인공지능을 위한 확률과 통계

1.1 탐색적 자료분석

서울대 통계학과 장원철

탐색적 자료분석이란?

- 데이터 분석을 위한 두가지 접근법
  - 탐색적 자료분석 (Exploratory Data Analysis)
    - 데이터의 특징과 구조에 대한 탐구
    - 인사이트의 생성, 가설과 모형의 도출
    - 선 데이터, 후 분석
- 확증적 자료분석 (Confirmatory Data Analysis)
  - 가설, 모형의 타당성, 재현성 평가
  - 모형 적합도, 가설검정, 신뢰구간
  - 분석계획 → 데이터확보 → 분석

허명회, 데이터 분석의 철학과 과학성, R User Conference 2015

탐색적 자료분석이란?

EDA vs CDA

[예제1] 감기에 걸리는 사람과 걸리지 않는 사람들 간에 어떤 차이가 있는가를 수십 가지 측면에서 살펴보았다. 그 결과, 비타민 C를 복용하는 사람들이 감기에 잘 걸리지 않음을 알게 되었다. 그렇다면, 비타민 C 복용이 감기를 예방하는 효과락 있다고 말할 수 있는가? [EDA]

EDA로 이에 대한 답을 하긴 어렵다. 비교실험을 설계하여 새롭게 자료를 수집하여 가설을 확인해볼 필요가 있다. [CDA]

[예제2] 대형마켓에서 고객들의 구매내역 자료를 분석한 결과, 일부 고객들은 다른 고객들에 비해 유기농 식재료 비중이 크게 나타났다. 그들이 어떤 생각을 하는 사람들인가? 이에 대해 몇 개의 추측이 생성되었다. [EDA]

추측이 맞는지 확인하기 위하여 전체 고객의 일부를 선택하여 몇가지 인구사회적 속성과 연소득, 그리고 소비와 삶에 대한 태도를 조사하여 구매내역과 연결해 확인하였다. [CDA]

정형화된 데이터의 요소

데이터 유형

- 데이터 유형
  - 수치형 (numerical)
    - 연속형 (continuous): 모든 실수값을 취하는 경우
    - 이산형 (discrete): 횟수와 같이 정수값만 취하는 경우
  - 범주형 (categorical)
    - 순서형 (ordinal): 값들 사이에 분명한 순위가 있는 경우 (예: 학년)
    - 명목형 (nominal): 주어진 범위에서 값을 가지는 경우 (예: 휴대폰 제조사)

테이블 데이터

- 데이터 프레임 (data frame):
- 피처 (feature)
- 결과 (outcome)
- 레코드 (record)

연습문제

- 전화번호는 다음중 어디에 속하는가?

1. 수치형, 연속형
2. 수치형, 이산형
3. 범주형, 명목형
4. 범주형, 순서형

테이블 형식이 아닌 데이터구조

- 공간데이터
- 네트워크 데이터

위치 추정

평균

- 데이터를 대표하는 값들에 대해서 알아보자!
- 평균은 모든 값의 총합을 개수로 나눈 값, 즉 산술평균을 의미한다.

$$\bar{x} = \frac{x_1 + x_2 + \cdot + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- 평균의 변형된 형태로 절사평균(trimmed mean)을 들 수 있다. $x_{(i)}$를 *i*번째로 작은 값이라고 하면

$$\bar{x}_{TM} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

- 절사평균을 사용한 예로 피겨스케이팅에서 심판들의 점수중 최고점과 최저점을 제외한 점수의 평균으로 선수의 성적을 산출하고 있다

위치추정

가중평균

- 가중평균: 각 데이터 값에 사용자가 지정한 가중치 w_i를 곱한 값들의 총합을 다시 가중치의 총합으로 나눈 값

- $\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

- 여론조사에서 전체 인구구성의 성별, 연령별로 비율을 고려하여 가중평균을 통하여 최종적으로 예상치를 계산한다.

