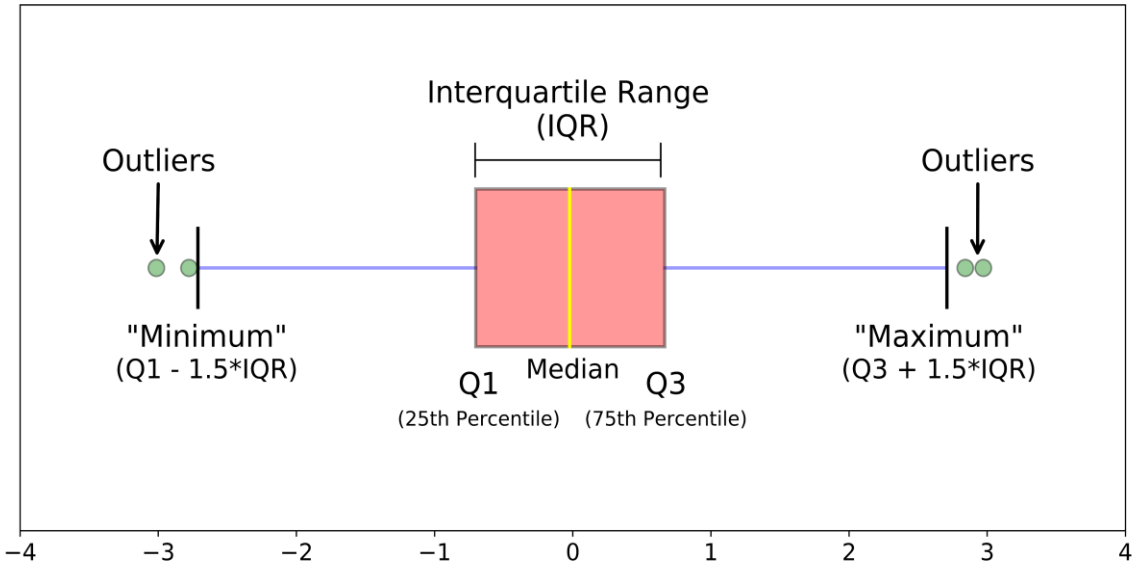


4. 데이터 시각화

3) 박스 플롯(Box-Plot)

- 박스 플롯(Bot-Plot)은 데이터의 분포, 중앙값, 이상치를 효과적으로 시각화하는 도구입니다.최소값, 1사분위수, 중앙값, 3사분위수, 최대값을 한 눈에 보여주며, 이상치도 함께 표시함. 금융 데이터 분석에서는 주가 변동성 분석이나 펀드 성과 비교 등에 활용되어 중요한 인사이트를 발굴할 때 활용되는 차트임
  - ✓대규모 거래 데이터셋에서 박스 플롯을 통해 거래 금액이나 빈도의 일반적인 패턴을 시각화하고, 이상치로 표시되는 데이터 포인트를 집중 조사함 : 이상거래, 자금세탁 등의 활동 식별 탐지에 활용
  - ✓신용 평가 모델에서 고객의 재무 지표를 분석할 때도 박스 플롯을 통해 소득, 지출, 부채 비율 등의 분포를 시각화해 비정상적인 패턴을 보이는 고객을 식별하고, 리스크 관리 전략을 수립

Box Plot에서 이상치 확인 방법



- 좌측의 Box-plot은 데이터의 분포와 이상치(outlier)를 시각적으로 표현하는 데 사용되며. 주요 요소로는 IQR, Q1, Q3 등이 있음
- **사분위수 범위(IQR) : 박스로 표시되며, 데이터의 중간 50%를 나타냄**
- 중앙값(Median) : 박스 안의 노란 선으로, 데이터의 중앙값을 나타냄
- Q1(25th Percentile)과 Q3(75th Percentile) : 박스의 왼쪽과 오른쪽 경계로, 각각 데이터의 25%와 75% 지점을 나타냄

