데이터 탐색과 시각화

시간 시각화

시간의 흐름에 따른 변화

비교 시각화

그룹별 차이를 나타냄

분포 시각화

전체 데이터에서 특정 항목이 차지하는 비중을 나타냄

관계 시각화

두 개 이상의 수치 데이터를 통해 서로 간의 관계를 나타냄

공간 시각화

실제 지리적 위치에 수치를 나타냄

공분산(Covariance)

2개의 확률변수의 선형 관계를 나타내는 값

$$COV(X_1, X_2) = \frac{\Sigma(각\,X_1의\,편차)(각\,X_2의\,편차)}{n(-1)} = \frac{1}{n(-1)}\Sigma(X_{1i} - \bar{x}_1)(X_{2i} - \bar{x}_2)$$

상관계수(Correlation coefficient)

2개의 연속성 변수 간의 연관성에 대한 척도

피어슨(Pearson) 상관계수

X_1과 X_2가 함께 변하는 정도(공분산)를 X_1과 X_2가 변하는 전체 정도로 나눔

$$P(X_1, X_2) = \frac{COV(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}}$$

시간 시각화

시점 요소가 있는 데이터는 시계열(Time series) 형태로 표현이 가능

효과)
전체적인 흐름을 한눈에 확인
데이터의 트렌드나 노이즈도 쉽게 발견 가능

1. 연속형(선그래프)

시간 간격의 밀도가 높을 때 사용

ex) 초 단위의 공정 센서 데이터, 일년 간의 일별 판매량 데이터
But, 데이터의 양이 너무 많거나 변동이 심하면 트렌드나 패턴을 확인하는 것이 어려움 -> 추세선 삽입(들쭉날쭉한 데이터 흐름을 안정된 선으로 표현) -> 전체 경향과 패턴을 쉽게 파악 가능

이동평균(Moving average) 방법
: 데이터의 연속적 그룹의 평균을 구하는 것(추세선을 그리는 가장 일반적인 방법)
ex) 2 -> 5 -> 3 -> 7 -> 4
(2, 5, 3의 평균) -> (5, 3, 7의 평균) -> (3, 7, 4의 평균)

2. 분절형(막대/누적/점 그래프)

시간의 밀도가 낮은 경우 활용.
ex) 1년 동안의 월 간격 단위 흐름

값들의 상대적 차이를 나타내는 것에 유리.
+ 막대에 색상을 표현하여 특정 시점에 대한 정보를 추가

비교 시각화

히트맵 차트(Heatmap chart)

그룹과 비교 요소가 많을 때 효과적으로 시각화를 할 수 있는 방법
각각의 셀은 색상이나 채도를 통해 데이터 값의 높고 낮음을 나타냄
각 행: 그룹 / 각 열: 요소

표현 방법
- 하나의 변수(그룹) X N개의 각 변수에 해당하는 값들(수치형)
- 하나의 변수(그룹) X 하나의 변수(그룹/수준) X 하나의 변수(수준)

주의)
현재 데이터의 구조와 확인 목적을 정확히 파악한 후 차트 생성
분류 그룹이나 변수가 너무 많으면 혼란을 유발할 수 있기 때문에 적정한 수준으로 데이터를 정제

방사형 차트(Radar chart)

하나의 차트에 하나의 그룹을 시각화
하나의 차트에 모든 그룹을 한 번에 시각화

평행 좌표 그래프(Parallel coordinates)

변수별 값을 정규화 -> 평행 좌표 그래프를 보다 효과적으로 표현

장점)
각 그룹의 요소별 차이 수준을 효과적으로 파악
집단적 경향성 표현에 용이

분포 시각화

구분)

- 양척 척도(연속형)
- 질적 척도(명목형)

양적 척도

- 막대그래프/선그래프
- 히스토그램(histogram)을 통한 분포 단순화

: 세세하게 나누어 분포를 살핀 다음, 시각적으로 봤을 때 정보의 손실이 커지기 전까지 조금씩 구간의 개수를 축소

Q: Why?