위치 추정

중간값과 로버스트 통계량

- 앞의 요약치들은 특잇값에 민감하다. 특이점에 민감하지 않은 통계량을 로버스트하다고 한다.
- 위치를 나타내는 요약치중 중간값(중앙값, 중위수)가 로버스트한 통계량이다.
- 전체 데이터를 크기순으로 나열한다면 가운데에 위치한 값이다.
- 만약 전체 데이터의 개수가 짝수라면 전체데이터를 정열한 후 정확히 반으로 나눈후 하위값이 속한 곳에서 가장 큰 값과 상위값이 속한 곳에서 가장 작은 값의 평균을 중간값으로 정의한다.
- 특잇값은 극단적인 값을 가지는 데이터를 의미하며 특이값을 정하기 위해 boxplot을 이용할 수 있다.

변이 추정

분산과 편차

- 데이터의 변동을 나타내는 대표적인 요약치는 다음과 같다.
  - (표본)분산: $S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
  - (표본) 표준편차: 분산의 제곱근
  - 평균절대편차: $\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$
- 분산은 왜 제곱합을 *n*이 아닌 *n-1*로 나누어 주는가?

변이 추정

- 앞의 대푯값들은 특잇값에 대해 민감하다. 즉 로버스트 하지 않다.
- 변이를 나타내는 대표적인 로버스트 통계량으로 중간값의 중위절대편차(MAD)를 들 수 있다. m를 중간값이라고 하면 MAD는 다음과 같이 정의된다. MAD = 중간값$(|x_1 - m|, |x_2 - m|, \cdots, |x_n - m|)$

변이 추정

백분위수에 기초한 추정

- 변이를 추정하는 또 다른 방식은 정렬된 데이터를 사용하여 구하는 방식이다.

- *n*개의 데이터를 순서대로 정렬한 후 가장 작은 값을 $X_{(1)}$, 가장 큰 값을 $X_{(n)}$이라 하자. 이렇게 정렬된 데이터를 순서통계량이라고 한다.

- 범위 $= X_{(n)} - X_{(1)}$

- 데이터의 *p*번째 백분위수는 전체 데이터의 *p*퍼센트의 값이 그 값보다 작거나 같은 것을 의미한다.

- 1사분위수는 25번째 백분위수, 3사분위수는 75번째 백분위수를 말한다.

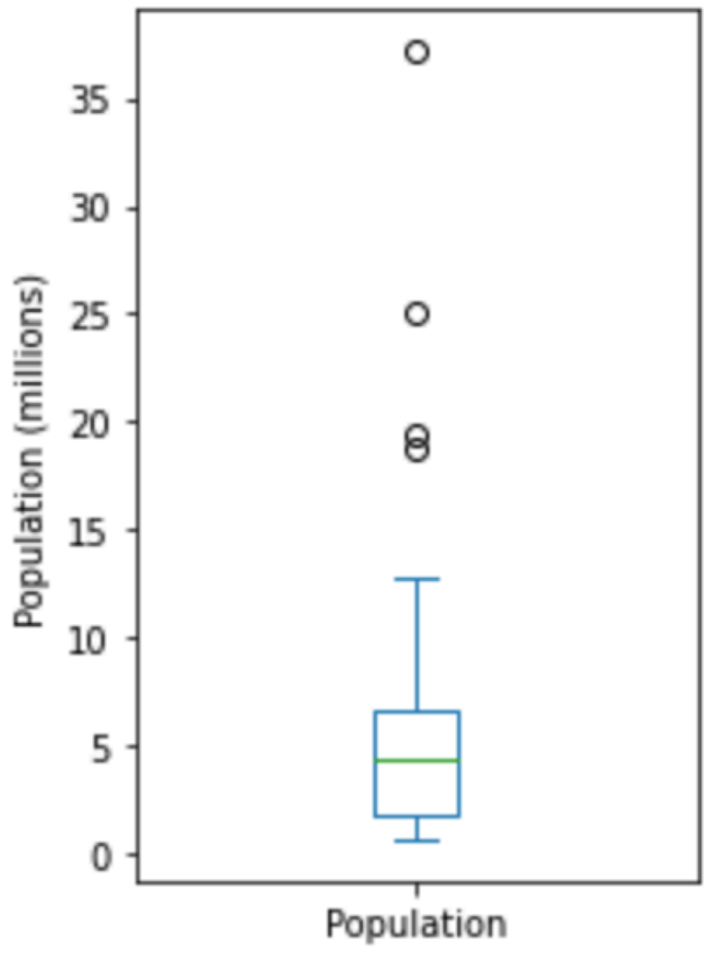
예제

- 3,1,5,3,6,7,2,9로 데이터가 주어졌을때 1사분위수, 3사분위수, 범위를 구하여라.

