EDA

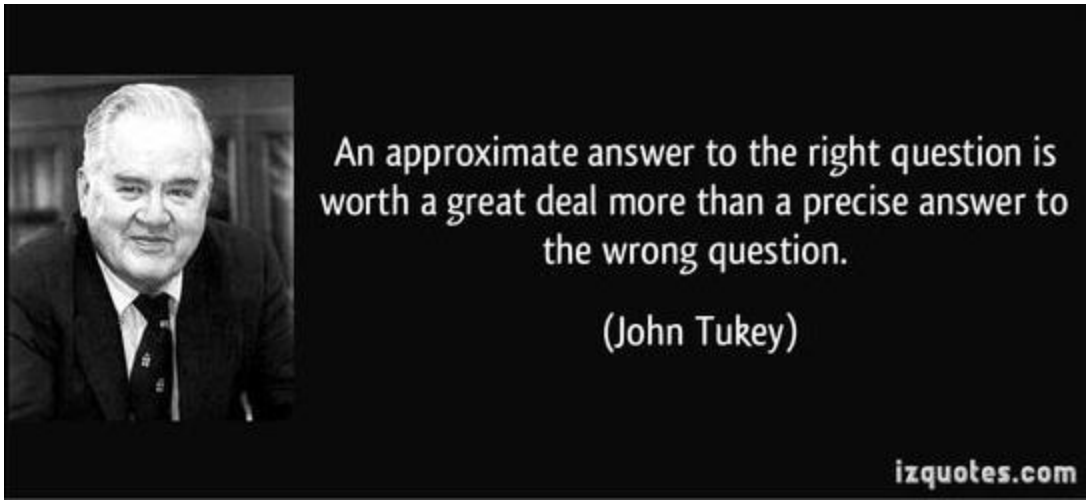
데이터 탐색 및 이해

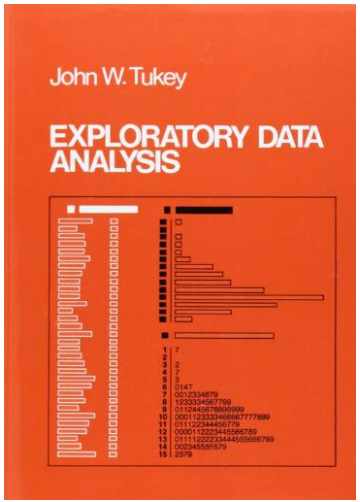
최희윤 강사

EDA 이해

EDA의 등장

연구소의 수학자이자 통계학자인 '존 튜키(John W. Tukey)'가
1977년 처음 **제안한 데이터분석 과정**

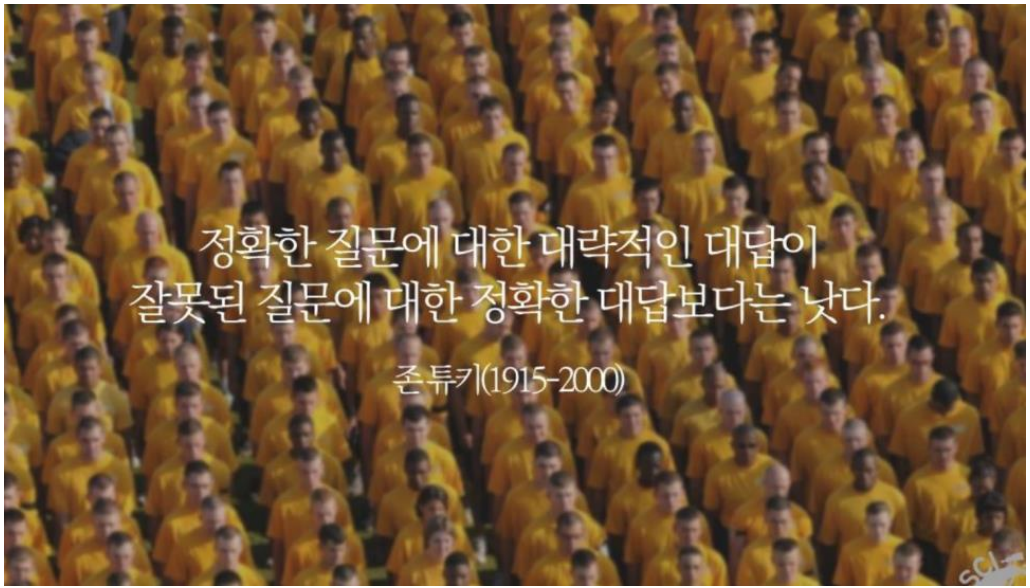




EDA의 개념

EDA는 데이터를 분석하고 결과를 내는 과정에 있어서

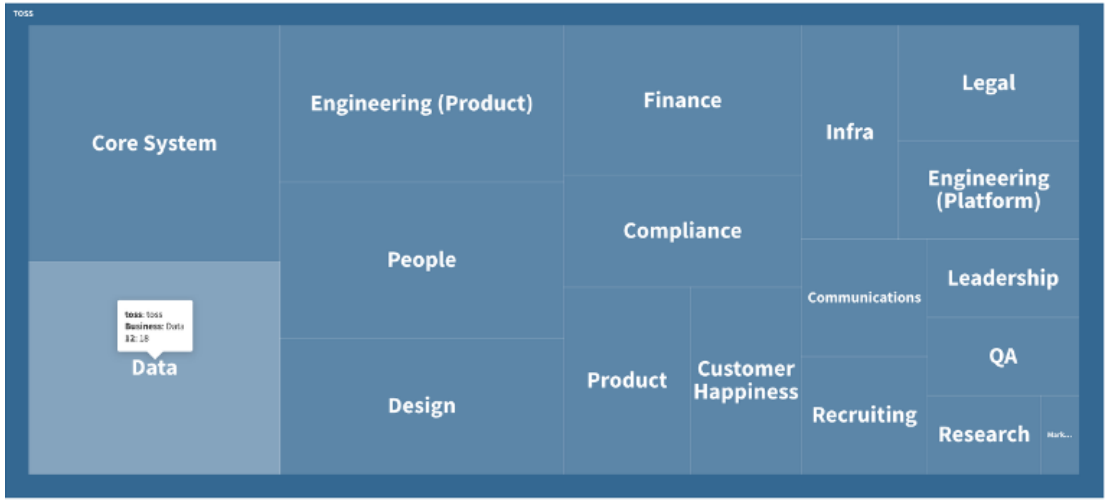
지속적으로 해당 **데이터에 대한 '탐색과 이해'**를 기본으로 가져야 함



이미지 출처
https://blog.naver.com/nabilera1/222993063024

EDA의 중요성

- ✓ 기술의 발전으로 **데이터가 많아**지고 **활용**할 수 있는 기회 증가
- ✓ 기업들은 데이터로 **의사결정**을 하고 고객들에게 **편리함**을 제공
