2과목.빅데이터 탐색

(Ch_02. 데이터 탐색 - SEC 01. 데이터 탐색의 기초

SEC 02. 고급 데이터 탐색)

빅데이터 분석 기사(2과목. 빅데이터 탐색)

CHAPTER 1. 데이터 전처리

CHAPTER 2. 데이터 탐색

CHAPTER 3. 통계 기법 이해

데이터 탐색

데이터 탐색 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 데이터 탐색의 기초
2. 고급 데이터 탐색

2. 데이터 탐색 – 데이터 탐색의 기초

01 데이터 탐색의 개요

1) 탐색적 데이터 분석(EDA: Exploratory Data Analysis)

; 수집한 데이터가 들어왔을 때, 다양한 방법을 통해서 자료를 관찰하고 이해하는 과정을 의미하는 것으로 본격적인 데이터 분석 전에 자료를 직관적인 방법으로 통찰하는 과정이다.

2) 탐색적 데이터 분석의 필요성

- 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 이해하며 내재된 잠재적 문제에 대해 인식하고 해결안을 도출할 수 있다.
  - ▶ 문제점 발견 시 본 분석 전 데이터의 수집 의사를 결정할 수 있다.
- 다양한 각도에서 데이터를 살펴보는 과정을 통해 문제 정의 단계에서 인지 못한 새로운 양상, 패턴을 발견할 수 있다.
  - ▶ 새로운 양상을 발견 시 초기설정 문제의 가설을 수정하거나 또는 새로운 가설을 설립할 수 있다.

2. 데이터 탐색 – 데이터 탐색의 기초

3) 분석과정 및 절차

- 분석의 목적과 변수가 무엇인지, 개별변수의 이름이나 설명을 가지는지 확인한다.
- 데이터의 문제성을 확인한다. 즉, 데이터의 결측치의 유무, 이상치의 유무 등을 확인하고 추가적으로 분포상의 이상 형태와 Head 또는 Tail 부분을 확인한다.
- 데이터의 개별 속성값이 예상한 범위 분포를 가지는지 확인한다(기초통계산출을 통한 확인과정을 거친다).
- 관계성 확인 절차를 가진다. 즉, 개별 데이터 간의 속성 관찰에서 보지 못한 데이터 간의 속성(예, 상관관계 등)을 확인한다.

4) 이상치의 검출

① 개별 데이터 관찰

▶ 데이터 값을 눈으로 살펴보면서 전체적인 추세와 특이사항을 관찰할 수 있다.

▶ 데이터가 많다고 앞부분만 보면 안 되고, 패턴이 뒤에서 나타날 수도 있으므로 뒤 or 무작위로 표본을 추출해서 관찰한다. 단, 이상치는 표본의 크기가 작은 경우 나타나지 않을 수도 있다.

2. 데이터 탐색 – 데이터 탐색의 기초

4) 이상치의 검출

② 통계값 활용

▶ 적절한 요약 통계지표(Summary Statistics)를 사용할 수 있다.

▶ 데이터의 중심을 알기 위해서는 평균(mean), 중앙값(median), 최빈값(mode)을 사용할 수 있다.

▶ 데이터의 분산도를 알기 위해서는 범위(range), 분산(variance)을 사용할 수 있다.

▶ 통계 지표를 이용할 때는 데이터의 특성에 주의해야 한다. 예를 들어, 평균에는 집합 내 모든 데이터 값이 반영되기 때문에, 이상값이 있으면 값이 영향을 받지만, 중앙값에는 가운데 위치한 값 하나가 사용되기 때문에 이상값의 존재에도 대표성이 있는 결과를 얻을 수 있다.

③ 시각화 활용

▶ 시각적인 표현은, 분석에 많은 도움을 준다. 시각화를 통해 주어진 데이터의 개별 속성에 어떤 통계 지표가 적절한지 결정할 수 있다.

- 시각화 방법에는 확률밀도함수, 히스토그램, 박스플롯(box plot), 워드 클라우드, 시계열 차트, 지도 등이 있다.

2. 데이터 탐색 – 데이터 탐색의 기초

개념 체크

01 탐색적 데이터 분석 및 필요성에 대한 설명으로 틀린 것은?

① 수집한 데이터가 들어왔을 때, 다양한 방법을 통해서 자료를 관찰하고 이해하는 과정을 의미하는 것이다.

② 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 이해할 수 있다.

③ 문제점 발견 시 본 분석 전 데이터의 수집 의사를 결정 할 수 있다.

④ 최초의 가설에 집중하여 원하는 패턴과 양상에 맞는지에 집중하여 검증하는 데 노력한다.

다양한 각도에서 데이터를 살펴보는 과정을 통해 문제 정의 단계에서 인지 못한 새로운 양상, 패턴을 발견할 수 있다. 그러므로, 새로운 양상, 패턴 발견 시에 초기 설정 문제의 가설을 수정을 하거나 혹은 새로운 가설을 추가적으로 설립할 수 있다.

02 탐색적 분석의 절차에 대한 설명이다. 다음 중 옳은 것을 모두 고른 것은?

(가) 분석의 목적과 변수가 무엇인지, 개별변수의 이름이나 설명을 가지는지 확인한다.
(나) 데이터의 문제성을 확인한다. 즉, 데이터의 결측치의 유무, 이상치의 유무 등을 확인하고 추가적으로 분포상의 이상 형태, Head 또는 Tail 부분을 확인한다.
(다) 데이터의 개별 속성 값이 예상한 범위 분포를 가지는지 확인한다.
(라) 관계속성 확인 절차를 가진다. 즉, 개별 데이터 간의 속성 관찰에서 보지 못한 데이터 간의 속성(예: 상관관계 등)을 확인한다.

① 가, 나

② 가, 다

③ 가, 나, 라

④ 가, 나, 다, 라

위의 내용은 탐색적 데이터 분석의 분석과정 및 절차의 4가지 항목이다.

2. 데이터 탐색 – 데이터 탐색의 기초

03 이상 발견의 통계적 기법 활용을 설명한 것 중 옳은 것은? .

① 데이터의 중심을 알기 위해서는 평균(mean), 중앙값 (median), 최빈값(mode), 첨도(kurtosis)를 사용할 수 있다.

② 데이터의 분산도를 알기 위해서는 범위(range), 분산 (variance), 왜도(skew-ness)를 사용할 수 있다.

③ 평균에는 집합 내 모든 데이터 값이 반영되기 때문에, 이상값의 영향을 받는다.

④ 중앙값은 전체변수의 범위 중에서 가운데 값을 사용 하므로 이상값의 크기에 영향을 받는다.

첨도, 왜도는 데이터의 분포모양에 해당된다.

첨도(Kurtosis) : 분포의 뽀족한(peakedness) 정도를 나타 내는 통계적 척도이다.

왜도(Skewness) : 왜도는 분포의 비대칭 정도를 나타내는 통계적 척도이다. 데이터 분포의 대칭성과 비대칭성을 정량화하여 평가하는데 사용된다.

중앙값은 전체변수의 범위에서 가운데가 아니라 관찰된 변수들 중에 가운데 값이므로 이상값의 영향을 받지 않는다.

2. 데이터 탐색 – 데이터 탐색의 기초

02 상관관계분석

1) 변수 간의 상관성 분석

; 두 변수 간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법이다. 두 변수는 서로 독립적인 관계이거나 상관된 관계일 수 있으며 이때 두 변수 간의 관계의 강도를 상관관계(correlation)라 한다.

① 단순상관분석(Simple Correlation Analysis) : 단순히 두 개의 변수가 어느 정도 강한 관계에 있는가를 측정한다.

② 다중상관분석(Multiple Correlation Analysis) : 3개 이상의 변수 간의 관계강도를 측정한다.

▶ 편상관 관계 분석(Partial Correlation Analysis) : 다중상관분석에서 다른 변수와의 관계를 고정하고 두 변수의 관계강도를 측정하는 것을 말한다.

2) 상관분석의 기본가정

① 선형성 : 두 변인 X와 Y의 관계가 직선적인지를 알아보는 것으로 이 가정은 분포를 나타내는 산점도를 통하여 확인할 수 있다.

② 동변량성 : X의 값에 관계없이 Y의 흩어진 정도가 같은 것을 의미한다. 반의어는 이분산성이다.

▶ 산포도가 특정 구간에 상관없이 퍼진 정도가 일정할 때 자료가 동변량성을 띤다고 얘기하며, 반대로 그 정도가 일정하지 않으면 이분산성을 보인다고 말한다.

산점도는 직교 좌표계를 이용해 두 개 변수 간의 관계를 나타내는 방법이며, 산포도는 변량이 흩어져 있는 정도를 하나의 수로 나타낸 값이다.