데이터 분포 탐색하기

백분위수와 상자그림

- Boxplot은 중간값, 1사분위수(Q_1), 3사분위수(Q_3), *IQR*($= Q_3 - Q_1$), 특이점등의 정보를 효율적으로 표현할 수 graphical summary이다.
- 왼쪽 그림은 주별 인구를 boxplot을 이용하여 보여주고 있다.
- 50개주의 인구 중앙값은 약 500만이라는 것을 알 수 있고 1사분위수가 약 200만, 3사분위수가 700만으로 보인다. 즉 25개정도의 주인구는 이 범위안에 있는 것이다.
- 구레나룻처럼 위아래로 나 있는 점선을 수염이라고 부르고 수염 바깥쪽에 있는 데이터를 특잇값으로 간주한다.
- 일반적으로 수염의 위치는 다음과 같다.
  $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$



데이터 분포 탐색하기

도수분포표와 히스토그램

- 도수분포표는 변수의 범위를 동일한 크기의 구간으로 나눈 후 각 구간마다 몇개의 변수값이 존재하는지 보여준다.

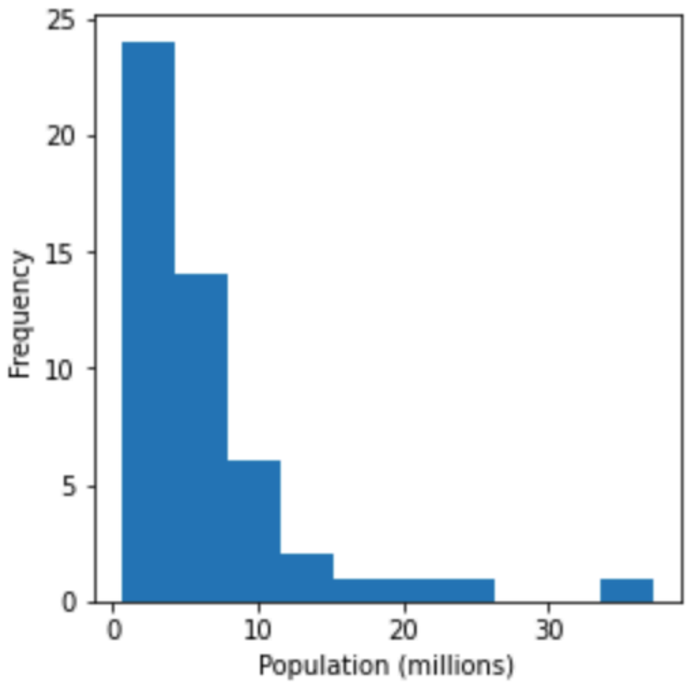
- 아래 그림은 주별인구를 총 10개의 구간으로 나눈 후 각 구간에 속하는 주의 개수를 보여준다. 보이지 않는 마지막 10번째 구간에는 CA가 속한다.

