2.2 데이터 탐색

개념 1

● 탐색적 자료 분석(EDA)

EDA란 데이터를 이해하고 의미 있는 관계를 찾아내기 위해 데이터의 통곗값과 분포 등을 시각화하고 분석하는 것을 말한다.

특징

1. 저항성

2. 잔차해석

3. 자료 재표현

4. 시각화

개념 2

● 데이터 파악

데이터 탐색을 들어가기에 앞서 분석하고자 하는 데이터가 어떤 데이터인지, 각 변수가 의미하는 바는 무엇인지를 파악하는 것이 선행돼야 한다.

개념 3

● 상관관계 분석

상관분석은 두 변수가 선형적 관계를 가지는 분석하는 통계적 분석 방법이다.

상관계수 해석

1. 상관계수의 절대값이 1에 가까울수록 강한 상관이 존재한다고 해석할 수 있다.

2. 상관계수의 부호는 관계의 방향을 의미한다.

개념 4

● 상관계수 종류

1. 피어슨 상관계수

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

특징: 모수 검정, 연속형 변수에 대한 것

2. 스피어만

$r_i = 1 - \frac{6\sum(x_i-y_i)^2}{n(n^2-1)}$

특징: 비모수 검정, 이산형 순서형 변수에 대한 것

기초통계량 추출 및 이해

개념 5

● 중심경향치

데이터의 특성을 파악하기 위해 중심경향치를 확인해보는 것이 우선이다.

1. 평균
2. 중앙값
3. 최빈값

개념 6

● 산포도 통계치

1. 범위 : Range=Max-Min
2. 사분위수 범위
3. 분산
4. 표준편차

개념 7

● 분포 모양의 이해

1. 왜도

왜도는 데이터 분포의 비대칭성을 나타는 지표이다.

왜도 <0 : 오른쪽으로 치우쳐져 있음을 의미한다.

왜도 ≈ 0 : 중앙 대칭적임을 의미

왜도 >0 : 왼쪽으로 치우쳐져 있음을 의미한다.

2. 첨도

첨도는 데이터가 중심으로 얼만큼 몰려있는지를 나타내는 지표이다.

첨도 <3 : 표준 정규분포보다 더 퍼져있다.

첨도 ≈ 3 : 표준 정규분포랑 비슷함

첨도 >3 : 표준 정규분포보다 더 좁게 분포되어 있다.

시각적 데이터 탐색

개념 8

● 히스토그램

히스토그램은 연속형 변수 데이터를 구간으로 나누고 해당 구간의 빈도 분포를 보여준다.

히스토그램을 그리는 방법은 아래와 같다.

1. 데이터의 범위를 구간으로 나눈다.

2. 각 구간에 속하는 데이터의 개수를 센다

3. 가로축에는 구간을 세로축에는 빈도수로 막대그래프를 이어 그린다.

개념 9

● 막대 그래프

범주형 데이터를 직사각형 막대그래프로 그린 그래프를 말한다.

히스토그램과 다른점 : 가로축이 수치형이 아니여도 된다.

개념 10

● 줄기-잎 그림

줄기 잎 그림은 큰 수 자릿값은 세로 선의 왼쪽 줄기에, 작은 수의 자릿값은 세로 선 오른쪽 앞에 나타내는 표 형태이다.

개념 11

● 상자그림(Box plot)

데이터에 대한 통계량인 최솟값, Q1, 중앙값, Q3, 최댓값을 가지고 그린 그래프이다.

사분위수 범위 (Q3-Q1)의 1.5배를 넘는 곳에 위치한 값을 이상치로 파악한다.

개념 12

● 산점도

산점도는 직교좌표계에서 점을 사용해 서로 다른 연속형 변수의 값을 나타낸 것을 의미한다.

개념 13

● 원그래프

원그래프는 전체에 대한 각 부분의 비율을 원 모양으로 나타낸 그래프이다.

개념 14

● 시간 데이터(Time series)

시간 데이터란 시간에 따라 발생하는 데이터를 의미한다.

개념 15

● 공간 데이터

공간 데이터는 지하, 지상, 수중, 수상 등에 존재하는 객체의 위치 및 공간 관계 정보와 관련한 데이터를 의미한다.

