DSL 21/01/26 세션
EDA 개요

4기 김동욱

Part1_Bayesian statistics

EDA란?

Exploratory **D**ata **A**nalysis

= 시각화와 각종 집계된 수치적 통계량들을 바탕으로

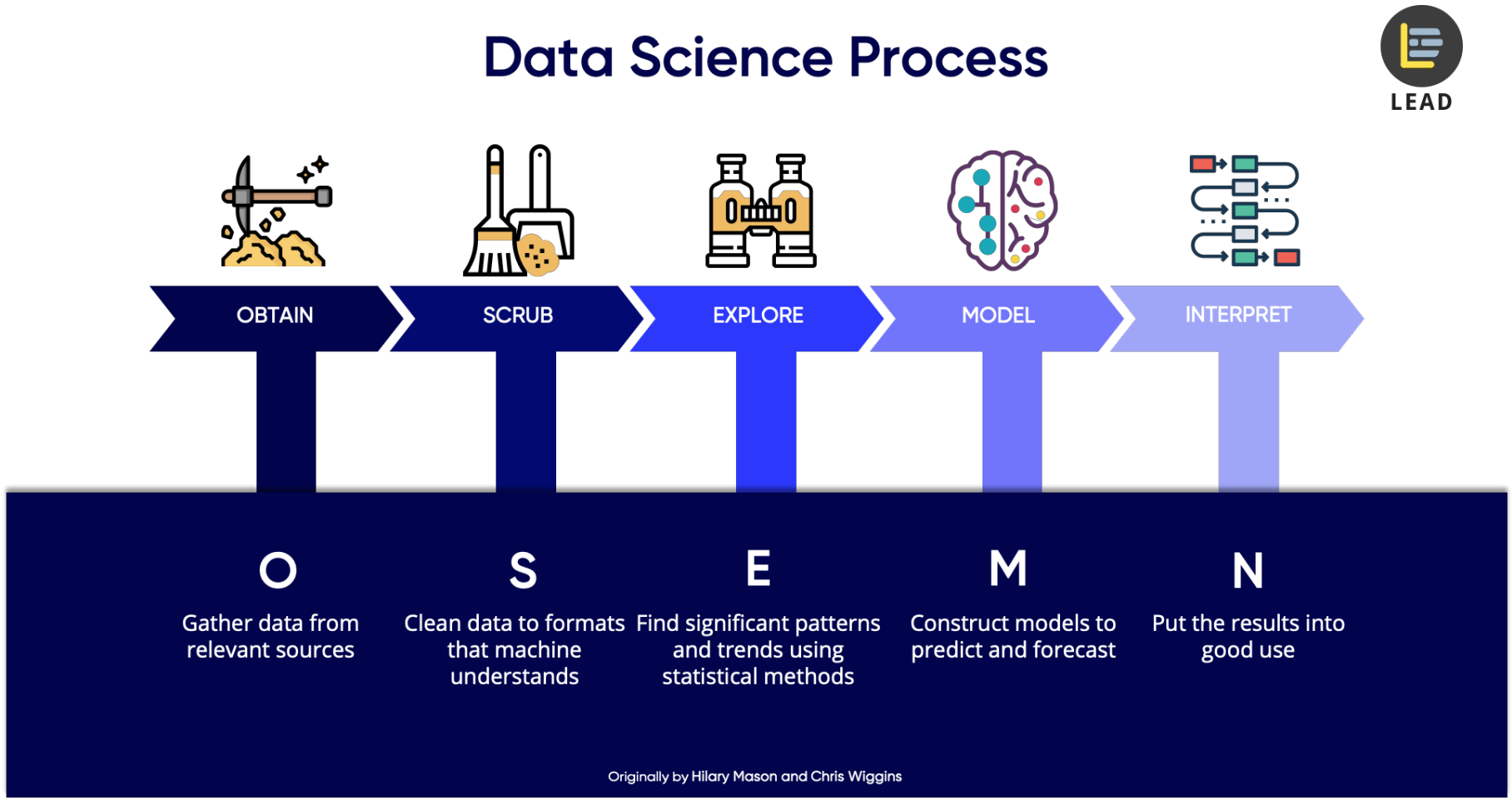
데이터의 패턴을 발견하고 이해하는 작업

- 데이터의 세계관에 대해서 파악하는 것!
- 데이터 분석의 출발점

Part1_Bayesian statistics

EDA

EDA란?