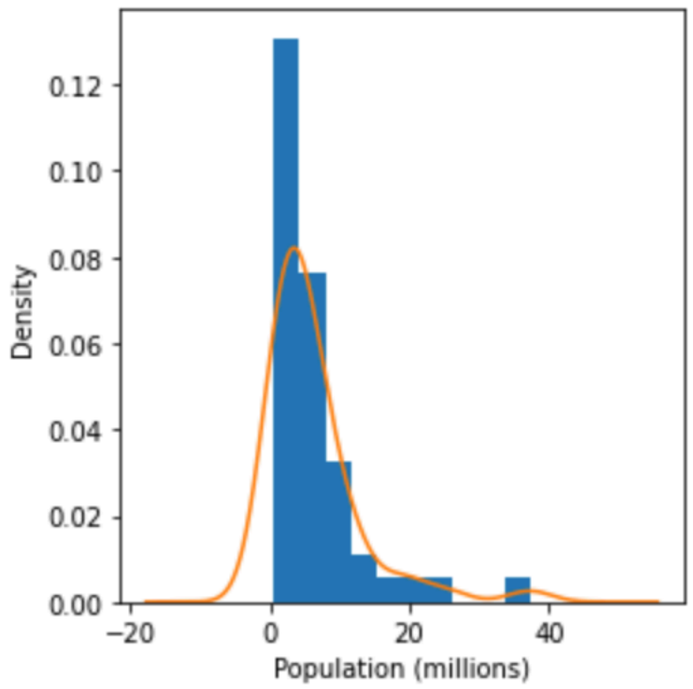
```
(526935.67, 4232659.0]      24
(4232659.0, 7901692.0]      14
(7901692.0, 11570725.0]      6
(11570725.0, 15239758.0]     2
(15239758.0, 18908791.0]     1
(18908791.0, 22577824.0]     1
(22577824.0, 26246857.0]     1
(33584923.0, 37253956.0]     1
(26246857.0, 29915890.0]     0
(29915890.0, 33584923.0]     0
Name: Population, dtype: int64
```

데이터 분포 탐색하기

밀도그림과 추정

- 도수분포표를 시각적으로 표현한 것으로 히스토그램과 밀도그림을 들 수 있다.

