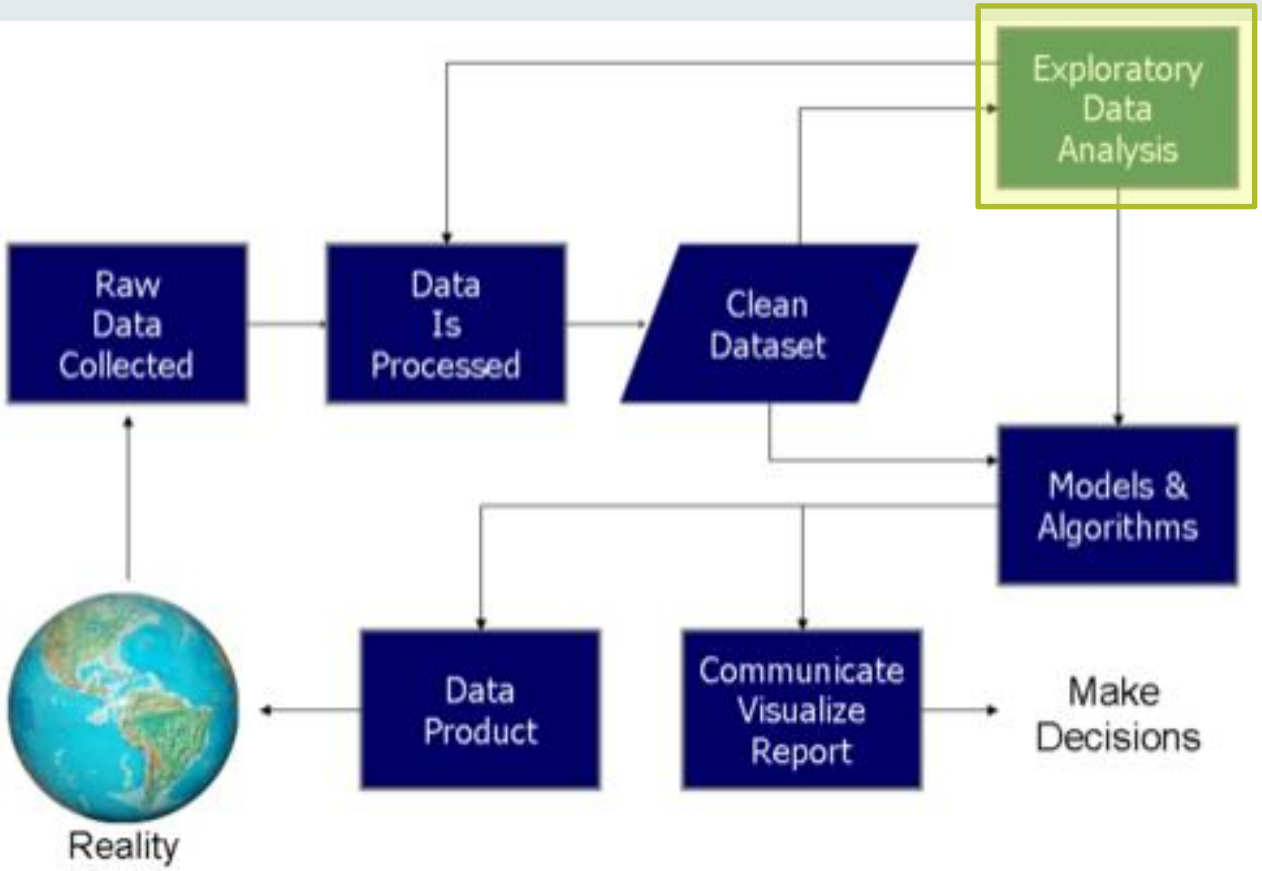
탐색적 데이터 분석

탐색적 데이터 분석 EDA

- EDA: Exploratory Data Analysis
- 존 튜키 미국의 통계학자
- 기존 통계학이 정보 추출 과정에 가설 검정 등에 치우쳐 자료가 가지고 있는 **본연의 의미를** 찾는데 어려움이 있어 이를 보완하고자 주어진 데이터 자체로 충분한 정보를 찾을 수 있도록 여러 탐색적 자료 분석 방법을 개발함