Part1_Bayesian statistics

EDA

EDA란?

EDA는 데이터 가공과 함께 진행.

데이터가공이란?

- 시각화, 모형화에 용이하도록 데이터를 **재구성** 하는 것
- 일반적으로 70~80%의 시간을 쏟아야 함...

Part1_Bayesian statistics

데이터 가공

데이터 가공

Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior
AUC2016XC	Aman	30	3.5	Jr

Missing Data

Inconsistent Data

Noisy Data

Step 1. 일변량 분석

Part1_Bayesian statistics

일변량 분석

일변량분석 - 개괄

★ 데이터의 전체적인 내용 파악 → head()

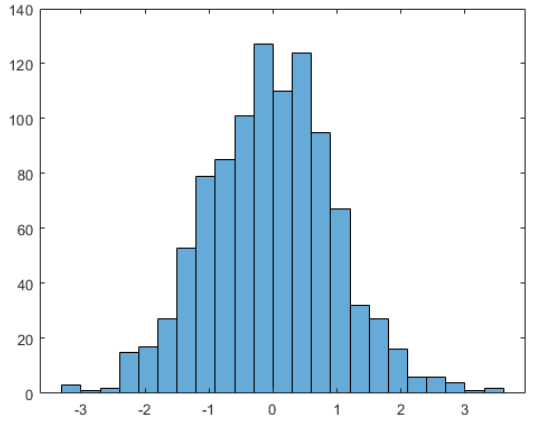
★ 결측치, 중복치, 이상치 파악 → isnull(), duplicated()

→ 모든 데이터에 대해 기본적으로 시행

Part1_Bayesian statistics

일변량 분석

일변량분석 - 수치형 변수 (numerical variable)

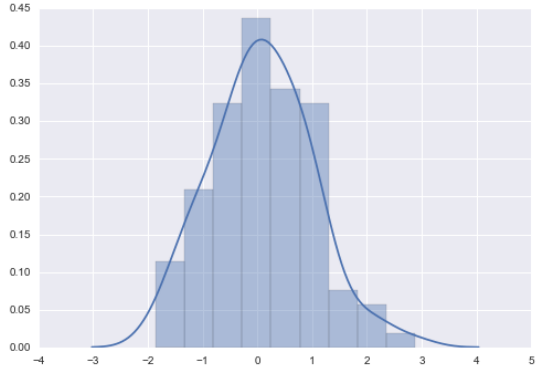


★ 요약 통계량을 확인한다.

★ 히스토그램을 통해 분포적 특징을 파악한다.
→ 정규성, 너무 marginal 하지는 않은지...

★ 박스플롯을 통해 이상치를 파악

→ 이상치를 단순한 입력오류로 규정할 수 없는 경우 robust한 방법론을 사용.



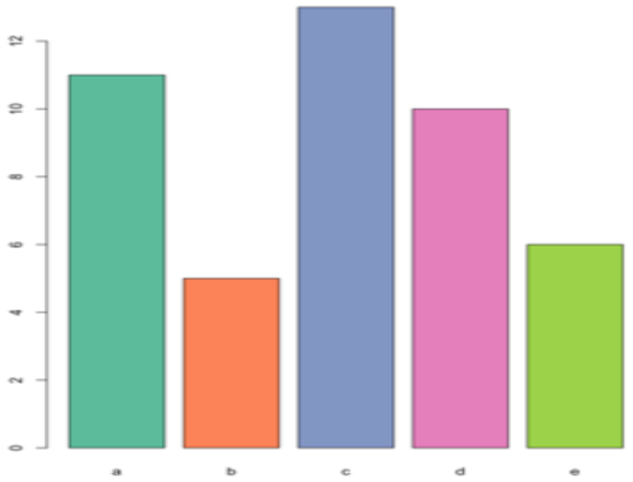
Part1_Bayesian statistics

일변량 분석

일변량분석 - 범주형 변수 (categorical variable)