2. 데이터 탐색 – 데이터 탐색의 기초

2) 상관분석의 기본가정

③ 두 변인의 정규 분포성 : 두 변인의 측정치 분포가 모집단에서 모두 정규분포를 이루는 것이다.

④ 무선독립표본 : 모집단에서 표본을 뽑을 때 표본대상이 확률적으로 선정된다는 것이다.

3) 상관분석 방법

① 피어슨 상관계수(Pearson Correlation Coefficient 또는 Pearson's r)

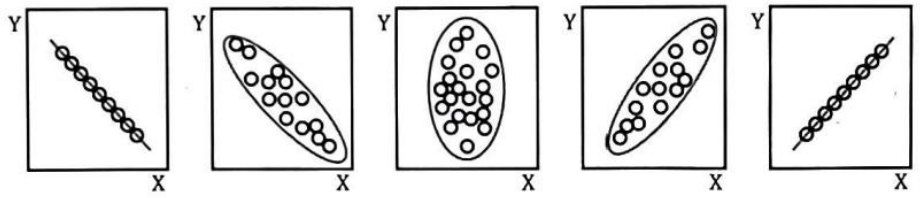
▶ 두 변수 X와Y 간의 선형 상관관계를 계량화한 수치이다.

▶ 피어슨 상관계수는 +1과 -1 사이의 값을 가지며, +1은 완벽한 양의 선형상관관계, 0은 선형 상관관계없음, -1은 완벽한 음의 선형상관관계를 의미한다.

$$\rho = \frac{C(x,y)}{\sigma_x \sigma_y}$$

$C(x,y) = Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$: 표본 간의 공분산

σ_x: 표본 x의 표준편차, σ_y: 표본 y의 표준편차, $\bar{x}, \bar{y}$: 각각의 평균



▲ 상관계수값과 데이터의 산점도

상관계수 : 두 변량 x, y 사이의 상관관계의 정도를 나타내는 수치
공분산(covariance) : 2개의 확률변수의 선형 관계를 나타내는 값이다

2. 데이터 탐색 – 데이터 탐색의 기초

3) 상관분석 방법

② 스피어만 상관계수(Spearman Correlation Coefficient)

▶ 데이터가 서열(순서형) 자료인 경우, 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다.

▶ 두 변수 간의 연관관계가 있는지 없는지를 밝혀 주며 자료에 이상점이 있거나 표본크기가 작을 때 유용하다.

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

▶ d_i^2 x_i의 순위(x_i의 관측치를 크기 순으로 정렬하였을 때 순위)와 y_i 의 순위 차이를 나타낸다. n은 표본의 개수이다.

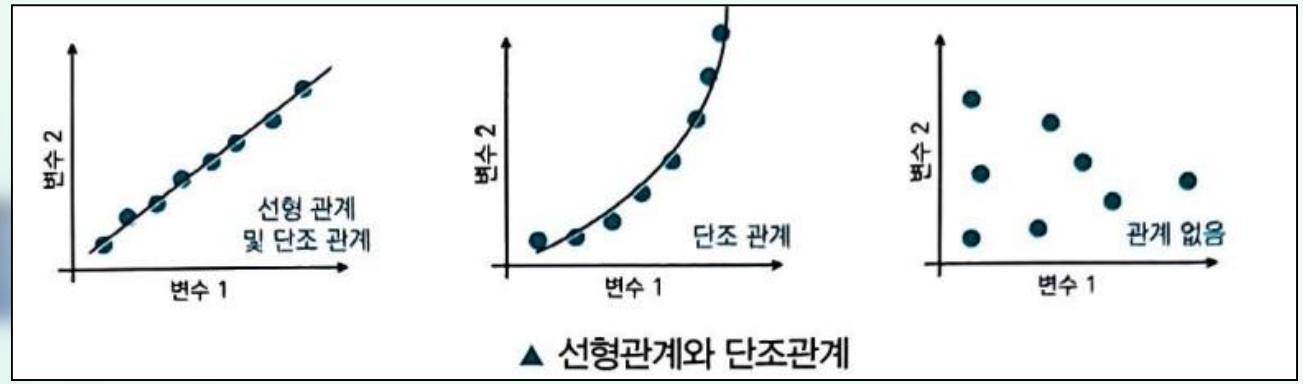
▶ 크기 순으로 정한 두 변수의 차이가 클수록 스피어만 상관계수의 값은 커진다. 즉 스피어만 상관계수는 한 변수의 값이 커지면 다른 변수의 값도 단조적으로 커지는지를 알아볼 수 있다.

스피어만 상관계수가 1에 가까울수록 두 변수는 단조적 상관성(커지면 같이 증가)을 가지는 것이고, 0에 가까우면 상관성이 없는 것으로 판단할 수 있다.

2. 데이터 탐색 – 데이터 탐색의 기초

3) 상관분석 방법

② 스피어만 상관계수(Spearman Correlation Coefficient)



▲ 선형관계와 단조관계

▶ 선형관계의 경우는 직선의 형태로 모형화가 가능한 것으로 해석될 수 있다. 단조 관계는 두 변수가 동일한 방향으로 변화는 하지만 직선의 형태로 모형화가 가능하지 않은(일정한 비율로 변화되는 것이 아닌) 것을 의미한다.

2. 데이터 탐색 – 데이터 탐색의 기초

개념 체크

01 상관분석의 기본가정에 대한 용어와 설명을 연결한 것 중 틀린 것은?

① 선형성 : 두 변인 X와 Y의 관계가 직선적인지를 알아보는 것으로 이 가정은 분포를 나타내는 산점도를 통하여 확인할 수 있다.

② 동변량성 : X의 값에 관계없이 Y의 흩어진 정도가 다른 정도를 의미한다.

③ 두 변인의 정규분포성 : 두 변인의 측정치 분포가 모집단에서 모두 정규분포를 이루는 것이다.

④ 무선독립표본 : 모집단에서 표본을 뽑을 때 표본대상이 확률적으로 선정된다는 것이다.

동변량성 : X의 값에 관계없이 Y의 값이 흩어진 정도가 같은 것을 의미한다. 반의어는 이분산성이다.

산포도가 특정 구간에 상관없이 퍼진 정도가 일정할 때 자료가 동변량성을 띤다고 얘기하고, 반대로 그 정도가 일정하지 않으면 이분산성을 보인다고 말한다.

02 피어슨 상관계수(Pearson Correlation Coefficient)에 대한 설명으로 옳은 것은?

① 두 변수 X와 Y 간의 비선형 상관관계를 계량화한 수치이다.

② 두 변수 간의 연관 관계가 있는지를 밝혀주며 자료에 이상점이 있거나 표본크기가 작을 때 유용하다.

③ 피어슨 상관계수는 +1과 -1 사이의 값을 가지며, +1은 완벽한 양의 선형 상관관계, 0은 선형 상관관계 없음, -1은 완벽한 음의 선형 상관관계를 의미한다.

④ 데이터가 서열자료인 경우 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다.

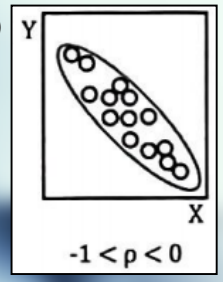
피어슨 상관계수

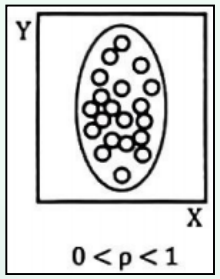
▶ 두 변수 X와 Y간의 선형 상관관계를 계량화한 수치이다.

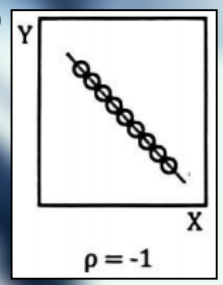
▶ 피어슨 상관계수는 +1과 -1사이의 값을 가지며, +1은 완벽한 양의 선형상관관계, 0은 선형상관 관계없음, -1은 완벽한 음의 선형상관관계를 의미한다.

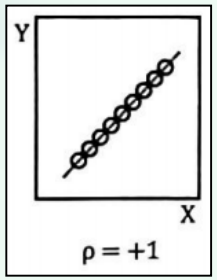
2. 데이터 탐색 – 데이터 탐색의 기초

03 피어슨 상관계수 값과 산점도 그림의 연결이 바르지 못한 것은?

① 

② 

③ 

④ 

04 다음 중 스피어만 상관계수에 대한 설명으로 틀린 것은?

① 데이터가 서열자료인 경우, 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수이다.

② 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다.

③ 두 변수 간의 연관관계가 있는지 없는지를 밝혀 준다.

④ 자료에 이상점이 있거나 표본크기가 클 때 유용하다.

스피어만 상관계수

▶ 데이터가 서열(순서형)자료인 경우, 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용하여 상관계수를 구한다.