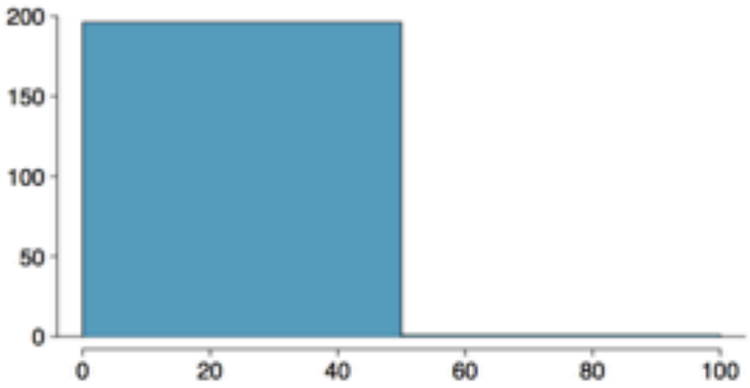


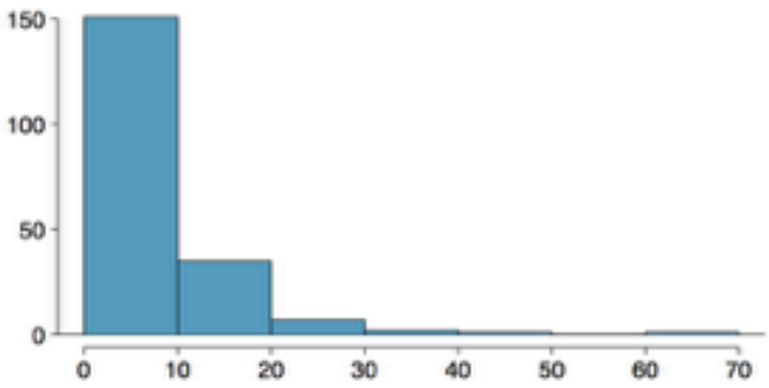


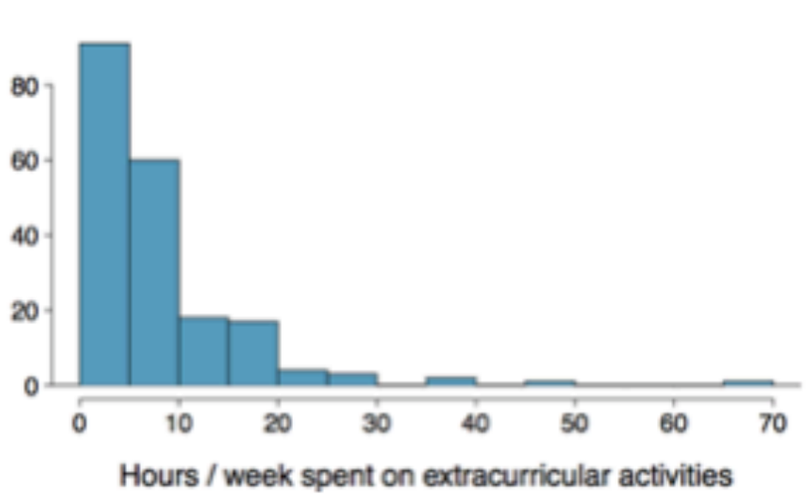
데이터 분포 탐색하기

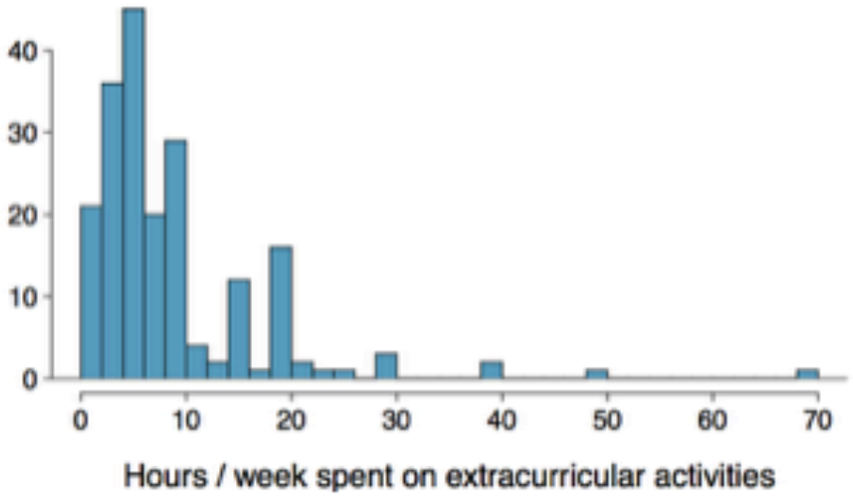
구간길이

- 아래 histogram중 어느 것이 자료를 가장 잘 표현하고 있는지 얘기해보자.