★ Barplot, value_counts() 등을 이용해 도수 분포 확인

★ Dummification → one-hot encoding (if necessary)



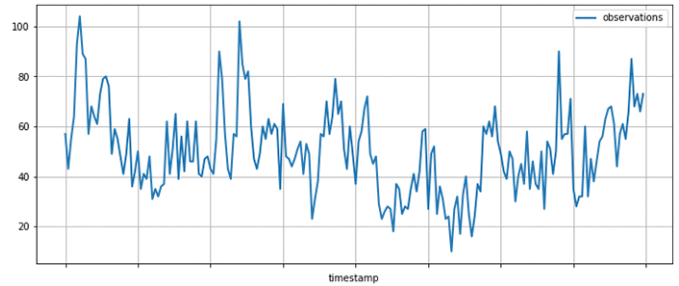
Part1_Bayesian statistics

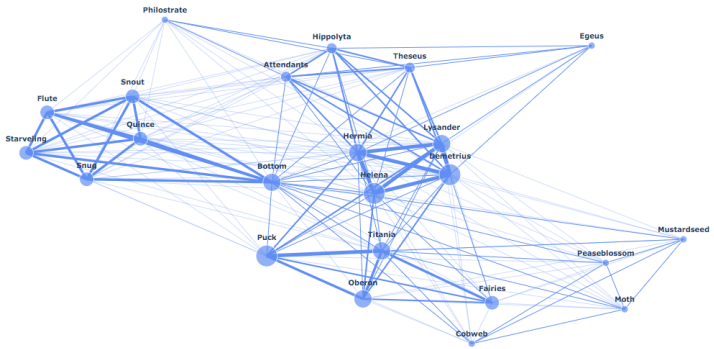
일변량 분석

일변량분석 – 비정형 데이터

★ 텍스트 데이터, 시계열 데이터, 네트워크 데이터 등...