▶ 두 변수 간의 연관관계가 있는지 없는지를 밝혀 주며 자료에 이상점이 있거나 표본크기가 작을 때 유용하다.

▶ 크기 순으로 정한 두 변수의 차이가 클수록 스피어만 상관계수의 값은 커진다. 즉 스피어만 상관계수는 한 변수 의 값이 커지면 다른 변수의 값도 단조적으로 커지는지를 알아볼 수 있다. 스피어만 상관계수가 1에 가까울수록 두 변수는

2. 데이터 탐색 – 데이터 탐색의 기초

03 기초통계량의 추출 및 이해

자료를 수집하여 요약·정리하는 기초통계(또는 기술통계)는 자료의 특성을 정량적인 수치에 의해서 나타내는 방법이다. 자료의 특성을 중심화 경향(Central Tendency), 퍼짐 정도(산포도·분산도), 자료의 분포형태(Shape of Distribution) 등의 수치적 결과로 나타낼 수 있다.

1) 중심화 경향 기초통계량

① 산술평균(Arithmetic Mean)

- ▶ 모든 자료들을 합한 후 전체 자료수로 나누어 계산하는 일반적인 평균을 의미한다.
- ▶ 모평균(Population Mean) : 모집단 전체 자료의 산술평균
- ▶ 표본평균(Sample Mean) : 모집단의 부분집합인 추출된 표본 전체의 산술평균

모집단 :통계적인 관찰의 대상이 되는 집단 전체를 의미하며, 표본을 뽑아내는 바탕이 된다.

2. 데이터 탐색 – 데이터 탐색의 기초

1) 중심화 경향 기초통계량

② 기하평균(Geometric Mean)

▶ N개의 자료에 대해서 관측치를 곱한 후 n 제곱근으로 표현한다.

$$\text{기하평균} = \sqrt[n]{x_1 \times x_2 \times x_3 \cdots \times x_n}$$

▶ 다기간의 수익률에 대한 평균 수익률, 평균물가상승률 등을 구할 때 사용한다.

예) 10,000원짜리 주식이 있다고 가정하고 10% 상승하고 10% 하락했다고 하면 산술평균적인 개념으로 봤을 때 가격변동이 없는 것으로 착각할 수 있다.
그러나 10,000 × 0.1 = 1000원이므로 11000원이 되고 이후 11,000 × 10% = 1,100원이므로 9,900원이 된다.
기하평균식으로 보면 다음 아래와 같다(수익률변환기준).

$$\sqrt{1.1 \times 0.9} = 0.995$$

0.995 – 1 = –0.005 즉, 평균 수익률은 –0.5% 손실로 해석할 수 있다.
최초 10,000원 – 50원 – 50원 ≒ 9,900원

2. 데이터 탐색 – 데이터 탐색의 기초

1) 중심화 경향 기초통계량

③ 조화평균(Harmonic Mean)

▶ 각 요소의 역수의 산술평균을 구한 후 다시 역수를 취하는 형태로 표현한다.

$$\text{조화평균} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$$

▶ 변화율 등의 평균을 구할 때 사용한다.

▶ 각 자료가 동일한 경우 자료에 대한 조화평균, 산술평균값과 기하평균의 값은 같다. 다만 자료가 서로 다를 경우 조화평균 ≤ 기하평균 ≤ 산술평균의 부등식 관계를 가진다.

④ 중앙값(Median)

▶ 중앙값은 자료를 크기순으로 나열할 때 가운데에 위치한 값이다.

▶ 자료의 수를 n이라 할 때, n이 홀수이면 (n+1)/2번째 자료값이 중앙값이 되고, n이 짝수이면, n/2번째와 1/2+1번째 자료의 평균을 중앙값으로 정의한다.

⑤ 최빈값(Mode)

▶ 가장 노출 빈도가 높은 자료를 최빈값이라 한다. 최빈값은 질적자료나 양적자료 모두에 사용된다.

역수 : 곱하여서 1이 되는 두 수의 각각을 다른 수에 대하여 이르는 말이다. 예를 들어 5의 역수는 1/5이다.

2. 데이터 탐색 – 데이터 탐색의 기초

1) 중심화 경향 기초통계량

⑥ 분위수(Quantile)

▶ 분위수는 자료의 위치를 표현하는 수치이다. 자료를 크기순서대로 배열을 한 후 그 자료를 분할하는 역할을 하는 위치의 수치를 계산한 것이다.

- 자료를 몇 등분 하느냐에 따라 사분위수(quartile), 오분위수(quintile), 십분위수(decile), 백분위수(percentile) 등이 있다.

- N개의 자료가 존재할 때 백분위수로 전환되는 분위수의 위치를 나타내는 식은 아래와 같다. 전체 y(%)가 해당 분위수의 하부에 위치한다.

$$\text{분위수의 위치} = (N+1)\frac{y}{100}$$

사분위수
자료를 동일한 비율로 4등분 할때 세 위치(25%, 50%, 75%) - Q1(제1사분위수) : 25% 지점, Q2(제2사분위수): 50% 지점, Q3(제3사분위수): 75% 지점

2. 데이터 탐색 – 데이터 탐색의 기초

2) 산포도(분산도, Degree Dispersion)

; 자료의 퍼짐 정도를 나타내는 기초 통계량이다. 중심 위치의 측도만으로 자료의 분포에 대한 충분한 정보를 얻을 수 없으므로 중심 경향도 수치에서 자료가 얼마나 떨어져 있는지를 측정하는 척도도 필요하다.

① 분산(Variance), 표준편차(Standard Deviation)

▶ 분산은 평균을 중심으로 밀집되거나 퍼짐 정도를 나타내는 척도이고, 표준편차는 분산의 제곱근으로 표현한다.

▶ 분산은 개개의 자료값과 평균과의 편차의 제곱을 이용하여 표현되므로 자료값의 단위를 제곱한 단위를 사용하게 된다. 분산으로 얻은 수치를 해석하기가 곤란하다는 단점을 보완하기 위하여 제곱근을 취한 척도가 표준편차이다.

예) 여러 자산에 대한 투자수익률이 다음과 같이 제시되었을 때 분산과 표준편차를 구하여 보면

12%, 20%, 23%, 25%, 30%

일단 평균은 $\frac{12\% + 20\% + 23\% + 25\% + 30\%}{5} = 22\%$

분산(모분산)은

$$\sigma^2 = \frac{(12\%-22\%)^2 + (20\%-22\%)^2 + (23\%-22\%)^2 + (25\%-22\%)^2 + (30\%-22\%)^2}{5} = 35.6$$

이 되고 표준편차는 $\sqrt{35.6} = 5.97\%$가 된다.

산포도 = 분산도 = 퍼짐 정도

2. 데이터 탐색 – 데이터 탐색의 기초

2) 산포도(분산도, Degree Dispersion)

▶ 분산의 특성

- 개개의 자료값에 대한 정보를 반영한다.
- 수리적으로 다루기 쉽다.
- 특이점에 매우 큰 영향을 받는다.
- 분산이 클수록 각 자료값이 평균으로부터 넓게 흩어진 형태를 갖는다.
- 미지의 모분산을 추론할 때 많이 사용한다.

② 범위(Range)

▶ 데이터 간의 최댓값과 최솟값의 차이를 나타내는 것으로 동일한 범위를 갖더라도 자료의 분포모양은 다를 수가 있음에 유의해야 한다.

③ 평균 절대 편차(평균 편차, 절대 편차, MAD: Mean Absolute Deviation)

▶ 각 자료값과 표본평균과의 편차의 절댓값에 대한 산술평균을 의미한다.

평균절대편차 : 관측값에서 평균을 빼고, 그 차이값에 절댓값을 취하고, 그 값들을 모두 더하여 전체 데이터 개수로 나눠 준 것

2. 데이터 탐색 – 데이터 탐색의 기초

2) 산포도(분산도, Degree Dispersion)

④ 사분위범위(Inter Quartile Range)

▶ 자료를 크기순으로 배열 후 자료의 1/4에 해당하는 1사분위수(Q1)를 구하고 3/4에 해당하는 3사분위수(Q3)를 구한다. 사분위범위는 Q3-Q1으로 정의되며 자료의 50% 범위 내에 위치하게 됨을 의미한다.

▶ 사분위범위는 주로 이상치의 판단 시에 사용되는 것으로 결정된 최대값보다 크거나 최소값보다 작은 값을 이상치로 간주한다.

- 최대값 = 75% percentile: 3 사분위수 + 1.5 × IQR
- 최소값 = 25% percentile: 1 사분위수 - 1.5 × IQR

⑤ 변동계수(CV: Coefficient of Variance)

▶ 평균을 중심으로 한 상대적인 산포의 척도를 나타내는 수치이다.