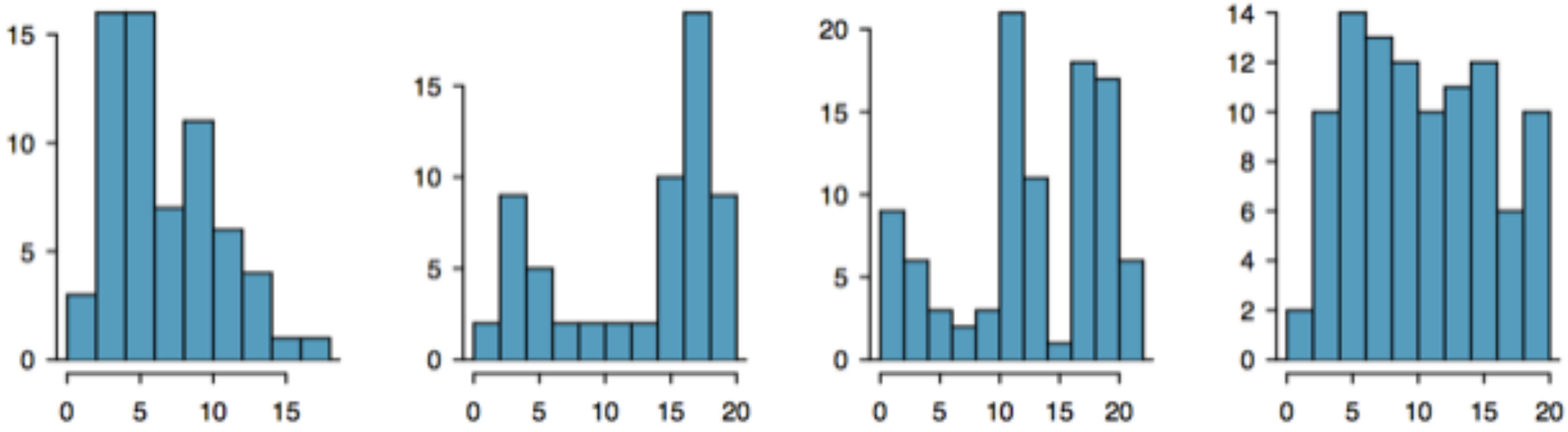




데이터 분포 탐색하기

Modality

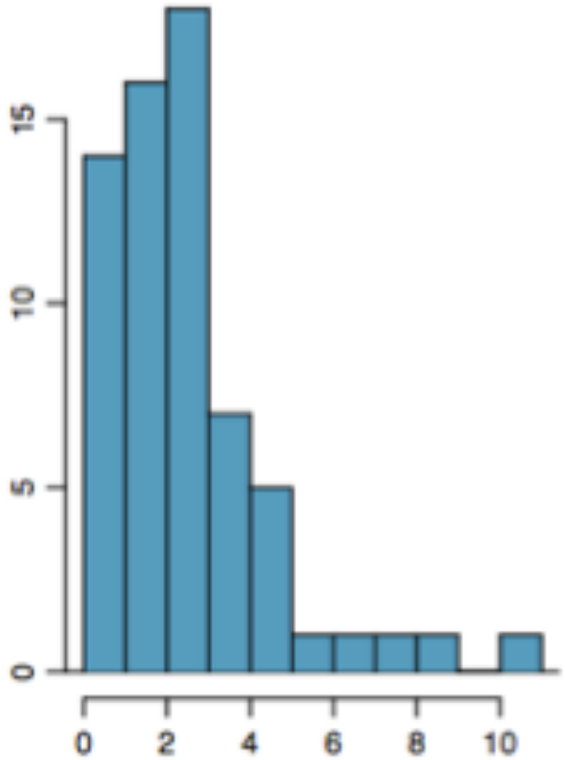
- Mode는 분포에서 peak가 있는 부분을 이야기한다. 아래그림에 unimodal, bimodal, multimodal, uniform인 분포는 어느것인지 얘기해 보자.

