데이터 탐색

Kyungsik Han

본 영상에서 다룰 내용

CODING

- Exploratory Data Analysis (EDA)

- Numerical summaries of data (통계적 특성 파악)
  - Descriptive statistics
- Graphical summaries of data (그래프를 통한 탐색)
  - Visualization

데이터 탐색

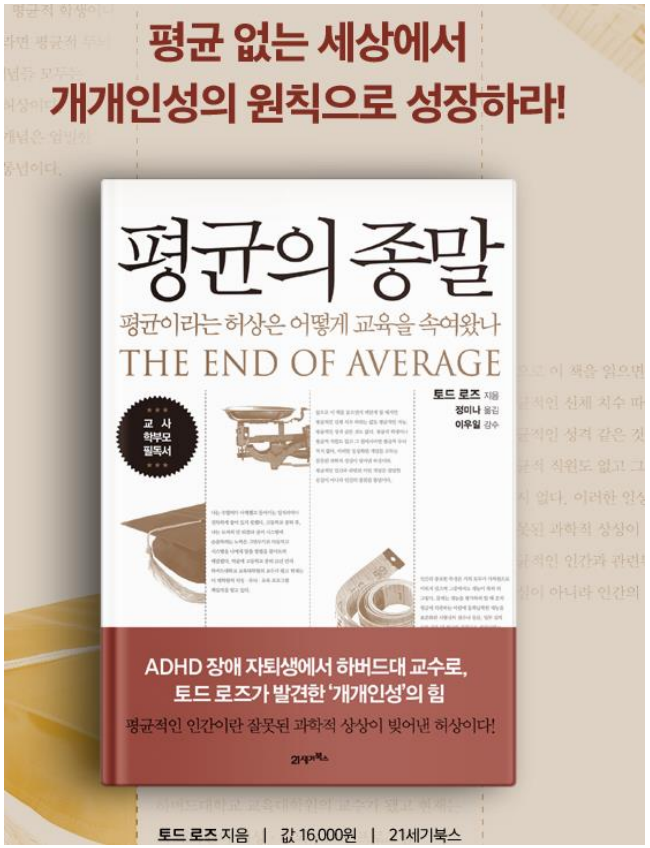
CODING

- 수집한 데이터의 전체적인 특성을 분석
- 본격적인 데이터 분석에 앞서 수집한 데이터가 분석에 적절한지 알아보는 과정

- 탐색적 데이터 분석 (Exploratory Data Analysis: EDA)
  - 기본적인 통계적 특성 파악: 숫자형 데이터 평균, 최대/최소값, 표준편차, 분산 등
  - 그래프를 통한 데이터 탐색: 데이터 시각화를 통해 데이터의 특성을 그래프로 나타내는 것이 탐색에 효과적

평균의 위험성


FLAW of AVERAGES
AVERAGE DEPTH 3ft.
DANZIGER


평균 없는 세상에서
개개인성의 원칙으로 성장하라!
평균의 종말
평균이라는 허상은 어떻게 교육을 속여왔나
THE END OF AVERAGE
교 사
학부모
필독서
토드 로즈 지음
정미나 옮김
이우일 감수
ADHD 장애 자퇴생에서 하버드대 교수로,
토드 로즈가 발견한 '개개인성'의 힘
평균적인 인간이란 잘못된 과학적 상상이 빚어낸 허상이다!
21세기북스
토드 로즈 지음 | 값 16,000원 | 21세기북스

중간값 (Median)

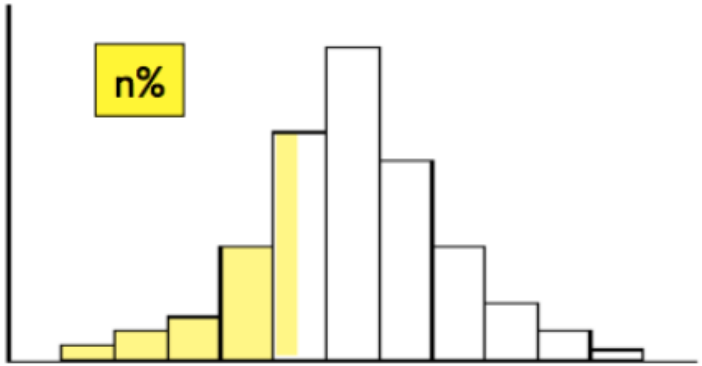
CODING

- 정확히 중간에 위치하는 값
  - Useful for skewed distributions or data with outliers
  - More robust than mean
  - Difficult to handle theoretically