▶ 측정 단위가 동일하지만 평균이 큰 차이를 보이는 두 자료집단 또는 측정단위가 서로 다른 두 자료집단에 대한 산포의 척도를 비교할 때 많이 사용한다.

▶ 변동계수가 클수록 상대적으로 넓게 분포를 이룬다.

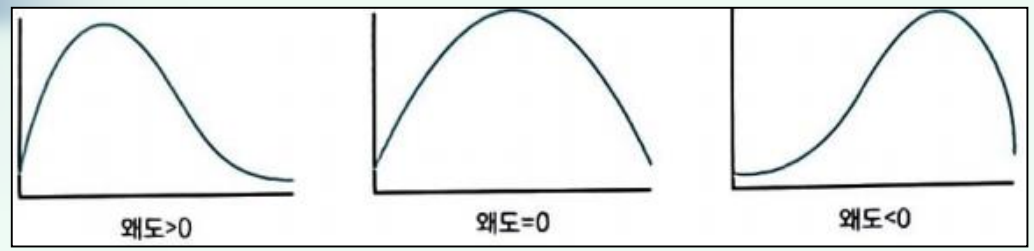
2. 데이터 탐색 – 데이터 탐색의 기초

3) 자료의 분포형태(Shape of Distribution)

① 왜도(Skewness)

▶ 왜도는 분포의 비대칭(asymmetry, 기울어진 정도) 정도를 나타내는 통계적 측도이다. 데이터 분포의 대칭성과 비대칭성을 정량화하여 평가하는데 사용된다.

▶ 분포가 대칭이면 왜도는 0이다. 왼쪽으로 치우친 경우 왜도는 양수, 오른쪽으로 치우친 경우 왜도는 음수이다.



▶ 왜도는 분포의 모양 뿐만 아니라 이상치의 존재 여부를 파악하는 데에도 도움을 줄 수 있다. 이상치는 분포의 비대칭성을 높이고, 왜도의 크기를 변화시킨다.

▶ 왜도의 값은 일반적으로 -3과 +3사이의 범위에 있으며, 보통 왜도의 절대값이 1.96보다 크면 비대칭성이 있다고 판단할 수 있다. 하지만 이것은 규칙적 기준은 아니며, 데이터 분포의 특성과 분석 목적에 따라 다르다.

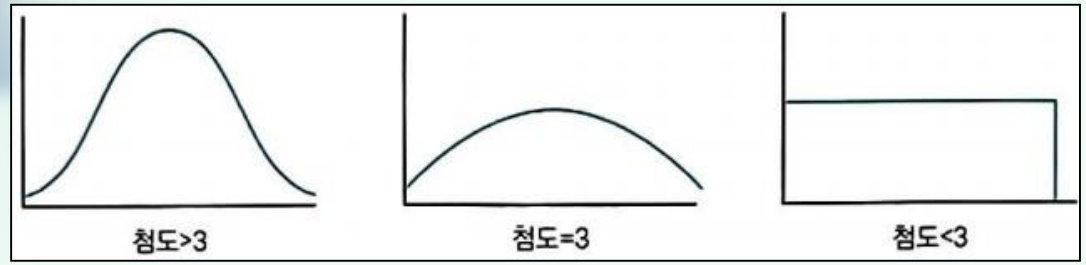
2. 데이터 탐색 – 데이터 탐색의 기초

3) 자료의 분포형태(Shape of Distribution)

② 첨도(Kurtosis)

▶ 분포의 뾰족한(peakedness) 정도를 나타내는 통계적 척도이다.

▶ 첨도의 값이 3 미만인 경우는 평평한 분포이고 3이면 정규분포를 나타내며 3이 넘는 경우는 뾰족한 분포의 형태를 가지는 것으로 판단할 수 있다.



2. 데이터 탐색 – 데이터 탐색의 기초

개념 체크

01 자료의 특성을 수치적 결과로 나타내는 기초통계 방법을 나열한 것이다. 다음 중 성질이 다른 하나는?

① 산술평균 ② 기하평균

③ 최빈값 ④ 범위

산술평균, 기하평균, 최빈값, 조화평균, 중앙값, 분위수는 중심화 경향 기초 통계량이고, 범위, 분산, 표준편차, 평균 절대 편차, 사분위범위, 변동계수는 산포도(퍼짐 정도)에 대한 기초 통계량이다.

산술평균

▶ 모든 자료들을 합한 후 전체 자료수로 나누어 계산하는 일반적인 평균을 의미한다.

기하평균

▶ N개의 자료에 대해서 관측치를 곱한 후 n제곱근으로 표현한다.

최빈값(Mode)

▶ 가장 노출 빈도가 높은 자료를 최빈값이라고 한다.

범위(Range)

02 다음 아래의 자료에서 분산, 평균, 중앙값을 구하시오.

101, 103, 105, 107, 109

① 평균 105, 분산 7.5, 중앙값 105

② 평균 104.5, 분산 8, 중앙값 104.5

③ 평균 105, 분산 8, 중앙값 105

④ 평균 106.5, 분산 8.5, 중앙값 106.5

평균 = 101 + 103 + 105 + 107 + 109 / 5

525 / 5 = 105가 평균이 된다.

분산 = $(101-105)^2$ + $(103-105)^2$ + $(105-105)^2$ + $(107-105)^2$ + $(109-105)^2$ / 5

16 + 4 + 0 + 4 + 16 = 40

40 / 5 = 8 이 분산값이 된다.

중앙값은 홀수이므로 105가 된다.

2. 데이터 탐색 – 데이터 탐색의 기초

03 포트폴리오의 투자수익률, GDP 성장률 등의 연간 자료에 대해서 알맞은 기술적 통계량인 평균은 무엇인가?

① 산술평균 ② 조화평균

③ 기술평균 ④ 기하평균

조화평균(Harmonic Mean)

▶ 각 요소의 역수의 산술평균을 구한 후 다시 역수를 취하는 형태로 표현된다.

▶ 변화율 등의 평균을 구할 때 사용한다.

04 다음 중 중심 경향성 통계량에 속하지 않는 것은?

① 중위수

② 최빈값

③ 분산

④ 사분위수

분산은 산포도(퍼짐 정도) 통계량에 속한다.

05 다음 데이터 중 최빈수는?

5, 7, 3, 2, 1, 2, 4, 2, 5, 6, 7, 2

① 2 ② 1

③ 7 ④ 5

최빈수는 가장 빈도수가 높은 수를 의미한다. 주어진 총 12개의 데이터 중 가장 많은 빈도수를 갖는 데이터는 빈도수 4를 갖는 2가 된다.

2. 데이터 탐색 – 데이터 탐색의 기초

06 다음 중 산포도 통계량에 대한 설명으로 틀린 것은?

① 분산은 평균으로부터 얼마나 떨어져 있는지를 나타내는 값이다.

② 표준편차는 분산에 양의 제곱근을 취한 값이다.

③ 범위는 데이터의 최댓값과 최솟값의 차를 나타낸다.

④ 산포도 통계량으로 데이터가 기울어진 정도를 확인할 수 있다.

산포도 통계량으로는 데이터의 흩어진 정도를 확인할 수 있다. 데이터의 기울어진 정도를 알 수 있는 통계량은 분포 통계량 중에 왜도에 대한 내용이다.

왜도(Skewness)

▶ 왜도는 분포의 비대칭(기울어진 정도) 정도를 나타내는 통계적 측도이다. 데이터 분포의 대칭성과 비대칭성을 정량화하여 평가하는데 사용된다.

▶ 분포가 대칭이면 왜도는 0이다. 왼쪽으로 치우친 경우 왜도는 양수, 오른쪽으로 치우친 경우 왜도는 음수이다.

▶ 왜도는 분포의 모양 뿐만 아니라 이상치의 존재 여부를 파악하는 데에도 도움을 줄 수 있다.

07 다음은 어떤 통계량에 대한 설명인가?

데이터 분포의 기울어진 정도를 설명하는 통계량

① 첨도 ② 왜도

③ 분산 ④ 표준편차

첨도(Kurtosis)

▶ 분포의 뾰족한(peakedness) 정도를 나타내는 통계적 척도이다.

▶ 첨도의 값이 3미만인 경우는 평평한 분포이고 3이면 정규분포를 나타내며 3이 넘는 경우는 뾰족한 분포의 형태를 가지는 것으로 판단할 수 있다.

2. 데이터 탐색 – 데이터 탐색의 기초

04 시각적 데이터 탐색

시각화를 통한 탐색적 자료분석은 기본적으로 전통적 통계차트 및 다이어그램에 의존하는 부분에 대해 설명 하며 좀 더 심화된 데이터 시각화는 뒤에서 다룬다.