- ✓ **데이터를 잘 다룰 줄 아는 사람**들이 필요
- ✓ 기업에서는 **데이터 관련 직무**들이 생겨나고 채용이 늘어남



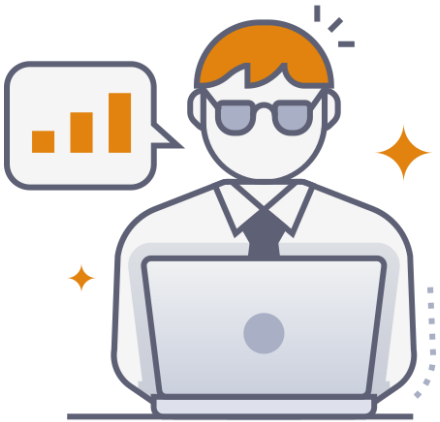
이미지 출처
https://brunch.co.kr/@hansungdev/7

EDA를 수행하는 사람들

- ✓ 데이터 과학자(Data Scientist)
- ✓ **데이터 분석가(Data Analyst)**
- ✓ 비즈니스 애널리스트(Business Analyst)
- ✓ 연구자(Researcher)

이미지 출처
https://boardmix.com/kr/news/data-science/

내가 데이터 분석가 라면?



John (32세), X이버 데이터분석가

John은 전자상거래 회사에서 데이터 분석가로 일하고 있다. 최근 회사는 고객 이탈율이 증가하고 있어, 이탈 고객의 특성을 분석하여 원인을 파악하고자 한다. John은 **EDA를 통해 다음과 같은 작업을 수행**한다.

