

탐색적 데이터 분석 1: 기초 통계 처리

기술통계를 이용한 데이터 요약 방법



부산대학교 정보·의생명공학대학
정보컴퓨터공학부



탐색적 데이터 분석

“데이터 기반의 문제 해결 단계 중 데이터 문제를 정의하고, 데이터를 수집하고, 분석을 위해 준비하는 과정을 (거치면) [...] 분석에 적합한 (테이블 형태) 양질의 데이터를 얻을 수 있을 것이다. [...] 경험 많은 데이터 과학자라면 [...] 우선 주어진 데이터의 모든 측면을 철저히 이해하려고 노력할 것이다. 데이터 수집 과정에서 세운 모든 가정이 맞는지, 혹시 기대하지 않았던 새로운 패턴이 발견되지 않는지, 추가적인 데이터가 필요하지 않은지 등을 알고 싶어할 것이다. 이처럼 주어진 데이터를 다양한 각도에서 관찰하고 이해하는 과정을 탐색적 데이터 분석 (Exploratory Data Analysis, EDA)이라고 부른다.”



▲ 탐색적 데이터 분석에는 원본 데이터, 요약 통계값, 시각화가 모두 필요하다.

출처: 헬로 데이터 과학

1. Business Problem

2. Data Acquisition

3. Data Preparation

4. Exploratory Data Analysis

5. Data Modeling

6. Visualization and Communication

[Source]

탐색적 데이터 분석 cont.

❖ “데이터 과학의 첫 단계는 **주어진 데이터를 이해**하는 겁니다. 데이터 종류에 관계없이 **확인해야하는 보편적인 질문**을 준비해두면 데이터와 멋지게 첫 데이트를 할 수 있습니다. 필자는 보통 아래와 같은 질문을 던집니다.

[...] 가장 중요한 질문은 ‘**이 데이터로 주어진 문제를 풀 수 있느냐**’는 겁니다”

- **데이터의 크기**는 얼마나 되나?
- 주어진 데이터가 데이터셋의 전부인가 아니면 일부인가?
- 이 데이터가 **모집단을 잘 대표하는가?** 예를 들어 특정 지역이나 연령대의 이용자를 과다 대표하는 데이터는 아닌가?
- **특이한 이상치나 잡음이 심하지는 않은가?** 예를 들어 웹 서버의 트래픽이 알고 보니 전체 데이터의 99%가 DDOS 공격 때문에 발생한 트래픽이라면 이 데이터를 포함해야 하는가?
- 원본 데이터가 아니라 임의로 가공한 데이터가 포함되고 있지는 않은가?
- 데이터별 **식별자가 존재하는가?**
- 식별자가 잘 설정되어 있는가? 혹은 중복되어 나타날 수도 있는가? 만일 그렇다면 어떻게 해야하는가?
- (두 데이터셋을 합쳐야 하는 경우라면) 이 둘은 정말 **같은 종류의 데이터셋인가?**
- 만일 값이 **누락된 데이터 표본**이 있다면 얼마나 그리고 왜 누락되었는가?