1) 통계적 시각화 도구

① 도수분포표(Frequency Table) : 수집된 자료를 적절한 계급에 의해 분류하여 정리한 표로 질적 자료의 경우는 각 자료값(범주)에 대하여 도수나 상대도수로 표현한다.

상품	도수	상대도수
콘 형태 아이스크림	65	65 / 100 = 0.65
막대 형태 아이스크림	25	25 / 100 = 0.25
기타	10	10 / 100 = 0.1
합계	100	1.0

▶ 도수(Frequency) : 질적 자료의 경우 각 범주별 빈도

▶ 상대도수(Relative Frequency) : 도수 / 전체 자료 수

▶ 양적 자료의 경우는 전체 자료를 그룹화(계급 구간)하고 각 그룹별 속하는 자료의 수를 계산하여 도수 및 상대도수로 표현한다.

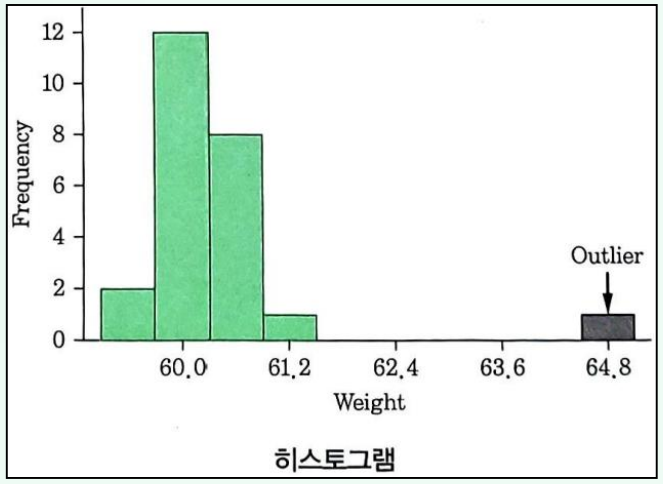
질적 자료 : 관찰 값이 수적 의미가 없이 범주만 나타내는 자료로 예시로는 성별, 연령층(10대, 20대) 등
양적 자료 : 관찰 값이 수적 의미를 나타내는 자료로써 예시로는 온도, 가격, 주가, 매출액 등 가능한 자료
명목형 자료 : 숫자 1, 2로 표기되는 기호적 의미만 있는 자료로 예시로는 남자, 여자를 1, 2로 수적 의미가 아니라 남자/여자를 구분해 주는 상징적 의미를 가지는 자료이다.

2. 데이터 탐색 – 데이터 탐색의 기초

1) 통계적 시각화 도구

② 히스토그램(Histogram) : 도수분포표를 이용하여 표본의 자료분포를 나타낸 그래프이다.

- ▶ 도수분포표의 각 계급의 양 끝 값을 가로축에 표시하고 그 계급의 도수를 세로축에 표시하여 직사각형 모양으로 나타낸다.
- ▶ 히스토그램은 가로축에 반드시 수량을 표시하지만, 막대그래프는 그렇지 않다.
- ▶ 히스토그램자료의 분포가 직사각형 형태이다.
- ▶ 막대는 서로 붙어있고, 막대의 너비는 일정하다.
- ▶ 이상값 확인이 가능하다(히스토그램 양쪽 끝, 고립된 막대).



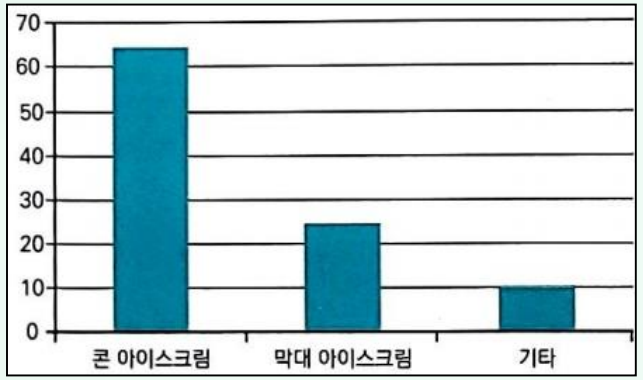
히스토그램

2. 데이터 탐색 – 데이터 탐색의 기초

1) 통계적 시각화 도구

③ 막대그래프(Bar Chart)

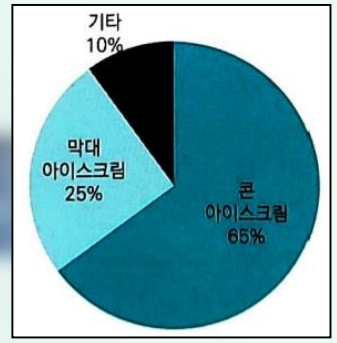
- 각 자료값에 대한 도수 또는 상대도수를 그림으로 표현한 것이다.
- 여러 가지 항목들에 대한 많고 적음을 표현한다.
- 막대그래프 가로축은 수치형 데이터가 아니어도 된다.
- 막대는 서로 떨어져 있다.
- 막대 너비는 같지 않을 수 있다.



2. 데이터 탐색 – 데이터 탐색의 기초

1) 통계적 시각화 도구

④ 파이차트(Pie Chart) : 각 자료값의 상대도수로 기입하여 원의 면적에 각 상대 크기별로 나타낸 그래프이다.

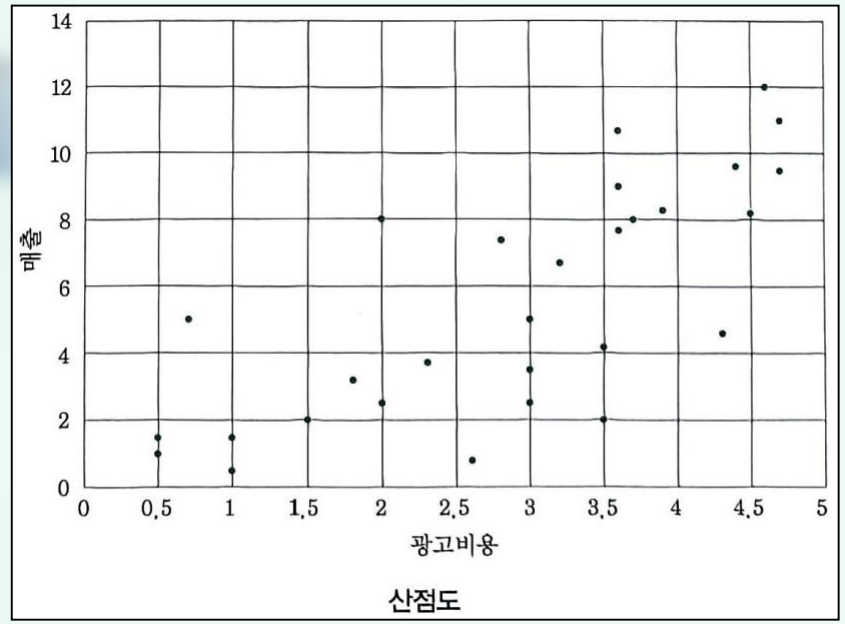


2. 데이터 탐색 – 데이터 탐색의 기초

1) 통계적 시각화 도구

⑤ 산점도(Scatter Plot)

- ▶ 가로축과 세로축의 좌표 평면상에서 각각의 관찰점들을 표시하는 시각화 방법이다.
- ▶ 2개의 연속형 변수 간의 관계를 보기 위해 사용한다.



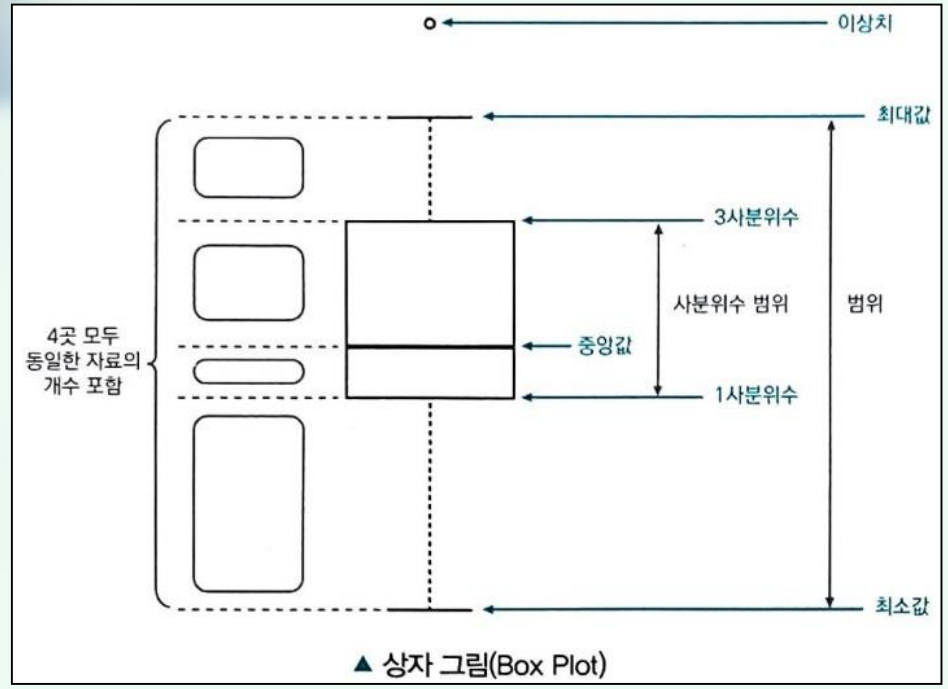
⑥ 줄기 잎 그림(Stem-and-Leaf Diagram) : 통계적 자료를 표 형태와 그래프 형태의 혼합된 방법으로 나타내는 것을 말한다. 줄기 잎 그림은 자료의 정리가 가능할 뿐 아니라 자료의 구조에 대한 정보도 파악이 가능한 도구이다.