★ Data-specific knowledge가 필요





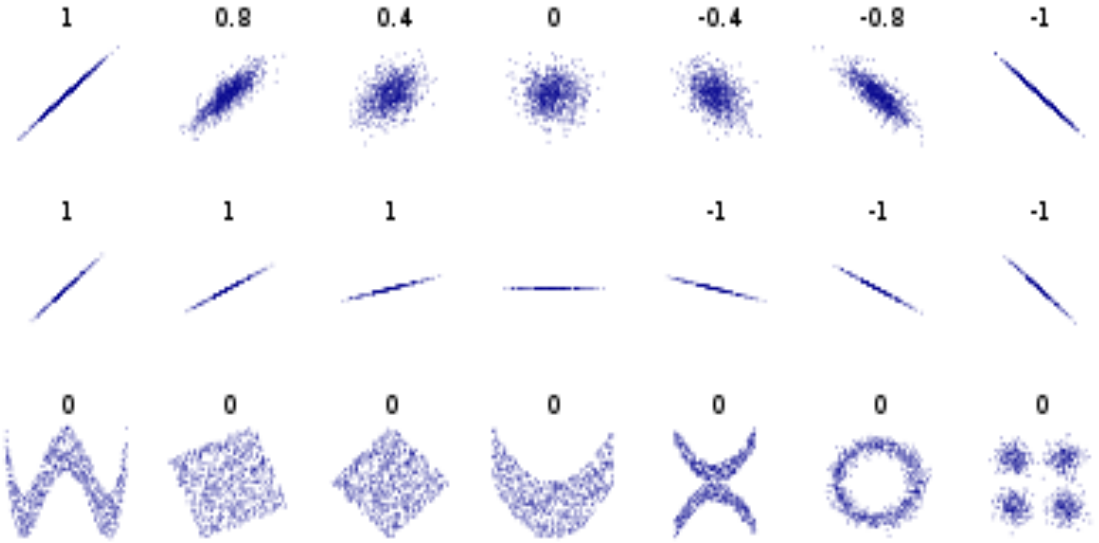
Step 2. 이(다)변량 분석

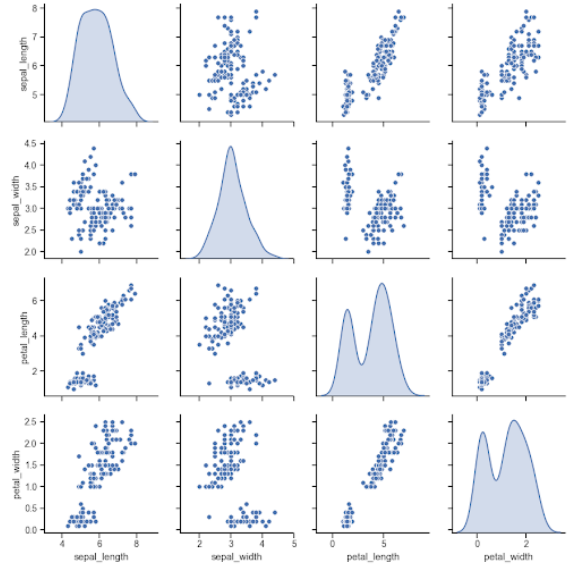
Part1_Bayesian statistics

이변량 분석

다변량분석 – numerical vs numerical

★ 산점도를 통해 상관관계 파악 (상관계수를 믿지 말자!)