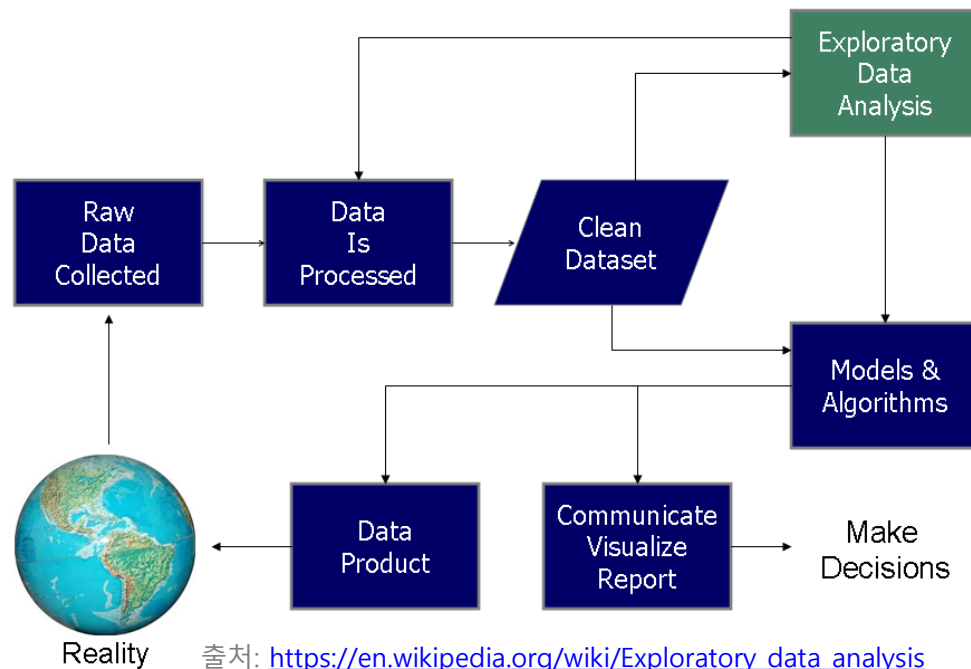
출처: 처음 배우는 데이터 과학

탐색적 데이터 분석 cont.

❖ Exploratory Data Analysis (EDA) Objectives 출처: towardsdatascience.com

1. Quickly describe a dataset; number of rows/columns, missing data, data types, preview.
2. Clean corrupted data; handle missing data, invalid data types, incorrect values.
3. Visualize data distributions; bar charts, histograms, box plots.
4. Calculate and visualize correlations (relationships) between variables; heat map.

Data Science Process



기초 통계 처리

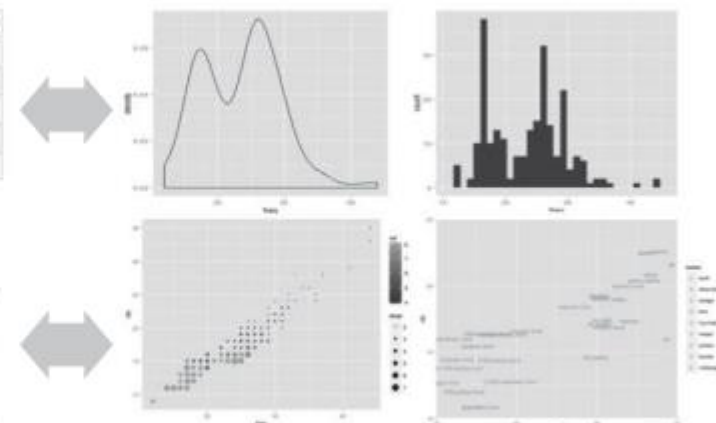
원본 데이터

maker	class	model	year	displ	cyl	drv	cty	hwy
audi	compact	a4	1999	1.8	4	f	18	29
audi	compact	a4 quattro	1999	1.8	4	4	18	26
audi	midsize	a6 quattro	1999	2.8	6	4	15	24
chevrolet	suv	c1500 suburban 2wd	2008	5.3	8	r	14	20
chevrolet	2seater	corvette	1999	5.7	8	r	16	26

요약 통계값

#	maker	class	model	year				
#	dodge	2seater	1	11	Min.	1999		
#	toyota	compact	147	cam 1500 pickup 4wd	10	1st Qu.	1999	
#	volkswagen	midsize	161	civic	1	9	Median	2004
#	ford	minivan	111	dakota pickup 4wd	1	9	Mean	2004
#	chevrolet	pickup	133	jetta	1	9	3rd Qu.	2008
#	audi	subcompact	135	mustang	1	9	Max.	2008
#	(Other)	suv	162	(Other)	1177			