Percentiles (Quantiles)

CODING

- The n^{th} percentile is a value such that n% of the observations fall at or below of it

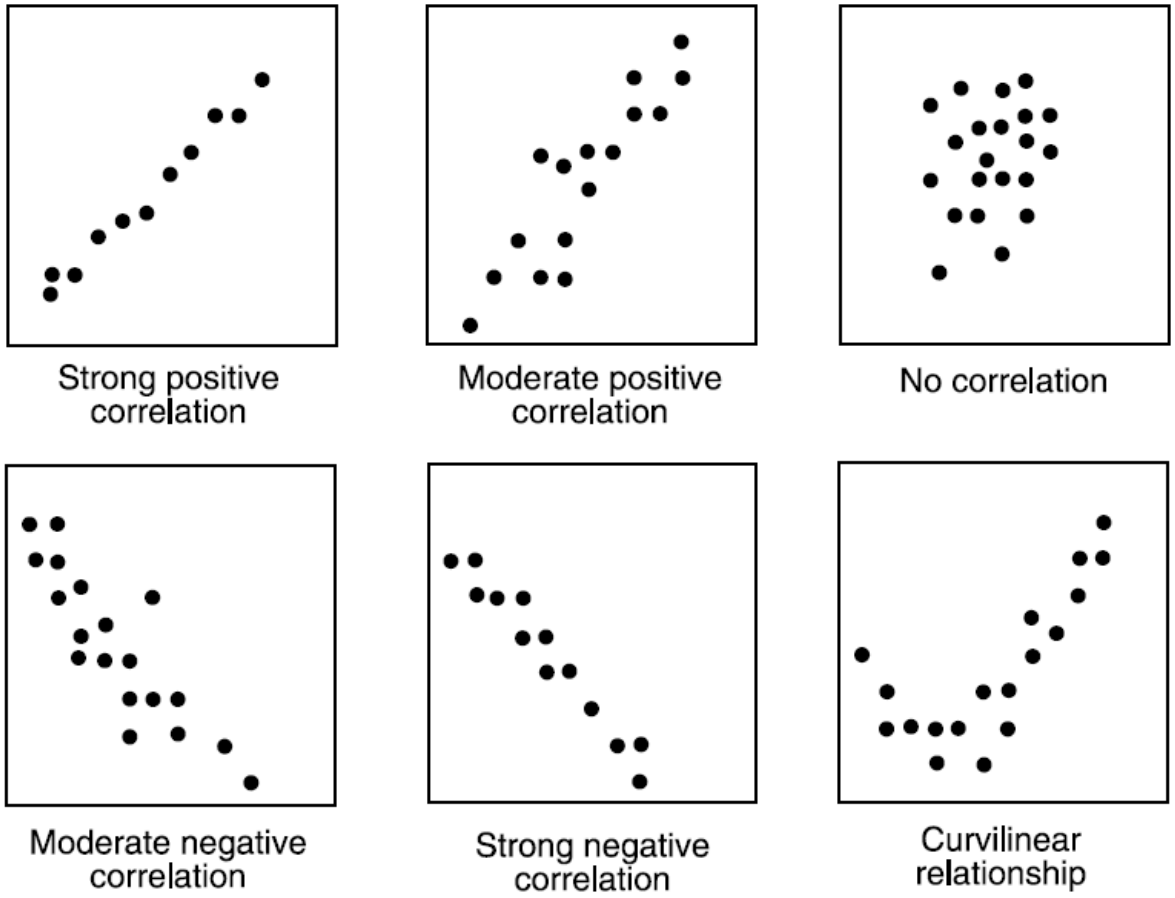


- Q1 : 25th percentile
- Median: 50th percentile
- Q3 : 75th percentile
- IQR: Interquartile range (25 to 75%: Q3-Q1)