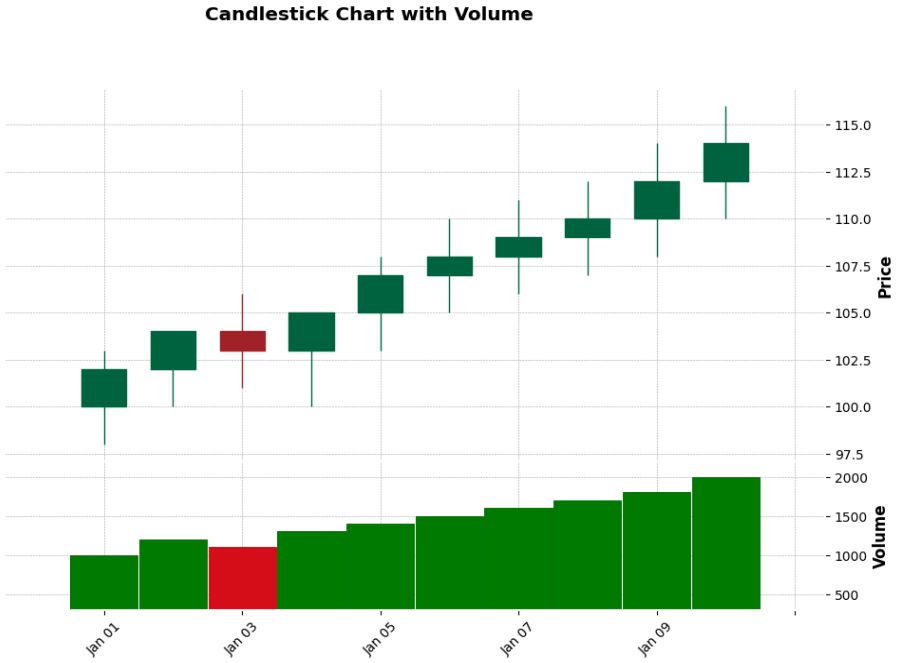
- **이상치 판정 기준 : 하한값(Q1-1.5*IQR), 상한값(Q3+1.5*IQR)**
  - ➢ 최대값, 최소값을 비롯해 위 기준에 보다 작거나 큰 값들 들임 전체적으로 그래프는 -4에서 4까지의 범위를 보여주며, 데이터가 대체로 -1에서 1 사이에 집중되어 있음. 이상치는 약 -3과 3 근처에 위치해 있음

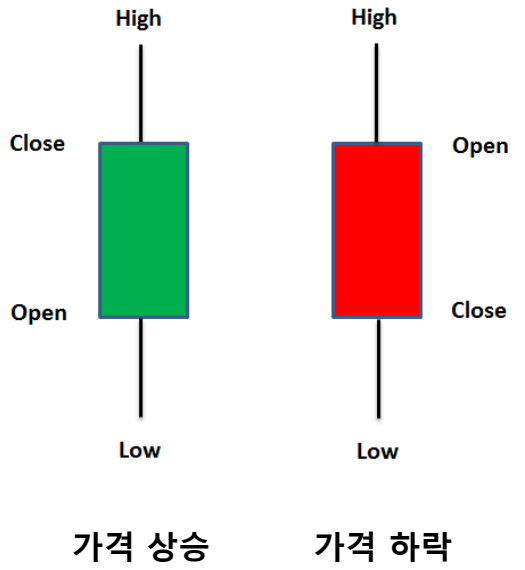
4. 데이터 시각화

4) 캔들스틱, OHLC 차트 #1

- 캔들스틱과 OHLC 차트는 금융시장에서 가격 움직임을 시각화하는 대표적인 차트로, 캔들스틱(Candlestick) 차트는 시가, 고가, 저가, 종가를 하나의 '양초(Candle) ' 로 시각적으로 표현한 것임.
- OHLC는 Open(시가), High(고가), Low(저가), Close(종가)의 앞 글자를 딴 차트로, 금융시장에서 가격의 움직임을 이 4가지 데이터 포인트를 막대로 표현해 가격변동과 추세를 쉽게 파악할 수 있게 함
- 캔들스틱 및 OHLC 차트는 거래량 정보와 함께 표시하는 경우가 많으며, 이를 통해 투자자는 매수/매도 시점 결정, 추세 파악, 지지/저항 수준 식별 등 중요한 투자 의사 결정을 함

