통계분석B 1강

2022.02.15 (화) 14:30 ~ 16:30

chwhint@gmail.com

목차

EDA (Exploratory Data Analysis, 탐색적 데이터 분석)

예제로 시작하기

대한 민국의 국민의 키 분포 통계

- (예상) 상식적으로 대부분이 대략 0.5 ~ 2m 사이로 일 것이므로, 대략 1.5 ~ 1.8m 사이로 추정 가능
- (확인) 통계청 성인 신장 자료
  - (분석) 19세 이하 평균 = 1.6949m → 통계가 잘못되었나?
- (의문 확인) 연령대별 인구 분포
  - 19세 이하 인구 숫자 , 10세 이상이 더 많음
  - 평균 계산시 다음과 같은 수식을 사용하게 됨
    - 19세 이하 평균 = (작은키 * 작은 인구수 + 큰키 * 많은 인구수)/전체 인구수
    - 평균이 큰 키에 가깝게 분포하게 될 것.
- (가정) 만약 수치값 범위가 1694.9 이라고 표현 되었다면?
  - (가설1) mm 단위로 표현되었다. —> 400 ~ 1800 사이에 값들이 위치
  - (가설2) 수집된 데이터가 이상하다 → 데이터가 널뛰기 할 것
- 기타 데이터 확인 프로세스(빈 데이터, out-lier 등 확인)
- (QA 완료 시) 다음 프로세스 진행
  - 결측치 값 처리
  - 없는 데이터에 대해 추정 수치로 사용 등

EDA는 무엇인가

- 데이터의 관상 보기
- 최종적인 데이터의 분석, 시각화, 커뮤니케이션 에는 상당히 많은 Trial & Error

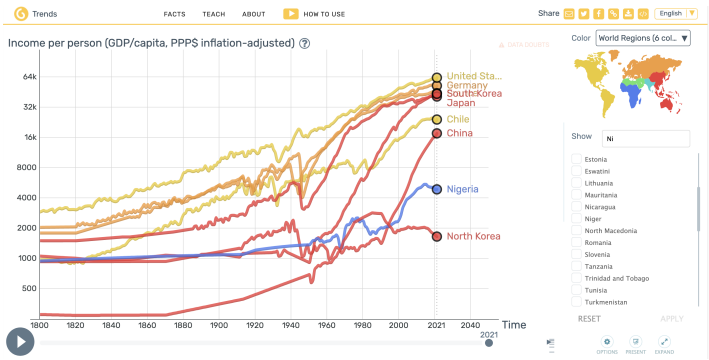
    - 결과의 Grand opening 이 아닌 짧은 결과를 여러번 도출하여 엮기
- 최소한의 노력으로 일단 데이터의 분포와 경향성을 파악
- 기초적인 데이터 QA(Quality Assurance) 실시하고, 어떤 전략으로 접근할지 결정
- 데이터를 수집한 직 후 뿐만 아니라 데이터 가공, 변형, 저장 후 등 필요할 때 마다 사용

데이터의 종류

데이터 속성에 따른 분류

- **연속형(Continuous data)**
    - ex) 센서가 측정하는 물리량, 국가별 GDP 데이터
- **범주형 데이터(Categorical data)**
    - 범주형 데이터 (Multi-class)
        - ex) 국가를
    - 순서형 데이터(Ordinal)
        - ex) 영화 평점 등
        - cf) Python 에서는 `Sci-kit learn` 라이브러리의 `sklearn.preprocessing.OrdinalEncoder` 로 순서형 데이터 셋을 지원한다.
    - 이산형 데이터 (Binary)
        - ex) 참/거짓으로 나타낼 수 있는 모든 데이터

- **연도별 국가별 인당 GDP**
    - 연속형 데이터
        - 연도별 GDP
    - 범주형 데이터
        - 대륙별 색깔 구분



https://www.gapminder.org/tools/#$model$markers$line$data$filter$dimensions$country$country$/$in@=usa&=chn&=nga&=kor&=prk&=jpn&=ch type=linechart&url=v1

데이터 형식에 따른 분류

- Table 데이터 (=Rectangular data)
    - `pandas.DataFrame` 과 같은 형태의 데이터
    - 사건을 나타내는 행(row)
    - 변수를 나타내는 열(column)