상관(관계) 분석

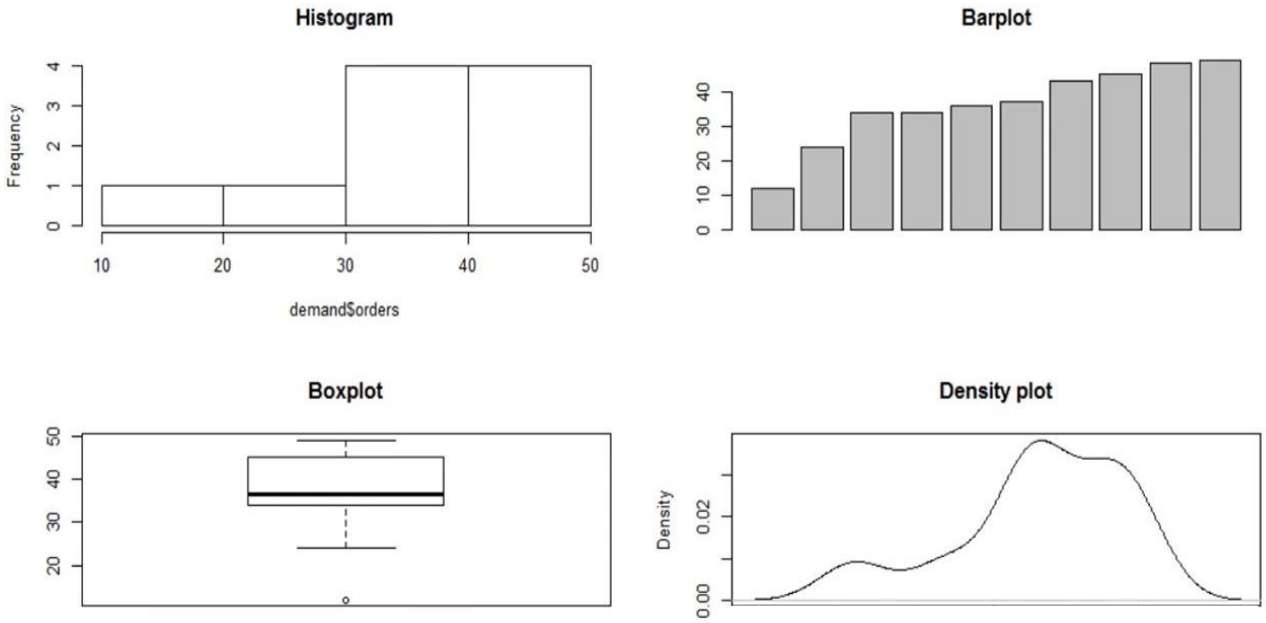
CODING

- 주어진 값들의 연관성을 파악
- -1 (음의 관계) ~ +1 (양의 관계) 사이의 값을 가짐



시각화

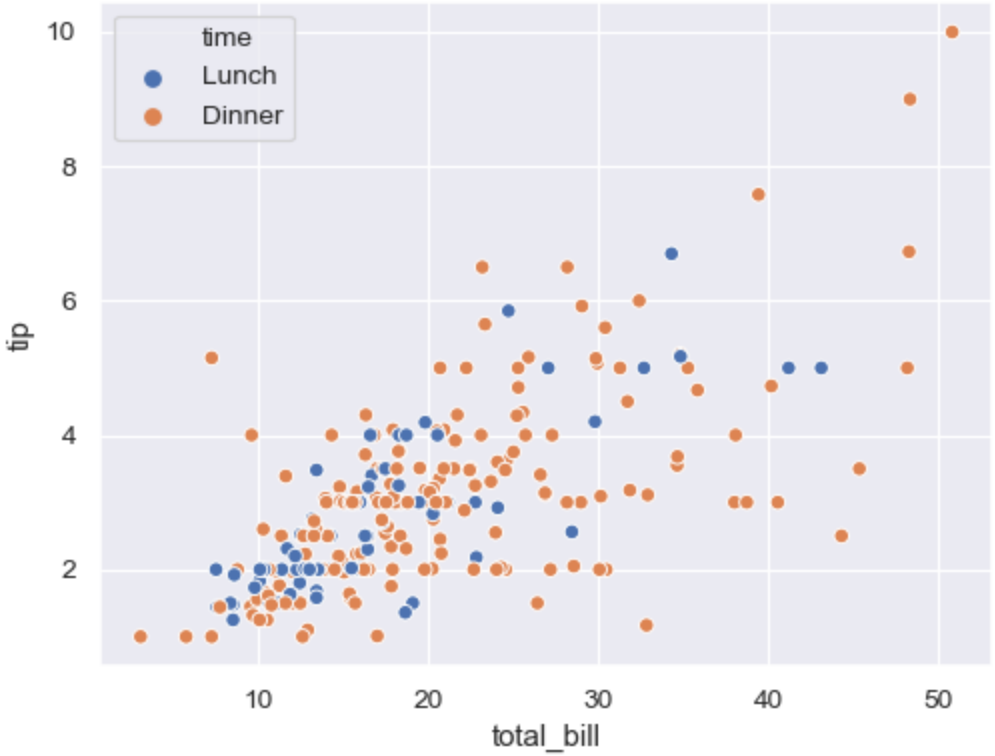
- Bar plot/line plot: 가장 많이 사용
- Box plot: 5가지 분석 결과 보여줌
- Histogram: 숫자의 분포를 보여줌



시각화

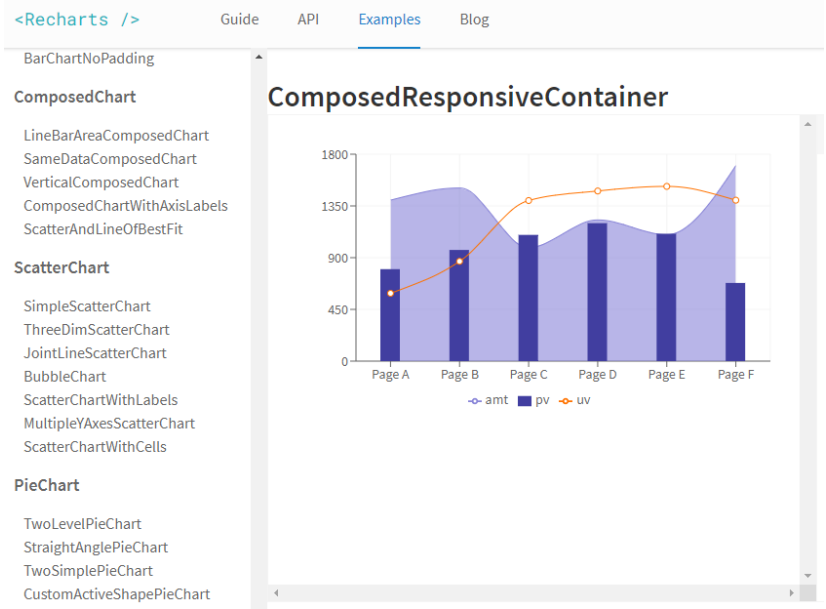
CODING

- Scatter plot



시각화 (more)

https://observablehq.com/@d3/gallery



http://recharts.org/en-US/examples/ComposedResponsiveContainer

데이터 전처리 (pre-processing)

데이터 전처리

CODING

- Data preprocessing
- 수집한 데이터를 분석하기 좋기 변환하는 모든 작업으로 데이터 정제(cleaning)라고도 함
- 분석 목적에 맞는지 데이터의 품질을 확인하고 필요하면 품질을 높이는 방법
- 데이터 전처리의 필요성
  - 데이터가 너무 크면 적절한 크기로 줄여야 함
  - 수집한 데이터가 비정형 데이터면, 정형 데이터로 바꾸어야 함
  - 데이터의 오류를 찾거나 빠진 값을 찾는 작업 필요
  - 목적에 맞게 데이터를 가공해야 함
  - 데이터의 상태와 가치를 파악할 필요가 있음