거래량과 함께 표시한 캔들스틱 차트



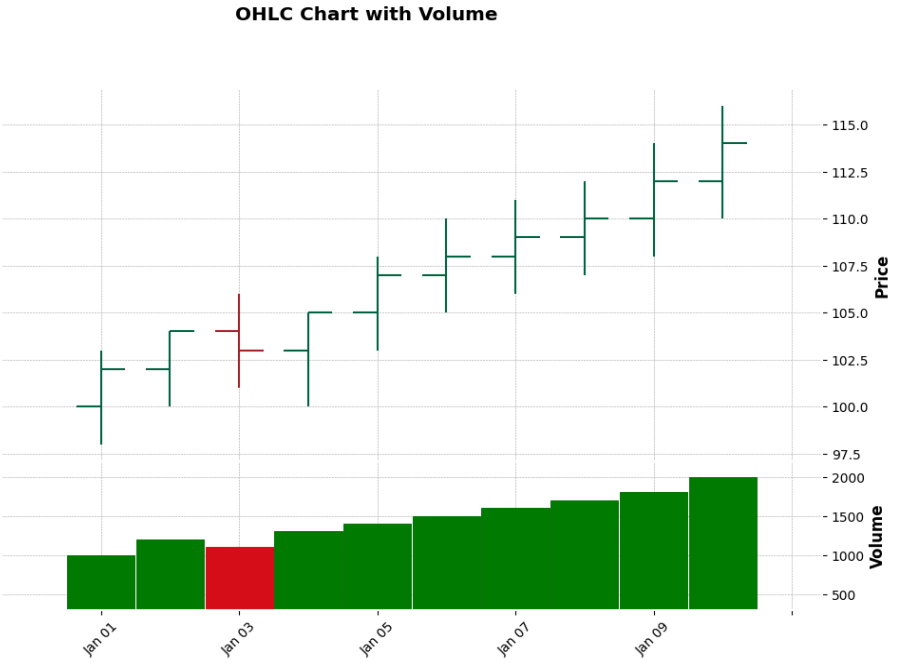


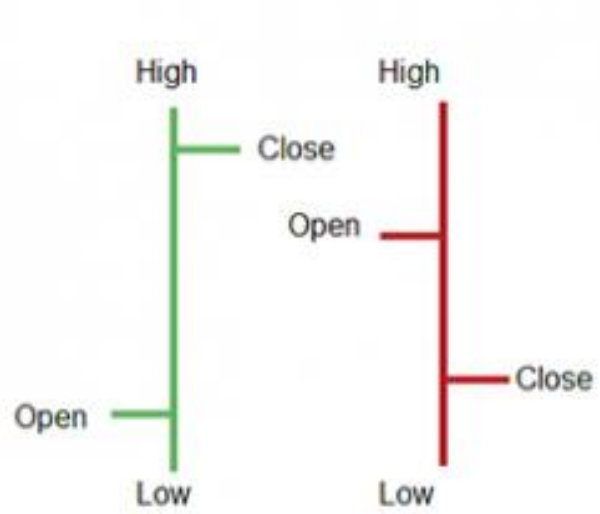
2. 핀테크 데이터 시각화를 위한 차트

4) 캔들스틱, OHLC 차트 #2

- 캔들스틱과 OHLC 차트는 금융시장에서 가격 움직임을 시각화하는 대표적인 차트로, 캔들스틱(Candlestick) 차트는 시가, 고가, 저가, 종가를 하나의 '양초(Candle) ' 로 시각적으로 표현한 것임.
- OHLC는 Open(시가), High(고가), Low(저가), Close(종가)의 앞 글자를 딴 차트로, 금융시장에서 가격의 움직임을 이 4가지 데이터 포인트를 막대로 표현해 가격변동과 추세를 쉽게 파악할 수 있게 함
- 캔들스틱 및 OHLC 차트는 거래량 정보와 함께 표시하는 경우가 많으며, 이를 통해 투자자는 매수/매도 시점 결정, 추세 파악, 지지/저항 수준 식별 등 중요한 투자 의사 결정을 함