시각화



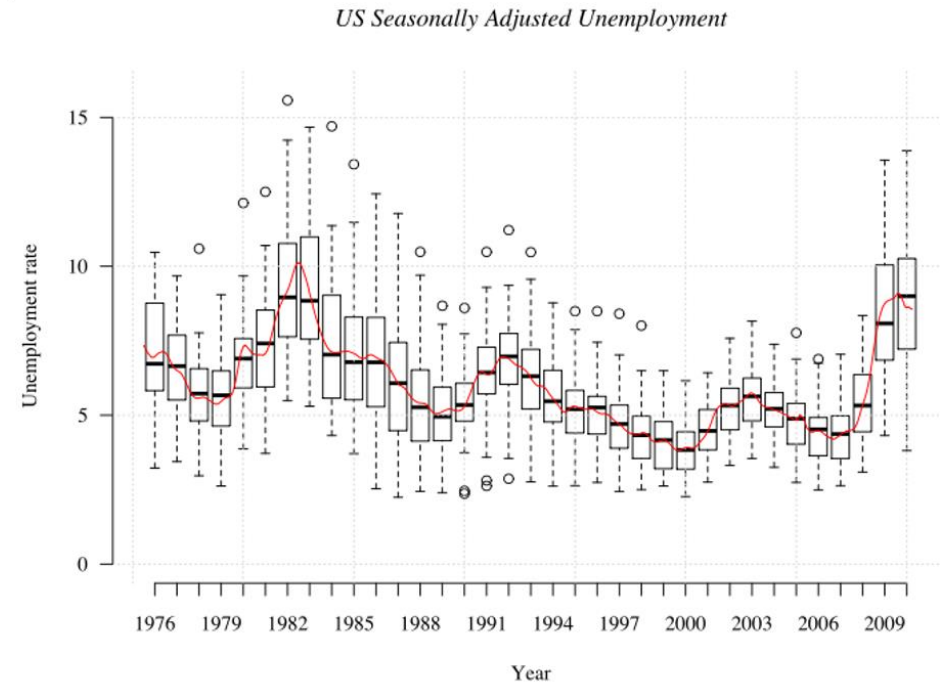
▲ 탐색적 데이터 분석에는 원본 데이터, 요약 통계값, 시각화가 모두 필요하다.

통계학의 목적

❖ 통계학의 목적(1): 가지고 있는 데이터의 설명

- 기술 통계 (Descriptive Statistics)
- 수집된 데이터의 정리와 요약을 목적으로 삼을 수 있다.

탐색적 데이터 분석(EDA)과정에서 많이 사용



❖ 통계학의 목적(2): 모르는 데이터의 예측

- 추리 통계 (Inferential Statistics)
- 가지고 있는 데이터를 이용하여 모르는 데이터를 추측할 목적



자료의 요약과 기술통계

❖ There are 3 main types of descriptive statistics:

- The **distribution(분포)** concerns the frequency(도수) of each value.
- The **central tendency(집중화 경향, 중심화 경향)** concerns the averages of the values.
- The variability or **dispersion(분산)** concerns how spread out the values are.

Frequency Distribution (도수 분포)

❖ A data set is made up of a distribution of values, or scores. In tables or graphs, you can summarize the frequency of every possible value of a variable in numbers or percentages.

■ Frequency Table (도수 분포표)

- Counts or frequency(도수)
- Relative frequency(상대도수)

■ Graphs

- Nominal & Ordinal Data (범주형)
:bar graphs & pie charts
- Interval & Ratio data(수치형)
:histograms

B+ Ao B+ Ao B+ Co Bo Ao C+ Bo B+ Bo C+ B+ Bo
B+ Ao B+ D+ A+ A+ Ao C+ Bo Ao B+ A+ Ao C+ Bo
C+ Co Bo Ao C+ Bo A+ Ao C+ B+ Ao B+ Bo C+ Ao
A+ A+ Ao C+ Bo C+ Ao A+ A+ Ao B+ Bo A+ Ao Bo