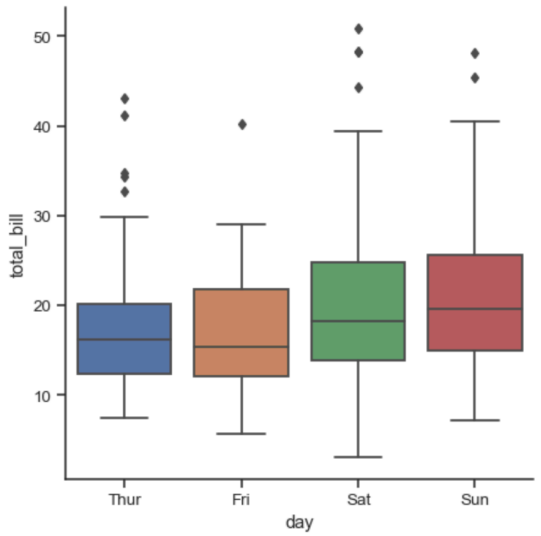


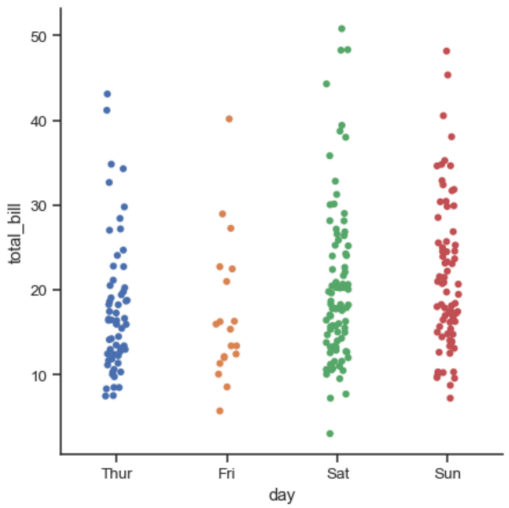
Part1_Bayesian statistics

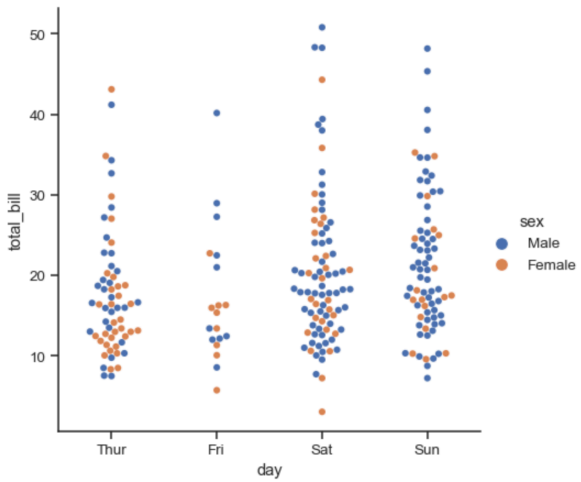
이변량 분석

다변량분석 – numerical vs categorical

- Side-by-side boxplot, jittered scatter plot 을 통해 시각화
  → 집단 간 차이를 중점적으로 파악할 것
- groupby() 를 사용해 그룹별 통계량 집계





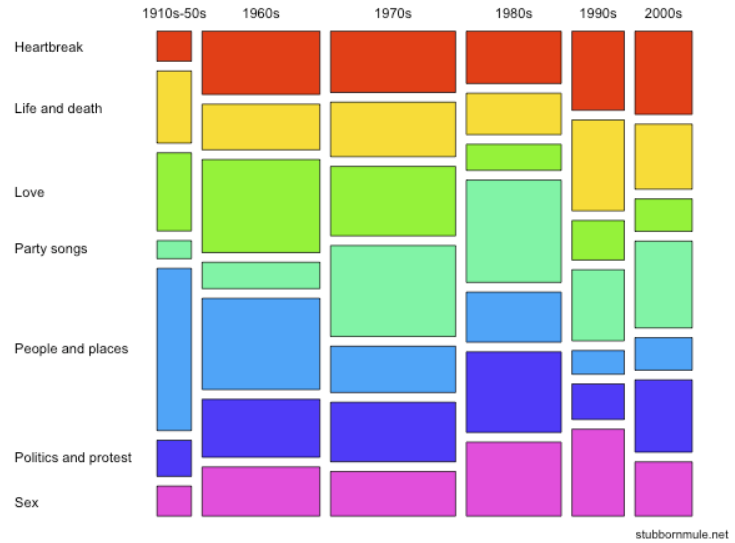


Part1_Bayesian statistics

이변량 분석

다변량분석 – categorical vs categorical

- Mosaic Plot 를 통해 시각화
  → 면적을 중심으로 볼 것
- pivot table, groupby() 등을 통해 그룹별 통계량 집계



Part1_Bayesian statistics

Tip

EDA Tip

- 편향되지 않은 시각으로 분석에 임하자.
- 주어진 데이터를 그대로 사용하는 경우는 드물다! 숲의 관점에서 데이터를 바라보자.
  → Latent variable finding, feature extraction, …
- 다변량 분석을 지향하자.
- 백문이 불여일견이다! 시각화를 적극 활용하자!

Part1_Bayesian statistics

Tip

좋은 시각화란? (by Edward Tufte)

1. 비교, 대조, 차이를 명확히 드러내자. (색깔을 적절히 사용)
2. 시각화를 위한 시각화는 지양하자.
3. 미니멀리즘을 지향하되, 하나의 그래프에 최대한 많은 변수를 담자.
4. 의미 있는 변수명을 사용하자.
5. 설명이 필요 없는 플롯을 지향하자. (labeling, legend)
6. 시각화 코드를 버전 컨트롤하자.
7. 모든 변수를 어떤 방식으로든 시각화 해보려고 노력하자.
8. 시각화를 위한 데이터를 준비할 때 집계 함수 (apply, groupby, pivot_table)를 적절히 사용하자.