(1단계) 데이터 수집: 최근 1년간의 고객 구매 데이터, 방문 빈도, 고객 서비스 문의 기록 등을 수집

(2단계) 데이터 구조 파악: 각 변수의 분포와 기본 통계량을 Pandas와 Seaborn을 사용하여 확인

(3단계) 이상치 탐지: 고객 서비스 문의 횟수나 구매 횟수가 비정상적으로 높은 데이터를 시각화하여 확인

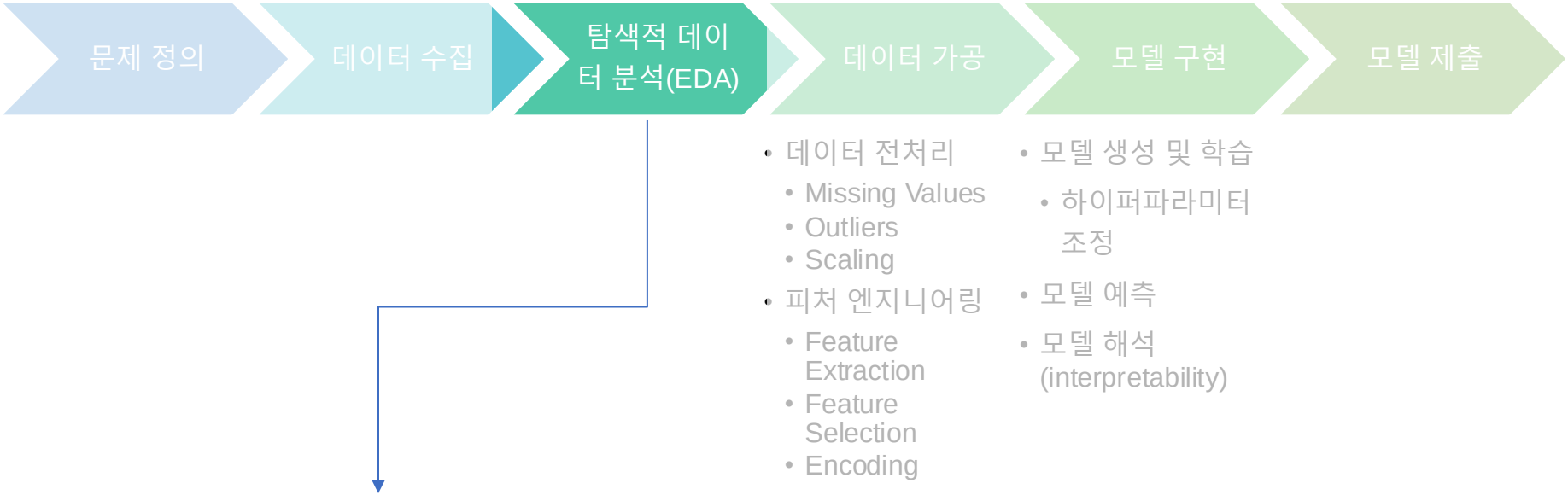
(4단계) 변수 간의 관계 분석: 고객 이탈 여부와 방문 빈도, 구매 횟수, 서비스 문의 횟수 간의 관계를 분석

(5단계) 결과 해석: 이탈 고객의 주요 특성을 파악하고, 이를 기반으로 고객 유지 전략을 제안

이미지 출처
https://www.flaticon.com/kr/free-icon/data-analyst_11460999

EDA의 중요성

데이터 분석 절차



탐색적 데이터 분석 (EDA, Exploratory Data Analysis)

"데이터를 다양한 각도에서 관찰하며 데이터를 이해하는 과정"

EDA 없이 하는 데이터 분석의 문제점

- ✓ **효율성 문제**
  데이터의 분포나 변수를 이해하지 못해 비효율적인 분석 절차를 따를 수 있음
- ✓ **재작업 증가**
  잘못된 가정에 기반한 결론으로 인해 분석 과정을 다시 시작해야 할 수 있음

이미지 출처
https://2runzzal.com/main/make/UmMzSkZSdVZpK0R0czJpaFYzV3BCUT09

EDA 없이 하는 데이터 분석의 문제점

- ✓ **과정 추적 어려움**
  데이터 탐색 과정에서 결정을 기록하지 않아 재현이나 오류 수정이 어려움
- ✓ **방향성 상실**
  분석의 진행 중간에 목적과 방향성을 잃기 쉬움