2. 데이터 탐색 – 데이터 탐색의 기초

1) 통계적 시각화 도구

⑦ 상자 수염 그림(Box Plot)

▶ 수치적 자료를 표현하는 그래프이다. 이 그래프는 가공하지 않은 자료 그대로를 이용하여 그린 것이 아니라, 자료로부터 얻어 낸 통계량인 5가지 요약 수치(다섯 숫자 요약, Five-number Summary)를 가지고 그린다.



▲ 상자 그림(Box Plot)

3사분위수(Q3) : 데이터를 정렬했을 때 75%에 위치한 수
1사분위수(Q1) : 데이터를 정렬했을 때 25%에 위치한 수
사분위범위(IQR) : 3사분위수 - 1사분위수
최댓값 : 3사분위수 + (1.5 * IQR)
최솟값 : 1사분위수 - (1.5 * IQR)

▶ 5가지 요약 수치 : 최솟값, 제1사분위(Q1), 제2사분위(Q2), 제3사분위(Q3), 최댓값을 일컫는다.

2. 데이터 탐색 – 데이터 탐색의 기초

개념 체크

01 다음과 같은 특징을 갖는 시각화 도구는?

- 자료의 분포가 직사각형 형태
- 가로축은 수치형 데이터
- 막대는 서로 붙어있고, 막대의 너비는 일정함
- 이상값 확인 가능

① 산점도

② 카토그램

③ 히스토그램

④ 막대형 그래프

히스토그램(Histogram) : 도수분포표를 이용하여 표본의 자료 분포를 나타낸 그래프이다.

▶ 도수분포표의 각 계급의 양 끝 값을 가로축에 표시하고 그 계급의 도수를 세로축에 표시하여 직사각형 모양으로 나타낸다.

▶ 히스토그램은 가로축에 반드시 수량(수치적 데이터)을 표시하지만, 막대그래프는 그렇치 않다.

▶ 히스토그램은 막대가 서로 붙어있고, 막대의 너비는

02 다음 중 막대형 그래프에 대한 설명으로 틀린 것은?

① 여러 가지 항목들에 대한 많고 적음을 표현하는 그래프이다.

② 막대그래프 가로축은 반드시 수치형 데이터로 표현한다.

③ 막대는 서로 떨어져 있다.

④ 막대 너비는 같지 않을 수 있다.

막대형 그래프의 가로축은 수치형이 아니어도 된다.

막대 그래프(Bar Chart)

▶ 각 자료값에 대한 도수 또는 상대도수를 그림으로 표현한 것이다.

▶ 여러 가지 항목들에 대한 많고 적음을 표현한다.

▶ 막대 그래프 가로축은 수치형 데이터가 아니어도 된다.

▶ 막대는 서로 떨어져 있다.

▶ 막대 너비는 같지 않을 수 있다.

03 다음 중 박스플롯에 대한 설명으로 틀린 것은?

① 최소값은 제1사분위에서 1.5×IQR을 뺀 위치를 의미 한다.

② 제1사분위(Q1)는 자료들의 하위 25%의 위치를 의미한다.

③ 제3사분위(Q3)는 자료들의 하위 75%의 위치를 의미한다.

2. 데이터 탐색 – 데이터 탐색의 기초

04 다음 중 의미가 다른 그래프는?

① 스타차트

② 박스플롯

③ 상자수염그림

④ 히스토그램

박스플롯, 상자수염그림, 히스토그램은 모두 같은 의미의 통계적 시각화 도구이며 그래프를 나타낸다.
스타차트는 비교 시각화의 한 종류로써 하나의 공간에 각각의 변수를 표현하는 몇 개의 축을 그리고, 축에 표시된 해당 변수의 값들을 연결하여 별 모양(또는 거미줄 모양)으로 표현하는 그래프이다.

05 다음은 어떤 시각화 도구에 대한 설명인가?

- 가로축과 세로축의 좌표 평면상에서 각각의 관찰점들을 표시하는 시각화 방법
- 2개의 연속형 변수 간의 관계를 보기 위해 사용

① 히스토그램

② 막대형 그래프

③ 산점도

④ 상자그림

2. 데이터 탐색 – 고급 데이터 탐색

01 시공간 데이터 탐색

1) 시공간 데이터(Spatio-Temporal Data)

- 시공간 데이터는 공간적 정보에 시간의 흐름이 결합된 다차원 데이터를 의미한다.
- 시공간 데이터는 점(Point), 선(Line), 면(Polygon)과 같은 형태로 시각화될 수 있다.

2) 시공간 데이터 탐색 방법

- 시공간 데이터의 탐색 절차는 주소를 행정구역 및 좌표계로 변환하고, 변환된 행정구역과 좌표계를 지도에 표시하는 순이다.
- 시공간 데이터를 지도에 표시하는 방법에는 코로플레스 지도, 카토그램, 버블 플롯맵 등이 있다.

2. 데이터 탐색 – 고급 데이터 탐색

2) 시공간 데이터 탐색 방법

① 코로플레스 지도(Choropleth Map)

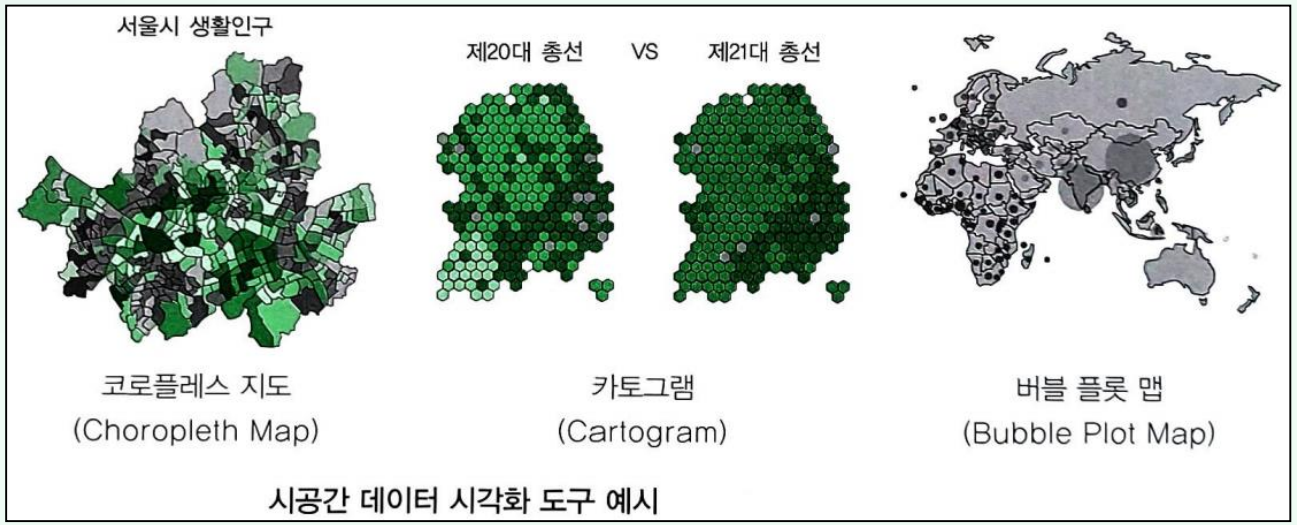
▶ 등치지역도라고도 하고, 어떤 데이터 값의 크기에 따라 해당 영역을 색칠해서 표현하는 방법이다.

② 카토그램(Cartogram)

▶ 변량비례도라고도 하고, 데이터 값 크기에 따라 면적을 왜곡하여 표현하는 방법이다.

③ 버블 플롯 맵(Bubble Plot Map)

▶ 위도와 경도를 적용하여 좌표를 원으로 표현하고, 원의 색깔과 크기로 데이터를 표현하는 방법이다.