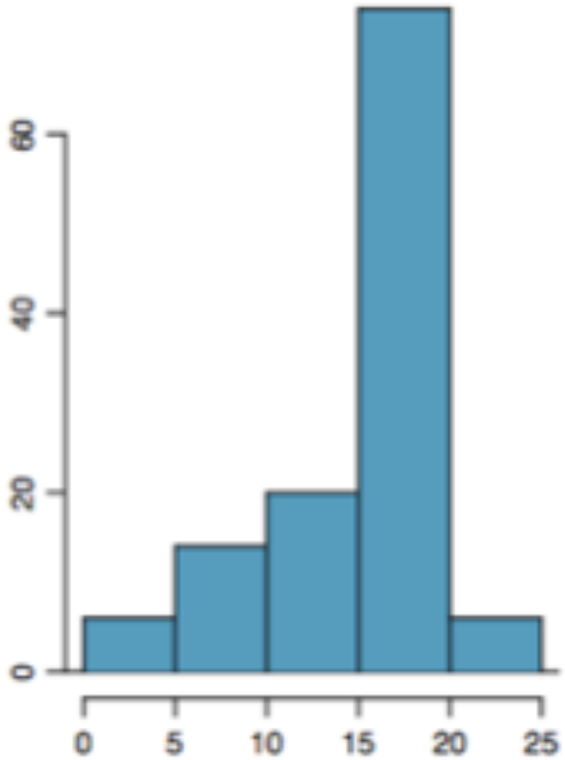


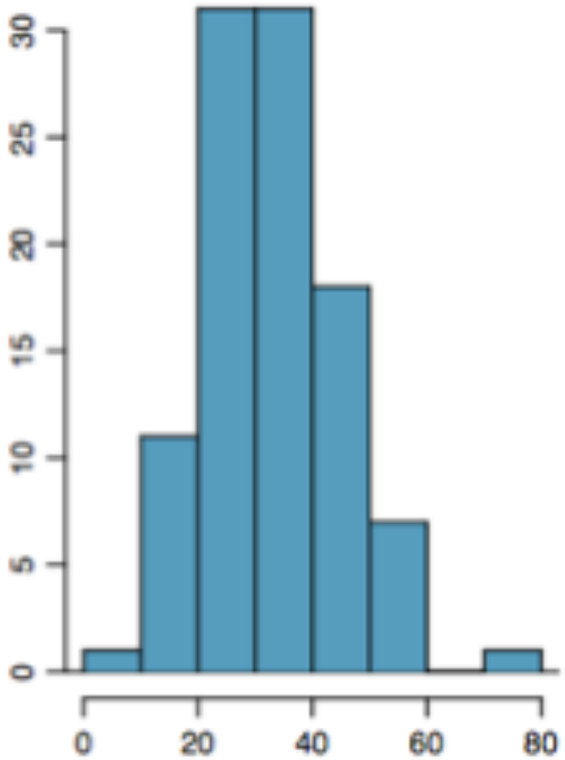
데이터 분포 탐색하기

분포의 형태

- 분포의 형태: right skewed, left skewed, symmetric





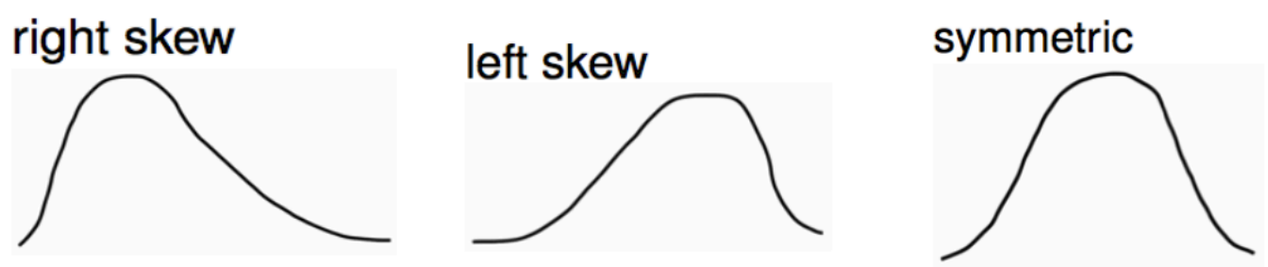


데이터 분포 탐색하기

일반적인 분포의 형태

Modality

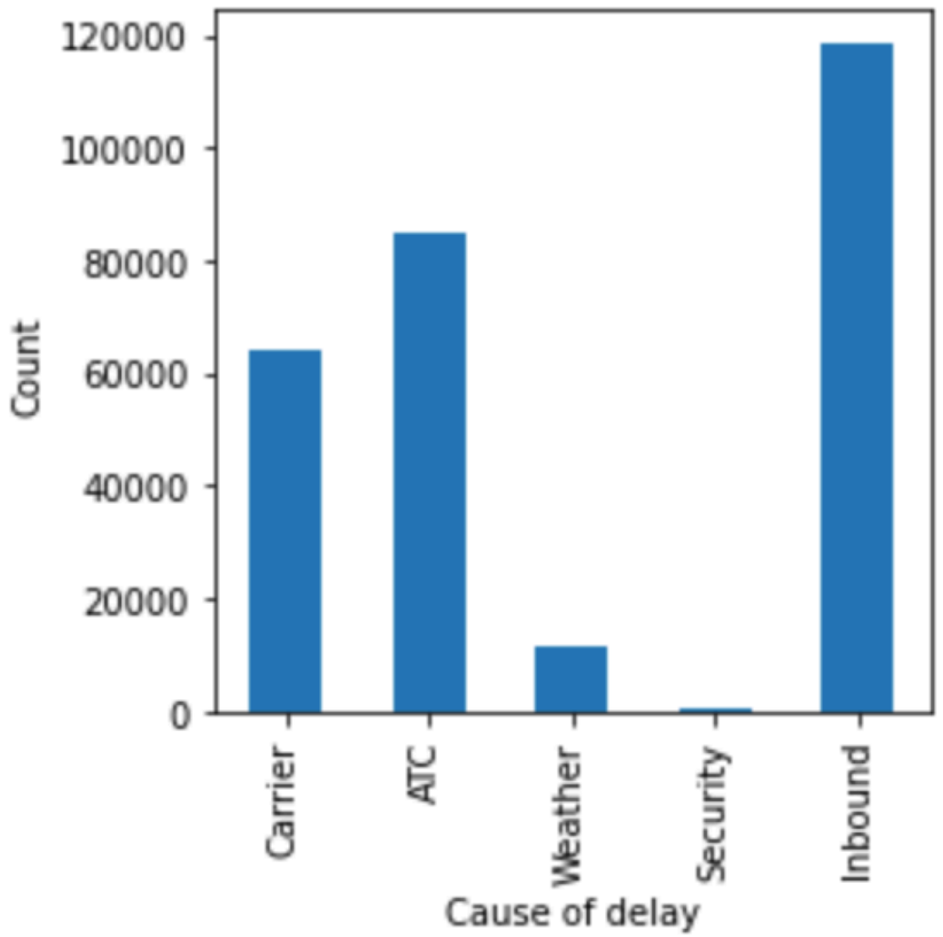
Skewness



이진 데이터와 범주 데이터 탐색하기