- Table 데이터가 아닌 데이터
    - 표로 표현하기 어려운 다소 추상적인 데이터

- 관계
- 그래프(수학에서의 graph data)
- 객체와 필드 값을 표현

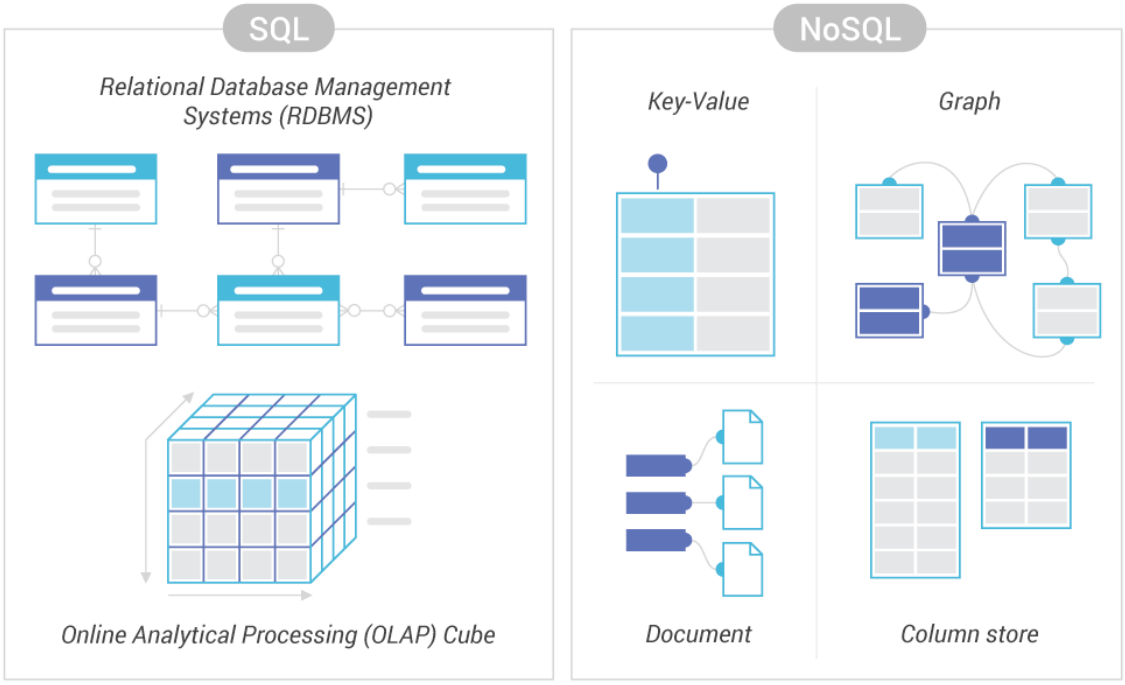
ex) json 데이터 (혹은 이의 변형) 형태 (:=python dictionary)

```
{
  "items":
    {
      "item":
        [
          {
            "id": "0001",
            "type": "donut",
            "name": "Cake",
            "ppu": 0.55,
            "batters":
              {
                "batter":
                  [
                    { "id": "1001", "type": "Regular" },
                    { "id": "1002", "type": "Chocolate" },
                    { "id": "1003", "type": "Blueberry" },
                    { "id": "1004", "type": "Devil's Food" }
                  ]
              },
            "topping":
              [
                { "id": "5001", "type": "None" },
                { "id": "5002", "type": "Glazed" },
                { "id": "5005", "type": "Sugar" },
                { "id": "5007", "type": "Powdered Sugar" },
                { "id": "5006", "type": "Chocolate with Sprinkles" },
                { "id": "5003", "type": "Chocolate" },
                { "id": "5004", "type": "Maple" }
              ]
          },

          ...

        ]
    }
}
```

https://opensource.adobe.com/Spry/samples/data_region/JSONDataSetSample.html#Example9