개념 16

● 지리정보시스템

지리정보시스템은 지리 공간적으로 참조 가능한 모든 형태의 정보를 효율적으로 수집,저장할수 있게 설계된 컴퓨터 하드웨어와 소프트웨어 등 통합체다.

지리정보시스템은 구성요소는 다음과 같다.

1. 컴퓨터 시스템

2. GIS 소프트웨어

3. 인력

4. 데이터

5. 인프라

개념 17

● 다변량 데이터

데이터의 차원이 3차원 이상일때 다변량 데이터라고 부른다.

개념 18

● 다변량 데이터 분석 방법

1. 상관분석

산점도 행렬을 그려 여러 변수를 조합한 산점도와 상관계수를 한 화면에서 확인한다.

2. 다차원 척도법

다차원 척도법은 객체 사이의 유사성을 유지하며, 2차원 또는 3차원 공간으로 시각화하는 방법이다.

3. 주성분분석

주성분분석은 데이터의 분포를 잘 설명함과 동시에 정보의 손실은 최소화하도록 고차원의 데이터를 저차원의 데이터로 변환하는 차원 축소기법이다.

4. 선형판별분석

어떤 그룹에 속할지를 판별하는 판별분석기법인 선형판별분석은 다변량 데이터에 판별 함수를 적용해 데이터의 클래스 분리를 최적으로 수행할 수 있게 데이터를 축소한다.

개념 19

● 텍스트 마이닝

텍스트 마이닝이란 다양한 문서 자료 내 비정형 텍스트 데이터에 자연어 처리 기술 및 문서처리 기술을 활용해 인사이트를 도출하는 기술이다.

코퍼스 : 분석 작업의 대상이 되는 대량의 텍스트 문서들을 모아놓은 집합

토큰화 : 구조화되어 있지 않은 문서를 단어로 나누는 과정

불용어 : 코퍼스에서 자주 등장하지만, 분석 프로세스에 있어 기여하는 바가 없는 단어

어간 추출 : 단어 내 접사를 제거하고 단어에서 의미를 담고 있는 어간을 분리 하는 것

개념 20

● 토픽 모델링과 LDA

토픽 모델링 : 대량의 문서 집합에 존재하는 추상적인 토픽을 추출하는 통계적 모델링 방법

Bag of Word : 단어의 순서는 무시하고 빈도만 고려하는 것을 말한다.

개념 21

● 소셜 네트워크 분석의 이해

소셜 네트워크 분석은 사회관계망 분석이라고도 부른다. 오늘날에는 인터넷, 도로, 유전정보, 화합물 등과 같이 우리 사회도 네트워크 형태로 구조화 되어 있다. 소셜 네트워크 분석을 통해 구성원들 간의 상호 의존성을 이해하고 사회 전체의 관계망 분석을 통해 사회현상을 설명하는 인사이트를 얻을 수 있다.

개념 22

● 네트워크 구조를 파악하기 위한 요소

1. 중심성

전체 네트워크에서 한 개체가 중심에 위치하는 정도를 표현하는 지표

2. 밀도

네트워크 내에 존재하는 노드 간의 연결정도의 수준을 의미

3. 집중도

네트워크 전체가 한 중심에 집중되는 정도를 의미

4. 연결정도

노드에 연결된 관계의 수를 의미

5. 포괄성

네트워크 내 연결되지 않은 노드들의 수를 뺀 연결된 노드들의 비율

* 개념정리

산점도 ⇒ 두 변수 사이의 상관관계를 알 수 있다.

Box plot ⇒ 평균, 분산 알수없다.

시공간 Data 에서 연속적 변화는 데이터의 수집 주기가 일정한 Data 로 표현.
이산적 변화는 시간의 변화에 따라 데이터가 추가 된다.

코로플레스 ⇒ 등치지역도 라고 함
Data 수치에 따라 지정한 색상 스케일로 영역을 색칠
가장 보편적인 방법.

버블 플롯맵 ⇒ 위도와 경도를 사용하여 좌표를 원으로 정의

카토그램 ⇒ Data 에 따라 지도의 면적이 왜곡.