데이터 전처리

- 데이터 전처리의 세부 작업
  - 데이터 필터링(filtering): 필요한 데이터만 선택함
  - 데이터 변환(transformation): 데이터의 형식 변경, 단위의 표준화
  - 데이터 통합(integration): 여러 소스의 데이터를 합치는 작업
- 데이터 품질
  - 데이터를 얼마나 믿고 쓸 수 있을지를 나타내는 신뢰성을 의미
  - 정확성과 적시성(최신성 등)이 보장되어야 함
  - 수동 입력 데이터의 신뢰성

데이터 정제(data cleaning)

CODING

구분	처리 방법
결측치 처리	• 결측치가 포함된 항목을 모두 버리는 방법 • 결측치를 적절한 값으로 대체 • 분석 단계로 결측치 처리를 넘김 • 별도의 범주형 변수를 정의하여 추적 가능하게 관리
틀린 값 처리	• 틀린 값이 포함된 항목을 모두 버리는 방법 • 틀린 값을 다른 적절한 값으로 대체 • 분석 단계로 틀린 값의 처리를 넘김
이상치 검출	• 값이 일반적인 범위를 벗어나 특별한 값을 갖는 경우 • 데이터 분석 과정의 활동이므로 분석 단계로 넘김

데이터 변환

CODING

- 데이터를 분석하기 쉬운 형태로 바꾸는 작업

구분	처리 방법
범주형으로 변환	• 수치 데이터의 세세한 구분이 오히려 혼란스러울 때
일반 정규화	• 수치 데이터의 범위가 각각 다를 때
z-score 정규화	• 일반 정규화에 표준 편차를 고려한 변환
로그 변환	• 로그를 취하는 것이 타당할 때 (로그 정규 분포 등)
역수 변환	• 역수를 사용하면 선형적인 특성이 파악 가능할 때
데이터 축소	• 불필요한 데이터의 제거

범주형으로 변환

- 수치 데이터의 개발 값 구분이 오히려 혼란스러울 때
- 수치 데이터를 범주형으로 변환하여 사용
  - 나이: 10, 20, 30, 40, 50대
  - 연간 소득: 고, 중, 저 소득층
- 범주형 변환 시 각 구간의 등급은 균등 or 차등 배정 가능
- 균등 배정보다 차등 배정이 더 자연스럽다고 느낄 때 차등 배정을 사용함

일반 정규화

CODING

- Normalization
- 수집 데이터의 데이터 범위가 서로 다를 경우, 이를 같은 범위로 변환해서 사용하는 방법
- 단순하게 비례화 시킨 값으로 최대치와 최소치를 고려

시험 A	시험 B
• 10점 만점 • 갑의 시험 점수: 7 ➔ 7/10: 0.7 • 을의 시험 점수: 8 ➔ 8/10: 0.8	• 50점 만점 • 갑의 시험 점수: 30 ➔ 30/50: 0.6 • 을의 시험 점수: 20 ➔ 20/50: 0.4
• 성적을 0(최소) – 1(최대)로 정규화 • 갑의 성적 평균 = 0.65 • 을의 성적 평균 = 0.60	

z-score 정규화 (1/2)

CODING

- 데이터 분포를 평균은 0, 표준 편차는 1이 되도록 정규화 하는 방법

100점 만점, 학급 평균 60점
표준편차: 과목 A=20, 과목 B=5

학생	과목 A	과목 B	평균
갑	90	80	85
을	80	90	85

성적이 더 높은 학생은?

z-score 정규화 (2/2)