시공간 데이터 시각화 도구 예시

2. 데이터 탐색 – 고급 데이터 탐색

02 다변량 데이터 탐색

1) 변량 데이터의 유형

- 변량(Variance)이란 조사 대상의 특징을 숫자나 문자로 나타낸 값을 의미한다.
- 변량은 종속변수(Y)의 수에 따라서 일변량, 이변량, 다변량으로 구분된다.

유형	설명
일변량 데이터 (단변량 데이터)	단위에 대해 하나의 속성만 측정하여 얻게 되는 변수에 대한 자료이다.
이변량 데이터	각 단위에 대해 두 개의 특성을 측정하여 얻어진 두 개의 변수에 대한 자료이며, 다변량 데이터에 속한다.
다변량 데이터	하나의 단위에 대해 두 가지 이상의 특성을 측정하는 경우 얻어지는 변수에 대한 자료이다.

2. 데이터 탐색 – 고급 데이터 탐색

2) 변량 데이터의 탐색 방법

① 일변량 데이터 탐색

- ▶ 일변량 데이터는 기술 통계량, 그래프 통계량을 활용하여 탐색한다.
- ▶ 기술 통계량에는 분산, 표준편차, 평균 등을 사용하고, 그래프 통계량에는 히스토그램, 상자그림을 사용한다.

② 이변량 데이터 탐색

- ▶ 조사 대상의 각 개체로부터 두 개의 특성을 동시에 관측한다.
- ▶ 일반적으로 두 변수 사이의 관계를 확인하는 것이 목적이다.

③ 다변량 데이터 탐색

- ▶ 산점도 행렬, 별 그림, 등고선 그림을 사용하여 데이터를 시각적으로 탐색한다.

2. 데이터 탐색 – 고급 데이터 탐색

03 비정형 데이터 탐색

1) 비정형 데이터(Unstructured Data)

; 비정형 데이터는 이미지, 영상, 텍스트 데이터와 같이 형태가 구조화 되지 않은 데이터를 의미한다.

2) 비정형 데이터 탐색 방법

; 데이터의 특징에 맞게 비정형 데이터를 탐색한다.

탐색 방법	설명
텍스트 탐색 방법	온라인상의 소셜 데이터의 텍스트와 같은 스크립트 파일 형태인 경우 데이터를 파싱(Parsing)한 후 탐색한다.
동영상, 이미지 탐색 방법	이진 파일 형태의 데이터의 경우 데이터의 종류별로 응용 소프트웨어를 활용하여 탐색한다.
XML, JSON, HTML 탐색 방법	XML, JSON, HTML 각각의 파서(Parser)를 이용하여 데이터를 파싱한 후 탐색한다.

파싱(Parsing) : 프로그래밍에서 특정 형식으로 구성된 데이터를 분석하고 그 의미를 이해하는 과정을 의미한다.
하지만 데이터 분석에서는 데이터를 조립하여 특정한 데이터를 추출할 수 있도록 프로그램 하는 작업을 칭한다.
파서(Parser) : 파싱(Parsing)을 수행하는 프로그램이다.

2. 데이터 탐색 – 고급 데이터 탐색

개념 체크

01 다음 중 시공간 데이터(Spatio-Temporal Data)로 알맞지 않은 것은?

① 점 ② 선

③ 면 ④ 행렬

시공간 데이터(Spatio-Temporal Data)

▶ 시공간 데이터는 공간적 정보에 시간의 흐름이 결합된 다차원 데이터를 의미한다.

▶ 시공간 데이터는 점(Point), 선(Line), 면(Polygon)과 같은 형태로 시각화될 수 있다.

02 다음 중, 시공간 데이터 탐색 방법으로 틀린 것은?

① 코로플레스 지도

② 카토그램

③ 버블 플롯맵

④ 박스플롯

시공간 데이터 탐색 방법

1. 코로플레스 지도(Choropleth Map)

▶ 등치지역도라고도 하고, 어떤 데이터 값의 크기에 따라 해당 영역을 색칠해서 표현하는 방법이다.

2. 카토그램(Catogram)

▶ 변량비례도라고도 하고, 데이터 값 크기에 따라 면적을 왜곡하여 표현하는 방법이다.

3. 버블 플롯 맵(Bubble Plot Map)

▶ 위도와 경도를 적용하여 좌표를 원으로 표현하고, 원의 색깔과 크기로 데이터를 표현하는 방법이다.

2. 데이터 탐색 – 고급 데이터 탐색

03 다음 중 변량 데이터의 유형이 아닌 것은?

① 일변량 데이터

② 이변량 데이터

③ 다변량 데이터

④ 공변량 데이터

변량 데이터의 유형

▶ 변량(Variance)이란 조사 대상의 특징을 숫자나 문자로 나타낸 값을 의미한다.

▶ 변량은 종속변수(Y)의 수에 따라서 일변량, 이변량, 다변량으로 구분된다.

1. **일변량 데이터(단변량 데이터)** : 단위에 대해 하나의 속성만 측정하여 얻게 되는 변수에 대한 자료이다.
2. **이변량 데이터** : 각 단위에 대해 두 개의 특성을 측정하여 얻어진 두 개의 변수에 대한 자료이며, 다변량 데이터에 속한다.
3. **다변량 데이터** : 하나의 단위에 대해 두 가지 이상의 특성을 측정하는 경우 얻어지는 변수에 대한 자료이다.

04 다음 중, 비정형 데이터 탐색 방법이 아닌 것은?

① HTML 탐색 방법

② 동영상 탐색 방법

③ 이미지 탐색 방법

④ 수치 데이터 탐색 방법

비정형 데이터의 탐색 방법

1. **텍스트 탐색 방법** : 온라인상 소셜 데이터의 텍스트와 같은 스크립트 파일 형태인 경우 데이터를 파싱(Parsing)한 후 탐색한다.
2. **동영상, 이미지 탐색 방법** : 이진 파일 형태의 데이터의 경우 데이터의 종류별로 응용 소프트웨어를 활용하여 탐색한다.
3. **XML, JSON, HTML 탐색 방법** : 각각의 파서(Parser)를 이용하여 데이터를 파싱한 후 탐색한다.

2. 데이터 탐색 예상 문제

예상 문제

01 다음은 어떤 방법에 대한 설명인가?

- 노키아 벨 연구소(Nokia Bell Labs)의 수학자 존 튜키(John Tukey)가 개발한 개념으로, 데이터를 분석하고 결과를 내는 과정에서 지속적으로 해당 데이터에 대한 '탐색과 이해'를 기본으로 가져야 한다는 것을 의미한다.
- 데이터 분석 분야에서 탐색적 데이터 분석은 수집된 데이터를 다양한 방법을 활용하여 탐색적으로 분석하여 데이터의 특징을 정확하게 파악하는 것이라고 할 수 있다.

① EDA ② PCA

③ LDA ④ ICA

탐색적 데이터 분석(EDA; Exploratory Data Analysis)에 대한 설명이다. PCA는 주성분 분석, LDA는 선형 판별 분석, ICA는 독립 성분 분석을 나타낸다.

02 다음 다차원 데이터 탐색 방법 중 데이터 간의 산점도와 기울기를 통해 변수 간의 상관성을 분석하는 데이터 조합 유형은?

① 범주형 ↔ 범주형

② 수치형 ↔ 수치형

③ 범주형 ↔ 수치형

④ 다중형 ↔ 일반형

다차원 데이터 탐색 방법 중 데이터 간의 산점도와 기울기를 통해 변수 간의 상관성을 분석하는 데이터 조합 유형은 수치형 ↔ 수치형이다.

03 개별 변수 탐색 방법 중 다음이 설명하는 데이터 유형은?

- 명목형 변수와 순서형 변수에 대한 데이터 탐색 방법
- 빈도수, 최빈값, 비율, 백분율 등을 활용하여 데이터 분포의 특징을 중심성, 변동성 측면에서 파악
- 시각화는 막대형 그래프(Bar Plot)를 주로 사용

① 포괄형 데이터

2. 데이터 탐색 예상 문제

04 다음 중 탐색적 데이터 분석의 특징에 속하지 않는 것은?

① 저항성

② 영구성

③ 현시성

④ 잔차 해석

탐색적 데이터 분석의 특징으로는 **저항성, 잔차 해석, 자료 재표현, 현시성**이 있다.

저항성(Resistance) : 오류의 영향을 적게 받는 성질로써 저항성이 큰 데이터를 사용한다.

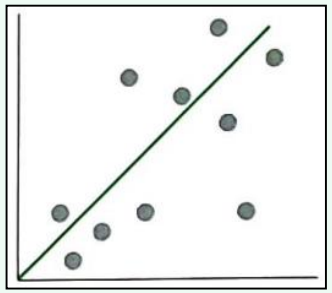
잔차 해석(Redidual) : 잔차는 관찰값들이 주 경향으로부터 벗어난 정도이며, 이를 해석하며 데이터의 특징을 파악한다.