거래량과 함께 표시한 OHLS차트





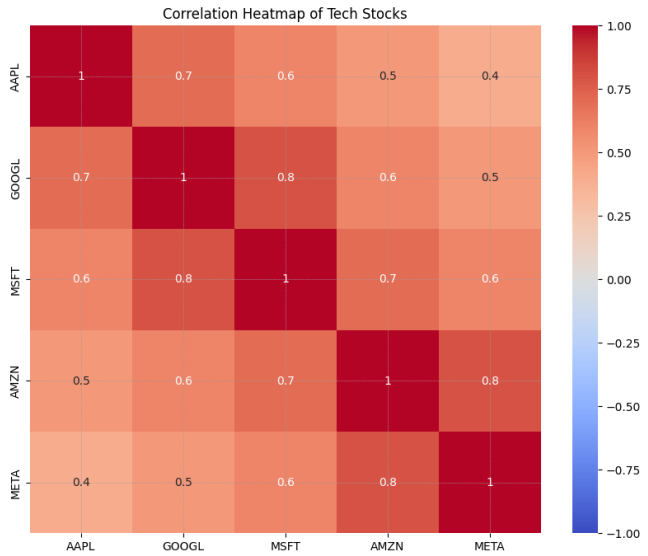
가격 상승　　가격 하락

4. 데이터 시각화

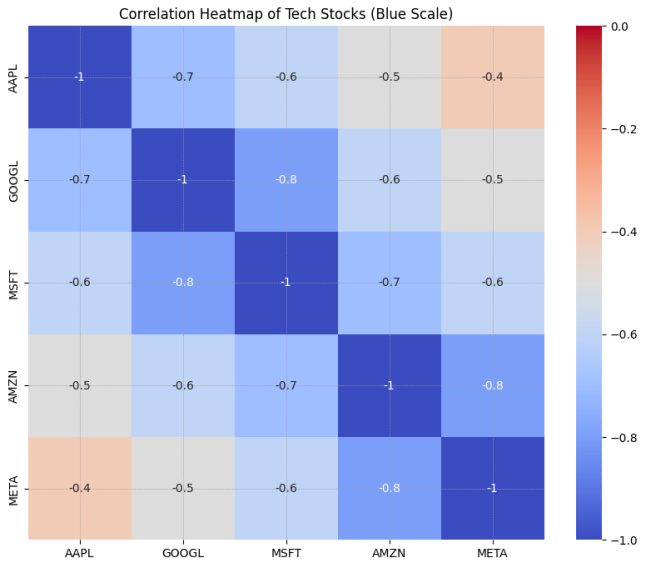
5) 6) 고급 분석 차트 : 히트맵, 트리맵 차트 및 네크워크 그래프 #1

- 히트맵, 트리맵, 네트워크 그래프는 일반적으로 다차원적이고 복잡한 데이터를 시각화하는 차트임.
  - ✓히트맵(Heatmap)은 여러 변수 간의 관계를 시각화한 차트로 데이터 값을 색상의 강도로 표현하는 2차원 그래프임. 행렬 형태의 데이터를 시각화하는 데 주로 사용되며, 각 셀의 색상은 해당 데이터 포인트의 값임
  - ✓트리맵(Treemap)은 여러 변수의 계층적 구조를 시각화한 차트로 계층적 데이터를 중첩된 직사각형들로 표현하는 차트임 각 직사각형의 크기는 해당 데이터 항목의 수량적 가치를 나타내며, 색상은 다른 범주나 그룹을 구분하는 데 사용될 수 있음
  - ✓네트워크 그래프(Network Graph)는 여러 변수의 연결 관계를 시각화한 차트로 노드(개체)와 엣지(관계)로 구성된 데이터 구조를 표현함. 노드는 점으로, 엣지는 선으로 표현되며, 복잡한 관계나 연결 구조를 보여줌