이미지 출처
https://www.seoul.co.kr/news/society/science-news/honey/2020/09/29/20200929500036

EDA 없이 하는 데이터 분석의 문제점

- ✓ **문제 정의 오류**
  문제를 정의하고 분석의 방향을 설정하는 데 중요한 역할을 함
- ✓ **데이터 이해 부족**
  데이터를 충분히 이해하지 못해 중요한 패턴이나 이상치를 놓칠 가능성이 높아짐
- ✓ **초기 가설의 한계**
  초기 가설이 편향될 수 있으며, 객관적인 검증이 필요

이미지 출처
https://m.blog.naver.com/skythe99/222422724690

EDA 없이 하는 데이터 분석의 문제점

- ✓ **설득력 부족**

  분석 결과를 설명하고 설득하는 데 필요한 논리적 근거가 부족해짐

- ✓ **신뢰성 결여**

  데이터를 충분히 이해하고 검증하지 않으면 분석 결과의 신뢰성이 떨어짐

이미지 출처

https://velog.io/@changhtun1/EDA%EB%A5%BC-%ED%86%B5%ED%95%9C-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EB%B6%84%EC%84%9D-%EA%B7%B8%EB%A6%AC%EA%B3%A0-%EC%9D%B8%EC%82%AC%EC%9D%B4%ED%8A%B8

EDA를 잘 수행하지 못한 경우

예측 모델 성능 저하 → 결과 왜곡 → 결과 신뢰성 하락

→ 인사이트 도출 불가 → 팀 내 커뮤니케이션의 어려움 → 비즈니스 의사결정 지연

위와 같은 문제 발생 위험이 있는지 점검해보세요.

이미지 출처
https://brunch.co.kr/brunchbook/workworkwork

EDA를 잘 수행한 경우

- ✓ 데이터 품질 보장
- ✓ 더 나은 분석 결과
- ✓ 인사이트 도출

이미지 출처
https://instablank.com/jjalbangMake/5664

EDA 실습

실습 범위

- 데이터 확인
  - 크기
  - 결측값
  - 이상치

- 데이터 전처리
  - Missing Values
  - Outliers
  - Scaling
- 피처 엔지니어링
  - Feature Extraction
  - Feature Selection
  - Encoding

- 모델 생성 및 학습
  - 하이퍼파라미터 조정
- 모델 예측
- 모델 해석 (interpretability)

EDA 체크리스트

✓ 데이터의 사이즈는 어느 정도 인지 ?

✓ 학습과 테스트 데이터는 어떻게 분리가 되어 있는지?

✓ 결측 값은 어느 정도 인지?

✓ 라벨이 있는 데이터라면 분포는 어떻게 되어있는지?

✓ 데이터의 특이점(이상치)이 있는지?

EDA 실습

Titanic.csv (from Kaggle)

Titanic 데이터 파이썬으로 분석하기 – 1(EDA).ipynb

사이트 URL
https://www.kaggle.com/datasets/mazenalattar/titanic-csv

EDA 정리

EDA가 어려운 이유

✓ **인지 편향(Cognitive Bias)**

사람이 **자신의 경험과 지식을 바탕으로 판단**을 내리는 경향.

데이터 분석에 있어서 **객관성을 유지하는 데 방해**가 될 수 있음.

“잘 듣고 잘 읽는 것이 중요하다.”

명확한 데이터 이해와

이해관계자의 요구를 명확히 파악하는 것이 중요

EDA에 사람이 필요한 이유

- ✓ 데이터의 **배경과 맥락 이해**
- ✓ 비판적 사고를 통해 **데이터를 평가하고 수정**할 수 있는 능력
- ✓ 분석 과정에서의 **문제/질문에 대한 창의적 해답과 가설** 제시
- ✓ 분석 결과를 **이해관계자에게 전달하고 피드백을 반영**

EDA 요약

'데이터 과학자, 데이터 분석가, 비즈니스 애널리스트, 연구자'가

데이터 분석 프로젝트의 초기 단계와 전처리 단계, 모델링 전에 EDA 수행

- **데이터의 구조와 특성을 이해**하고 **이상치와 결측치 식별**
- **변수 간의 관계를 탐색**

EDA 결과를 **데이터 전처리 및 클렌징, 모델링,**
비즈니스 인사이트 도출에 활용