자료 재표현(Re-expression) : 데이터 분석 및 해석의 용이성을 위해 변수를 적당한 척도로 바꾸는 것이다.

현시성(Graphic Representation) : 데이터 시각화라고도 할 수 있으며, 분석 결과를 쉽게 이해할 수 있도록 데이터를 시각적으로 표현하는 것이다.

05 다음 중 데이터 항목들을 그룹으로 간주하고, 각 그룹에 따라 수치형 변수의 기술 통계량 차이를 상호 비교하고

06 다음과 같은 그래프의 형태가 갖는 상관관계는?



① 상관관계 없음

② 약한 음(-)의 상관관계

③ 강한 양(+)의 상관관계

④ 약한 양(+)의 상관관계

양(+)의 상관관계

▶ 하나의 변수 값이 증가할 때 다른 변수의 값도 함께 증가하는 경향을 보이는 관계

▶ 관계의 정도에 따라 강한 양의 상관관계, 약한 양의 상관관계로 구분된다.

음(-)의 상관관계

▶ 하나의 변수 값이 증가할 때 다른 변수의 값이 감소하는 경향을 보이는 관계

▶ 관계의 정도에 따라 강한 음의 상관관계, 약한 음의

2. 데이터 탐색 예상 문제

08 다음 상관계수에 대한 설명 중 틀린 것은?

① 상관계수는 γ로 표시할 수 있다.

② 상관계수의 범위는 0 ~ 1이다.

③ 상관계수를 통해 변수의 상관관계를 확인할 수 있다.

④ 상관계수가 0에 가까울수록 상관관계가 없다고 해석할 수 있다.

상관계수(Correlation Coefficient)는 두 변수 X, Y 사이의 연관성을 수치로 나타낸 상관계수를 활용하여 변수 사이의 관계를 확인하는 방법이다. 상관계수의 -1 ~ +1의 범위를 갖고, 1에 가까울수록 강한 양의 상관관계를, -1에 가까울수록 강한 음의 상관관계를 가지며, 0에 가까울수록 상관관계가 없음을 의미한다.

09. 다음 중 보통의 양의 상관관계를 갖는 상관계수 범위는?

① $0.7 < \gamma < 1.0$

② $0.1 < \gamma < 0.3$

③ $0.3 < \gamma < 0.7$

④ $-0.1 < \gamma < 0.1$

상관계수(γ) 범위

10 다음 중 변수의 속성이 다른 하나는?

① 성별 ② 키

③ 몸무게 ④ 나이

키, 나이, 몸무게는 수치형 데이터이고, 성별은 명목형 변수이다.

명목형 데이터

▶ 명목형 데이터는 여러 카테고리(분류)들 중 하나의 이름으로 분류된 데이터를 의미한다.

▶ 명목형 데이터 분석 방법에는 카이제곱 검정(교차 분석)이 있다.

▶ 카이제곱 검정은 명목형 변수 사이의 관찰된 빈도가 기대되는 빈도와 의미있게 다른지의 여부를 검정하기 위해 사용되는 검정 방법이다.

11 다음 중 분석변수 속성과 분석 방법이 잘못 짝지어진 것은?

① 수치형 데이터 - 피어슨 상관계수

② 순서형 데이터 - 스피어만 상관계수

③ 명목형 데이터 - 카이제곱 검정

④ 명목형 데이터 - T 검정

2. 데이터 탐색 예상 문제

12 다음 중 중심 경향성 통계량에 속하지 않는 것은?

① 중위수 ② 최빈값

③ 분산 ④ 사분위수

중심 경향성 통계량에 속하는 것은 평균값, 중위수, 최빈수, 사분위수이다. 분산은 산포도 통계량에 속한다.

13 100명의 여자에 대한 신장과 체중을 비교한 자료이다. 체중의 개인차가 신장의 개인차보다 크다고 할 수 있는가?

	평균	표준편차
체중	52.3kg	2.54kg
신장	152.7 cm	2.28cm

① 체중에 대한 개인차가 크다.

② 신장에 대한 개인차가 크다.

③ 체중에 대한 개인차와 신장에 대한 개인차는 동일하다.

④ 체중과 신장의 개인차는 알 수 없다.

변동계수(CV: Coefficient of Variance)

▶ 평균을 중심으로 한 상대적인 산포의 척도를 나타내는 수치이다.

14 다음 중 시공간 데이터(Spatio-Temporal Data)로 알맞지 않은 것은?

① 수 ② 선

③ 면 ④ 점

시공간 데이터(Spatio-Temporal Data)

▶ 시공간 데이터는 공간적 정보에 시간의 흐름이 결합된 다차원 데이터를 의미한다.

▶ 시공간 데이터는 점(Point), 선(Line), 면(Polygon)과 같은 형태로 시각화할 수 있다.

15. 다음 중 시공간 데이터 시각화 도구가 아닌 것은?

① 코로플레스 지도

② 카토그램

③ 버블 플롯 맵

④ 스타차트

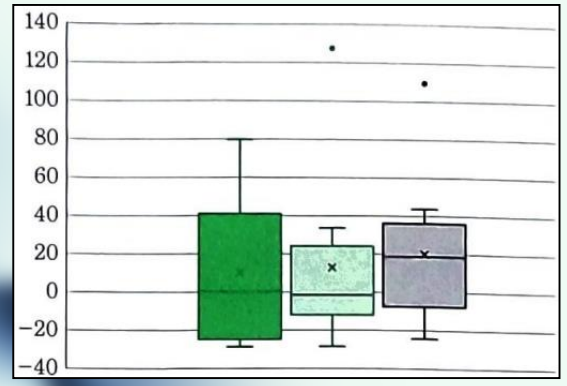
시공간 데이터 탐색 방법

1. 코로플레스 지도(Choropleth Map)

▶ 등치지역도라고도 하며, 어떤 데이터 값의 크기에 따라 해당 영역을 색칠해서 표현하는 방법이다.

2. 데이터 탐색 예상 문제

16 다음과 같은 시각화 도구에서 확인할 수 없는 값은?



① 중앙값 ② 이상값

③ IQR ④ 분산

주어진 시각화 도구는 박스플롯(상자수염그림)이다.

박스플롯에는 **하위경계(최솟값), 제1사분위수(Q1), 제2사분위수(Q2), 제3사분위수(Q3), IQR(Q3 - Q1), 상위경계(최댓값), 이상값이** 있다. 박스플롯에서는 분산을 포함하지 않는다.

17 다음 중 왜도 > 0일 때 올바른 순서는?

① 최빈수 < 중위수 < 평균

② 최빈수 < 평균 < 중위수

③ 평균 < 중위수 < 최빈수

18 다음 중 변량 데이터의 유형이 아닌 것은?

① 일변량 데이터

② 이변량 데이터

③ 다변량 데이터

④ 공변량 데이터

변량 데이터의 유형

▶ 변량(Variance)이란 조사 대상의 특징을 숫자나 문자로 나타낸 값을 의미한다.

▶ **변량은 종속변수(Y)의 수에 따라서 일변량, 이변량, 다변량으로 구분된다.**

19 다음 중 박스플롯에 대한 설명으로 틀린 것은?

① 최소값은 제1사분위에서 1.5×IQR을 뺀 위치를 의미 한다.

② 제1사분위(Q1)는 자료들의 하위 25%의 위치를 의미한다.

③ 제3사분위(Q3)는 자료들의 하위 75%의 위치를 의미한다.

④ 최댓값은 제3사분위에서 IQR의 0.5배 위치를 의미한다.

최댓값은 제3사분위에서 IQR의 1.5배 위치를 의미한다.

상자 수염 그림(Box Plot)

▶ 수치적 자료를 표현하는 그래프이다. 이 그래프는 가공

2. 데이터 탐색 예상 문제

20 다음 중 비정형 데이터의 특징으로 틀린 것은? .

① 비정형 정보는 일반적으로 텍스트 중심으로 되어 있다.

② 날짜, 숫자, 사실과 같은 데이터도 포함할 수 있다.

③ 구조화 되지 않은 데이터를 의미한다.

④ 변칙과 모호함이 발생하지 않는다.

비정형 데이터의 특징

▶ 비정형 정보는 일반적으로 텍스트 중심으로 되어 있으며 날짜, 숫자, 사실과 같은 데이터도 포함할 수 있다.

▶ 변칙과 모호함이 발생하므로 데이터베이스의 칸 형식의 폼에 저장되거나 문서에 주석화된(의미적으로 태그된) 데이터에 비해 전통적인 프로그램을 사용하여 이해하는 것을 불가능하게 만든다.

▶ 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어렵다.

감사합니다.