히트맵 : 상관계수 값이 클 수록 적색



히트맵 : 상관계수 값이 클 수록 청색

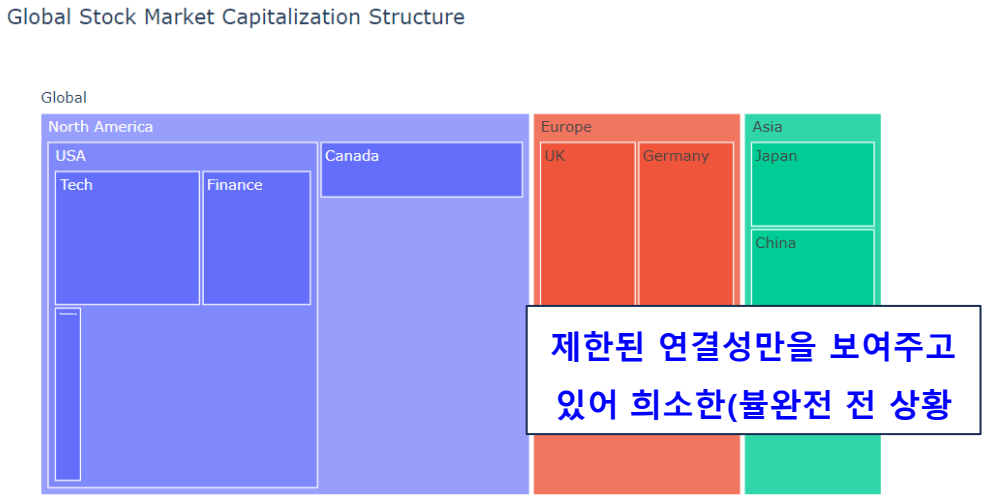


4. 데이터 시각화

6) 고급 분석 차트 : 히트맵, 트리맵 차트 및 네크워크 그래프 #2

- 히트맵, 트리맵, 네트워크 그래프는 일반적으로 다차원적이고 복잡한 데이터를 시각화하는 차트임.
  - ✓히트맵(Heatmap)은 여러 변수 간의 관계를 시각화한 차트로 데이터 값을 색상의 강도로 표현하는 2차원 그래프임. 행렬 형태의 데이터를 시각화하는 데 주로 사용되며, 각 셀의 색상은 해당 데이터 포인트의 값임
  - ✓트리맵(Treemap)은 여러 변수의 계층적 구조를 시각화한 차트로 계층적 데이터를 중첩된 직사각형들로 표현하는 차트임 각 직사각형의 크기는 해당 데이터 항목의 수량적 가치를 나타내며, 색상은 다른 범주나 그룹을 구분하는 데 사용될 수 있음
  - ✓네트워크 그래프(Network Graph)는 여러 변수의 연결 관계를 시각화한 차트로 노드(개체)와 엣지(관계)로 구성된 데이터 구조를 표현함. 노드는 점으로, 엣지는 선으로 표현되며, 복잡한 관계나 연결 구조를 보여줌

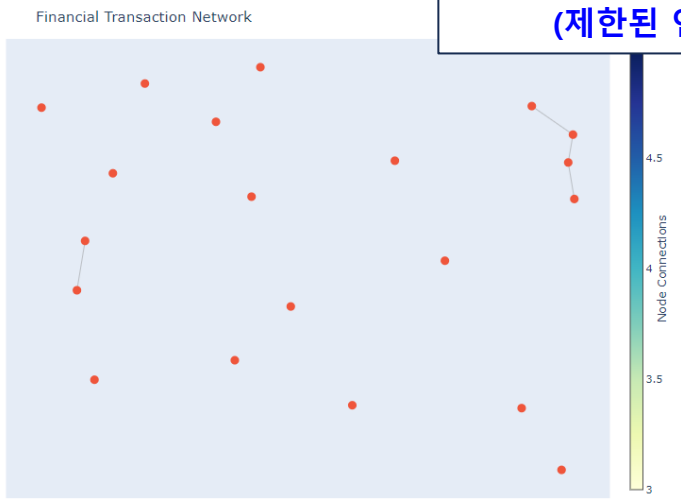
트리맵 : 글로볼 주식시장의 시가총액 구조



제한된 연결성만을 보여주고 있어 희소한(불완전 전 상황

네트워크 그래프 :

전체를 다 연결 못하고 소수의 데이터 포인트 간에 연결 (제한된 연결성)



Thank you