A: 구간이 너무 많으면 보기가 어렵고 너무 적으면 정보의 손실이 크기 때문에 시각화의 이점이 사라짐

질적 척도

<구성이 단순한 경우>

- 파이차트(시각적 표현만으로는 비율을 정확히 알기 힘들기 때문에 수치를 함께 표시)
- 도넛차트(비어 있는 가운데 공간에 전체 값이나 단일 비율 값 등의 추가적인 정보 삽입)

<구성 요소가 복잡한 경우>

- 트리맵 차트

: 하나의 큰 사각형을 구성 요소의 비율에 따라 작은 사각형으로 쪼개어 분포를 표현

장점)

사각형 내에 더 작은 사각형을 포함시켜 위계구조 표현 -> 한정된 공간 안에서 많은 구성 요소들의 분포를 체계적으로 표현

단점)

구성 요소들 간의 규모 차이가 크면 표현이 어려울 수 있다는 단점

- 와플 차트

와플처럼 일정한 네모난 조각들로 분포를 표현, but 트리맵 차트처럼 위계구조는 표현 불가

관계 시각화

산점도(scatter plot)

각 요소를 X축, Y축에 대입하여 일치하는 지점에 점을 찍음

장점)
점들의 분포와 추세를 통해 두 변수 간의 관계를 파악 가능

Tip)
극단치로 인해 주요 분포 구간이 압축되어 시각화의 효율이 떨어지기 때문에 극단치를 제거하고 그리기
다량의 데이터로 점들이 서로 겹칠 때, 각 점에 투명도(or 농도/색상)를 주어 점들의 밀도 또한 함께 표현

단점)
두 개의 변수 간 관계만 표현 가능

- ➔ 버블차트를 이용하여 세 가지 요소의 상관관계를 표현(X축, Y축 +버블의 크기)

주의) 버블차트를 해석할 때는 원의 지름이 아닌 면적을 통해 크기를 판단

공간 시각화

위치 정보인 위도와 경도 데이터를 지도에 매핑하여 시각적으로 표현
혹은, 시각화 프로그램에 따라 위도와 경도 정보가 없어도 가능
ex) Google의 지오맵(GeoMap)은 지명만으로도 공간 시각화가 가능

장점)
데이터를 훨씬 명확하고 직관적으로 볼 수 있음
지도를 확대하거나 위치를 옮기는 등 interactive한 활용이 가능

대표 기법
도트맵, 코로플레스맵, 버블맵, 컨넥션맵, etc.

도트맵(Dot map)
지리적 위치에 동일한 크기의 작은 점을 찍어서 해당 지역의 데이터 분포나 패턴을 표현
시각적으로 데이터의 개요 파악에 유리
정확한 값은 전달 불가

버블맵(Bubble map)
데이터 값이 원의 크기로 표현되기 때문에 코로플레스맵보다 비율을 비교하는 데에 효과적
!버블 크기 조절 필수!

코로플레스맵(Choropleth map) = 단계구분도
데이터 값의 크기에 따라 색상의 음영을 달리하여 해당 지역에 대한 값을 시각화
경우에 따라 여러 색상 혼합, 투명도/명도/채도 등 다양한 표현

커넥션맵(Connection map) = 링크맵(Link map)
지도에 찍힌 점들을 곡선 또는 직선으로 연결하기 지리적 관계를 표현
연속적 연결을 통해 지도에 경로 표현
연결선의 분포와 집중도 -> 지리적 관계의 패턴을 파악
ex) 지역 간 무역 관계, 항공 경로, 통신 정보 흐름, etc.

박스 플롯(Box-and-Whisker Plot)

네모 상자 모양에 최댓값과 최솟값을 나타내는 선이 결합된 모양의 데이터 시각화 방법

장점)
하나의 그림으로 양적 척도 데이터의 분포 및 편향성, 평균과 중앙값 등 다양한 수치를 보기 쉽게 정리
특히 두 변수의 값을 비교할 때 효과적
데이터의 대체적인 분포 형태를 쉽게 파악
카테고리별 분포 비교

<5가지 수치>

1. 최솟값: 제1사분위에서 1.5 IQR을 뺀 위치
2. 제1사분위(Q1): 25%의 위치
3. 제2사분위(Q2): 50%의 위치(중앙값(median)을 의미)
4. 제3사분위(Q3): 75%의 위치
5. 최댓값: 제3사분위에서 1.5 IQR을 더한 위치

+평균값(경우에 따라)

또한, 각 최솟값과 최댓값의 범위를 넘어가는 값은 이상치(outlier)로서 작은 원으로 표시

<분위수를 구하는 수식>

$$Q^1 = \frac{1}{4}(n-1)th\ value$$

$$Q^2 = \frac{2}{4}(n-1)th\ value$$

$$Q^3 = \frac{3}{4}(n-1)th\ value$$