데이터 사이언스의 프로세스



필요성

- 데이터를 검토함으로써 **데이터가 의미하는** 현상을 잘 이해할 필요가 있음
- 본격적인 분석에 들어가기 앞서 **데이터 재수집이나 추가 수집** 등의 결정을 내릴 수 있음
- 데이터를 다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 발생하지 못했을 다양한 패턴을 발견하고, 이를 바탕으로 **기존의 가설 수정 혹은 새로운 가설**을 세울 수 있음
- 데이터에 대한 지식은 이후 **통계적** 추론이나 **예측 모델을** 만들 때 유용하게 사용됨

분석 과정

분석 계획**에는 어떤 속성 및 속성 간의 관계를** 집중적으로 관찰해야 할지, 이를 위한 **최적의 방법**은 무엇인지가 포함되어야 한다

1. 분석 목적과 변수를 확인하고 개별 변수 이름이나 설명을 확인
2. 데이터를 전체적으로 살펴본다
    - Head 혹은 tail 확인
    - 이상치 결측치 등확인
    - 데이터에 문제가 있는지
3. 데이터의 개별 속성 관찰
    - 대칭/ 비대칭 분포
    - 실제 값의 주요 분포 범위
    - 값의 표준편차
    - 각 속성 값이 예측 범위와 분포를 가지는지
    - 그렇지 않다면 이유가 무엇인지
4. 속성 간의 관계에 초점을 맞추고 개별 속성 관찰에서 찾지 못했던 패턴 발견 (상관관계, 시각화 등)

분석 과정

분석 계획**에는 어떤 속성 및 속성 간의 관계를** 집중적으로 관찰해야 할지, 이를 위한 **최적의 방법**은 무엇인지가 포함되어야 한다

1. 분석 목적과 변수를 확인하고 개별 변수 이름이나 설명을 확인
2. 데이터를 전체적으로 살펴본다
    - Head 혹은 tail 확인
    - 이상치 결측치 등확인
    - 데이터에 문제가 있는지
3. 데이터의 개별 속성 관찰
    - 대칭/ 비대칭 분포
    - 실제 값의 주요 분포 범위
    - 값의 표준편차
    - 각 속성 값이 예측 범위와 분포를 가지는지
    - 그렇지 않다면 이유가 무엇인지

시각화 필요

4. 속성 간의 관계에 초점을 맞추고 개별 속성 관찰에서 찾지 못했던 패턴 발견 (상관관계, 시각화 등)

속성 파악

- Categorical Variable
  - Nominal Data : 숫자로 표시가능하나 편의상 숫자화 (강아지-0 고양이-1)
  - Ordinal Data: 순위의 개념이 있는 클래스 (학점 A>B>C>D)
- Numeric Variable
  - Continuous Data : 연속 데이터 (키, 몸무게, 속도 등)
  - Discrete Data: 비연속 데이터 (가족 구성원)

독립변수와 종속변수

구분	입력 데이터	출력 데이터
개념	• 분석의 기반이 되는 데이터	• 추정하거나 예측하고자 하는 목적 데이터
표기	• 보통 알파벳 X 로 표기합니다. • 보통 x, x_1, x_2, x_n등으로 표시합니다. • y의 변화를 회귀방정식으로 표현하고 설명하기 위해 필요한 변수	• 보통 알파벳 Y 로 표기합니다.
유사용어	• **독립변수(independent variable)** • 특징(feature) • 설명변수(explanatory variable) • 예측변수	• **종속변수(dependent variable)** • 반응변수 • 목표변수 • 목적 값(Target Value) • 종속변수가 카테고리값이면 라벨(label) 또는 클래스(class)라고도 합니다.
영향	• 영향을 주는 변수	• 영향을 받는 변수