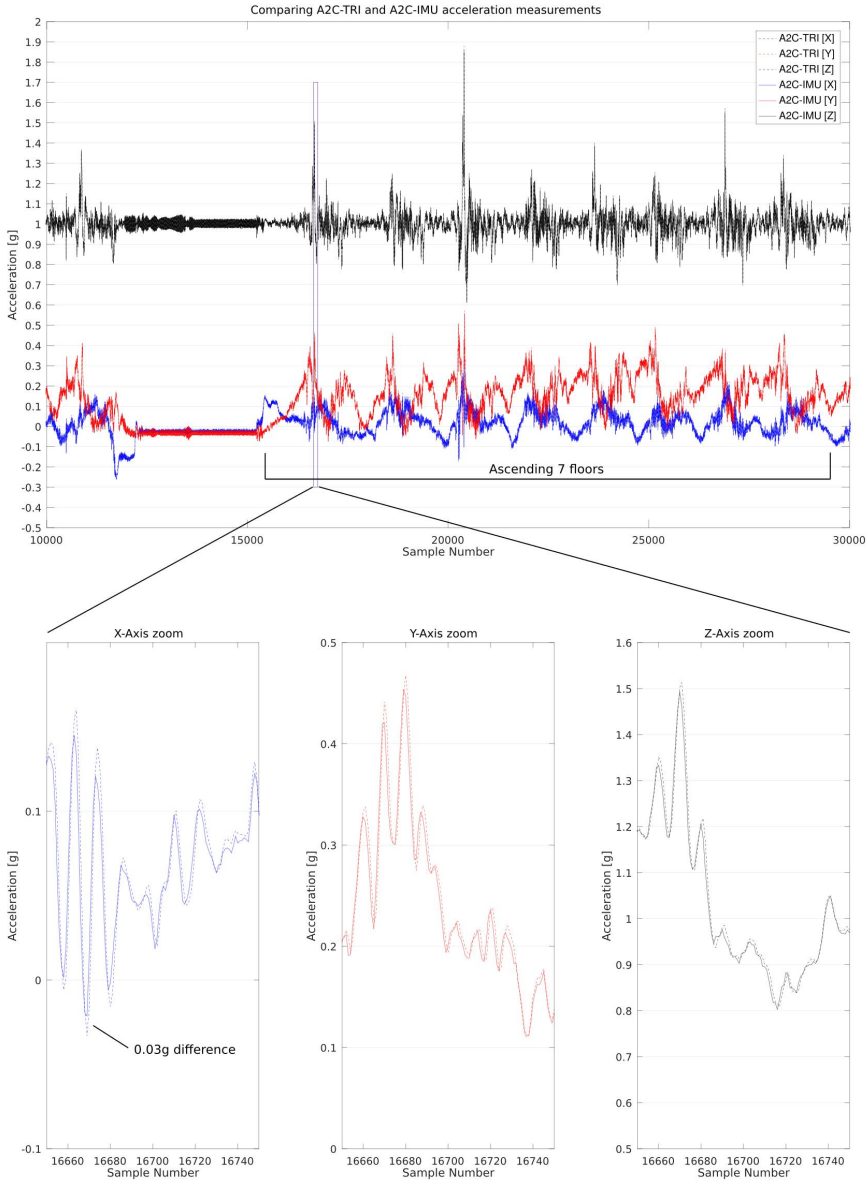
https://www.scylladb.com/resources/nosql-vs-sql/

EDA의 중요 포인트

1. **데이터에 대한 정확한 이해**
    - **Domain - Specific Knowlege 필요**
    - 각 row, column 의 의미는 무엇인가?
    - 데이터 값의 Feasibility를 정량/정성적으로 판단할 수 있는 기술

      (대한 민국 사람의 신장 통계 예시) 신장의 범위는 적절한가?

      ▼ ex) 가속도 측정
        - (선험 지식) 중력가속도 g = 10 m/s^2
        - (선험 지식) 자동차가 최대 3~5g 정도, 전투기가 9g 정도



https://lilliesystems.com/products/inertial-measurement-unit-imu-with-can-bus-a2c-imu-c/

- What if, 데이터에 이상이 있다??
  - 측정 주기의 문제?
    - 장비 설계/ 스펙
  - 측정 단위의 문제?
    - 센서 스펙과 통신 프로토콜
  - 센서의 문제?
    - 다른 센서와 비교
  - 통신의 문제?
    - (유선통신) 접촉불량, (무선통신) 전파 간섭 문제
    - 그외 기기 고장이 있는지 확인
  - HW(보드 고장, 전원 등)
    - 전원 확인 및 전력 측정 장비 활용

        - 데이터 저장의 문제?
            - raw 데이터의 패턴 확인
            - 단계별 저장 데이터 비교 분석

2. **결측치에 대한 합리적인 처리**
    - **Domain - Specific Knowlege 필요**
    - 다른 관측치를 활용
        - `pandas.DataFrame.interpolate()` **활용 (외삽/내삽 모두 가능)**
        - 내삽(interpolation)
        - 외삽(extrapolation)
        - 오류의 가능성이 있음

          (인당 GDP 통계 예시) 1990년대 초반까지의 아시아 데이터, 2021년 이후의 데이터?
    - 한가지 값으로 처리
        - ex) 시계열 데이터 padding
    - 결측치 버리기
        - 결측치를 버려도 상관없는 경우에만 사용

3. **시각화**
    - 목적
        - **데이터에 대한 통찰**
        - **커뮤니케이션**
    - 대상
        - 데이터 처리자를 위해서 - 나도 오해/착각 할수 있다.
            - 기본적인 것은 꼭 표시하자
            - legend, 축, 단위 등
        - 청자/독자를 위해서 - 사소한 것 까지 알려줘야 한다. 저들은 모른다.
            - 누구나 이해하기 쉽도록 시각화 하기
            - 초등 학생 정도가 들어도 이해할 수 있도록
            - 시간이 꽤 많이 소요됨