성적	(빈)도수
A+	9
Ao	15
B+	11
Bo	12
C+	10
Co	2
D+	1
계	60


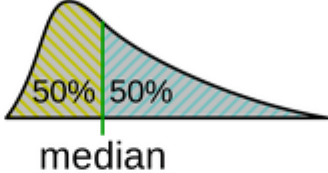
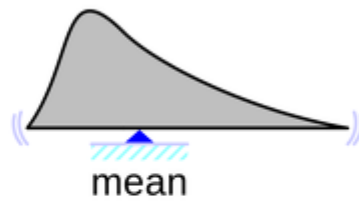
Gender	Number
Man	182
Woman	235
No answer	27

Library visits in the past year	Percent
0-4	6%
5-8	20%
9-12	42%
13-16	24%
17+	8%

Measures of Central Tendency

❖ Measures of central tendency help you find the middle, or the average, of a data set. The 3 most common measures of central tendency are **the mode, median, and mean**.

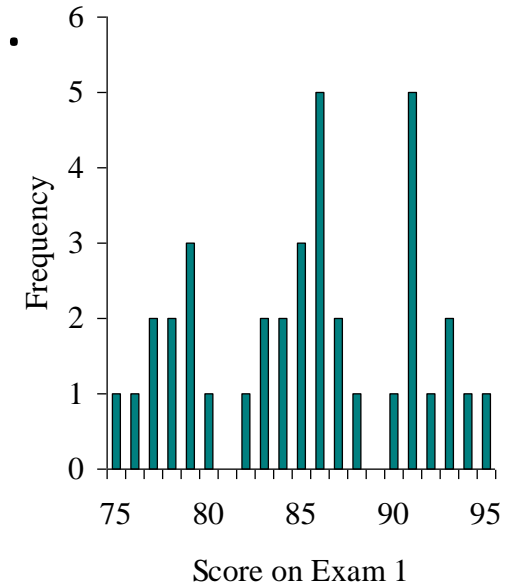
- Mode: the most frequent value.
- Median: the middle number in an ordered data set.
- Mean: the sum of all values divided by the total number of values.

	최빈치(mode)	중앙치(median)	산술평균(mean)
의미	<p>• 가장 빈번하게 나타나는 값</p> 	<p>• 자료를 크기 순으로 나열했을 때, 중앙에 위치하는 값</p> 	<p>• 자료를 모두 더해서 자료의 개수로 나눈 값</p> 
특징	<p>• 명목자료에서는 최빈치가 대푯값이다.</p>	<p>• 서열자료의 경우 평균을 사용할 수 없으므로 중앙치를 사용한다.</p>	<p>• 일부 극단적인 값들에 크게 영향을 받는다. • 수학적 연산에 의해 계산되므로 수리적인 조작이 용이하다.</p>
예	<p>유행하는 가방 인기 투표</p>	<p>학교 석차 100명 중 50 등</p>	<p>년간 평균 강우량 기말 고사 평균 점수</p>

The mode

❖ The **mode** is the most frequent number in a collection of data.

- Example A: 3, 10, 8, 8, 7, 8, 10, 3, 3, 3
 - The mode of the above example is 3, because 3 has a frequency of 4.
- Example B: 2, 5, 1, 5, 1, 2
 - This example has no mode because 1, 2, and 5 have a frequency of 2.
- Example C: 5, 7, 9, 1, 7, 5, 0, 4
 - This example has two modes 5 and 7. This is said to be bimodal.



❖ The mode **works best with categorical data**. It is the only measure of central tendency for **nominal variables**, where it can reflect the most commonly found characteristic (e.g., demographic information). The mode is also useful with **ordinal variables** – for example, to reflect the most popular answer on a ranked scale (e.g., level of agreement).

TO BE CONTINUED