독립변수와 종속변수

- `age` : 나이
- `workclass` : 고용 형태
- `fnlwgt` : 사람 대표성을 나타내는 가중치 (final weight의 약자)
- `education` : 교육 수준
- `education_num` : 교육 수준 수치
- `marital_status` : 결혼 상태
- `occupation` : 업종
- `relationship` : 가족 관계
- `race` : 인종
- `sex` : 성별
- `capital_gain` : 양도 소득
- `capital_loss` : 양도 손실
- `hours_per_week` : 주당 근무 시간
- `native_country` : 국적
- `income` : **수익 (예측해야 하는 값)**

- 다양한 feature들을 이용해서 이 사람의 income을 예측하고 싶다

-> **종속변수**는 income 이 되며 **독립변수**는 나머지 feature 들이 된다.

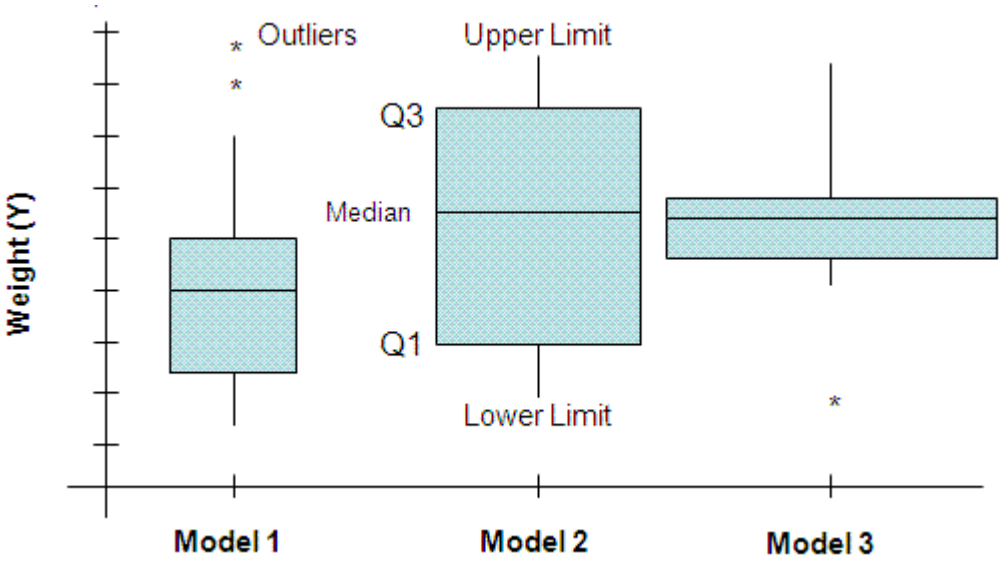
- Y값인 income 이 아니더라도 독립변수간 종속-독립이 될 수 있음

-> 교육수준과 교육수준 수치는 서로 영향이 있을 수 있다.

5가지 숫자 요약 (Five-number summary)

데이터 집합에 대한 정보를 제공하는 통계량으로 가장 중요한 표본 백분위수 5가지로 구성

1. **최대값**
2. **상위 사분위수**
   - (전체 데이터의 상위 25%)
3. **중앙값**
4. **하위 사분위수**
   - (전체 데이터의 하위 25%)
5. **최소값**



Descriptive Statistics

속성 관계 분석

데이터 조합	요약 통계	시각화
Categorical-Categorical	교차 테이블	모자이크 플롯
Numeric-Categorical	카테고리별 통계 값	박스플롯
Numeric-Numeric	상관계수	산점도

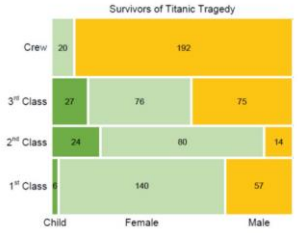


그림4 모자이크 플롯 예시