CODING

z 변환 공식

$$z = \frac{x-\mu}{\sigma}$$

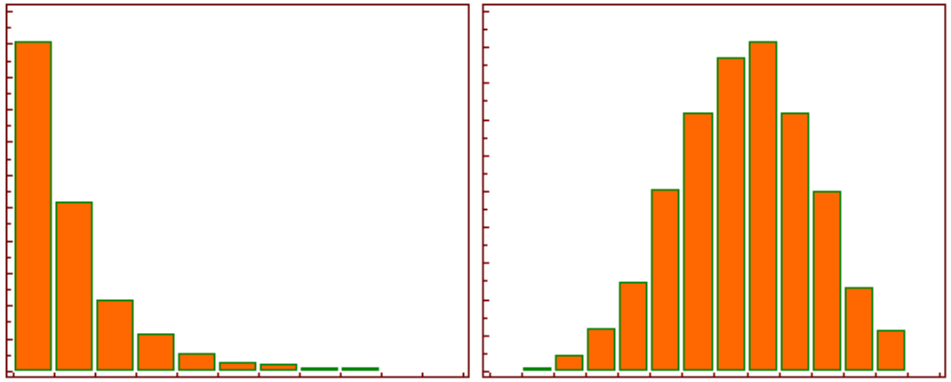
- x (해당 값)
- μ (뮤, 평균)
- σ (시그마, 표준편차)

	학생	과목 A	과목 B	평균
변환 전	갑	90	80	85
	을	80	90	85
변환 후	갑	(90-60)/20 = 1.5	(80-60)/5 = 4.0	2.75
	을	(80-60)/20 = 1.0	(90-60)/5 = 6.0	3.50

로그 변환, 제곱근 변환

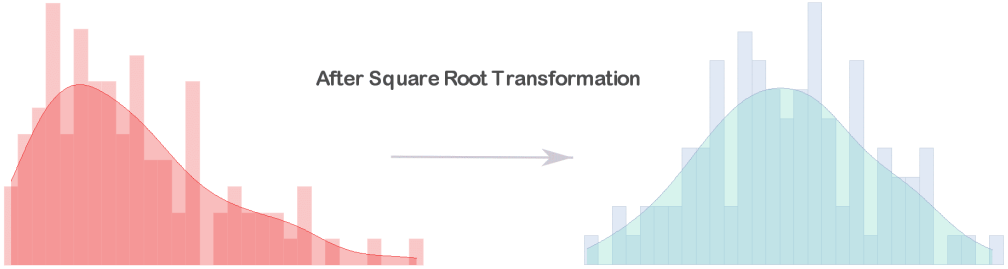
CODING

- 로그를 취한 값을 사용
  - 로그를 취했을 때 정규 분포(로그 정규 분포)에 가깝게 분포



로그 변환 전 로그 변환 후

- 제곱근 변환



제곱근 변환 전 제곱근 변환 후

역수 변환

CODING ✖

- 역수를 사용했을 때 오히려 선형적인 특성을 가지므로 의미를 해석하기가 쉬워지는 경우
- x의 역수는 x와 곱해서 1이 되는 수 ($1/x$ or x^{-1})
- 원하는 목적 변수와 선형적인 관계인 변수를 선택하는 것
- 속도와 시간의 관계에서 어떤 것을 특성 변수로 잡는 것이 좋을지 선택하는 경우

데이터 축소 (Data reduction)

CODING ✖

- 같은 정보량을 가지면서 데이터의 크기를 줄이는 것
- 데이터를 줄이면 데이터를 다루기가 편리
- 불필요한 데이터를 제거하여 분석 속도와 성능 개선
- 데이터 축소 기법으로 PCA (Principal Components Analysis) 사용
  - 기존의 속성들을 대표하는 속성 값을 추출하는 것
- 주어진 여러 데이터를 대표할 수 있는 새로운 변수를 만들어 사용 가능
  - BMI (비만도) 지수 = 몸무게 / 키*키
  - BMI >= 30 ➔ BMI = 1 (비만), otherwise BMI = 0 (정상)

샘플링 (Sampling)

CODING

- 구할 수 있는 전체 데이터 중에서 분석에 필요한 데이터만을 취하는 것
- 모두 사용하여 모델을 만들고 분석하는 것이 항상 좋은 것은 아님 (시간과 자원의 낭비)
- 최소한의 샘플 데이터를 가지고 사전 타당성 조사
- 분석 모델(알고리즘)의 큰 방향성 결정에 유용
- 샘플 데이터가 전체 데이터의 특징을 계속 유지할 수 있어야 함
  - 나이별, 소득별, 성별 등 균일성 유지

다음 영상에서 배울 내용

CODING X

- 데이터 탐색 실습
- practice_step_by_step 코드 기반

수고하셨습니다