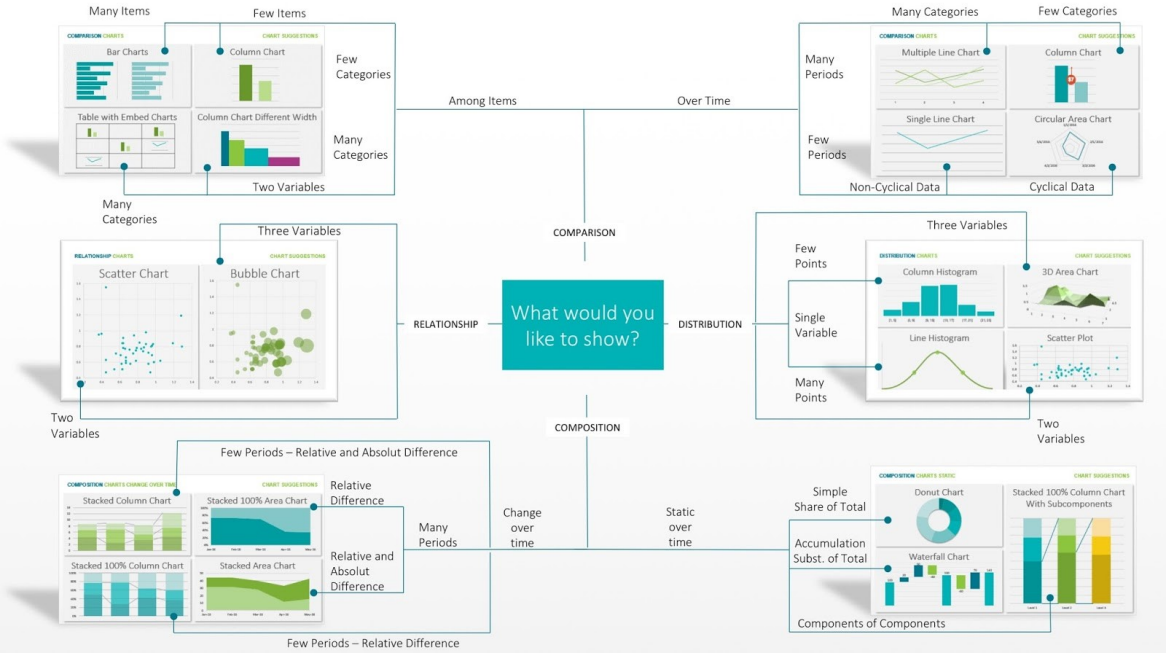
데이터에 대한 정확한 이해

데이터 확인

- 대표 위치 값 (산술, 기하, 조화, 가중 평균, 중앙, 최빈값 or its variation)
- 분산 (표준편차 or its variation)
- outlier 숫자와 분포 확인

시각화 방법

- AndrewAbela 의 Choosing the right chart page



https://www.techprevue.com/decision-tree-perfect-visualisation-data/

- 시각화에 대한 카테고리
  - 보여주고 싶은 것
    - 분포
    - 비교
    - 관계
    - 구성
  - 변인의 개수
    - 변인이 하나
    - 변인이 둘
    - 변인이 셋 이상
  - 변인의 특성
    - 순환적
    - 계층적
  - 데이터의 숫자
    - 많음
    - 적음
- 위 도표가 모든 방법을 나타내지는 않음
  - 다양한 루트를 통해 적절한 방법을 탐구/개발할 필요

  - 주로 많이 쓰이는 시각화 라이브러리
    - matplotlib
      - Matplotlib Tutorial - 파이썬으로 데이터 시각화하기
      - seaborn - 통계적, 확률적 데이터
    - Plotly

- 시각화 방법 references
  - Datavizcatalogue
  - The Python Graph Gallery
  - 변성윤님 블로그
  - Google it!

관상과 데이터

관상은 과학이다!

- 얼굴의 종횡비 가로/세로 수치가 높은 (납작한) 관상의 경우 사회적으로 지배자적 위치로 가고자 하는 성취지향성이 대단히 크다고 합니다.
- 경영학계 최신 연구에 의하면 이 비율이 넓은 CEO들이 M&A를 많이 하고, ROA가 높으며, 차입경영을 많이 합니다. 또한 해당 CEO가 소속된 기업의 주가변동성이 높다고 합니다.
- 그리고, 이 비율 넓은 헤지펀드 매니저들은 너무 과욕을 부려서 펀드 성과 (alpha)가 낮다고 하는 Yan Lu와 Melvyn Teo의 연구논문이 최근 Finance 탑저널인 JFQA에 게재됐습니다.

  http://www.fwhrmeasuring.com/