범주형 자료의 요약치와 시각화

- 범주형 자료를 대표하는 요약 통계량(summary statistics)으로 최빈값(mode)를 들 수 있다.
- 범주형 자료의 시각화는 막대도표를 사용한다.
- 파이차트는 어떤 경우라도 절대 사용해서는 안된다!

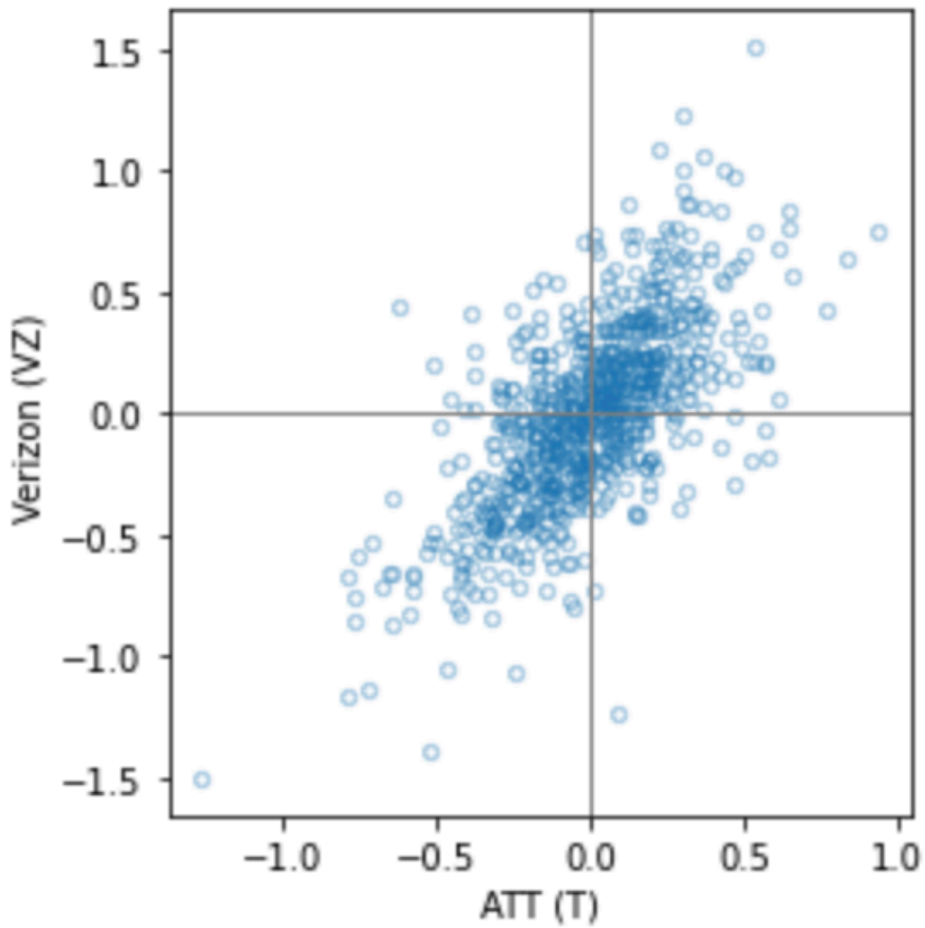


상관관계

산점도

- 두 연속형 변수의 관계를 시각화하는 가장 기본적인 방법은 산점도를 그리는 것이다.

- 왼쪽그림은 두 개 통신사 주식의 일간 수익간의 관계를 보여준다.



상관관계

상관계수

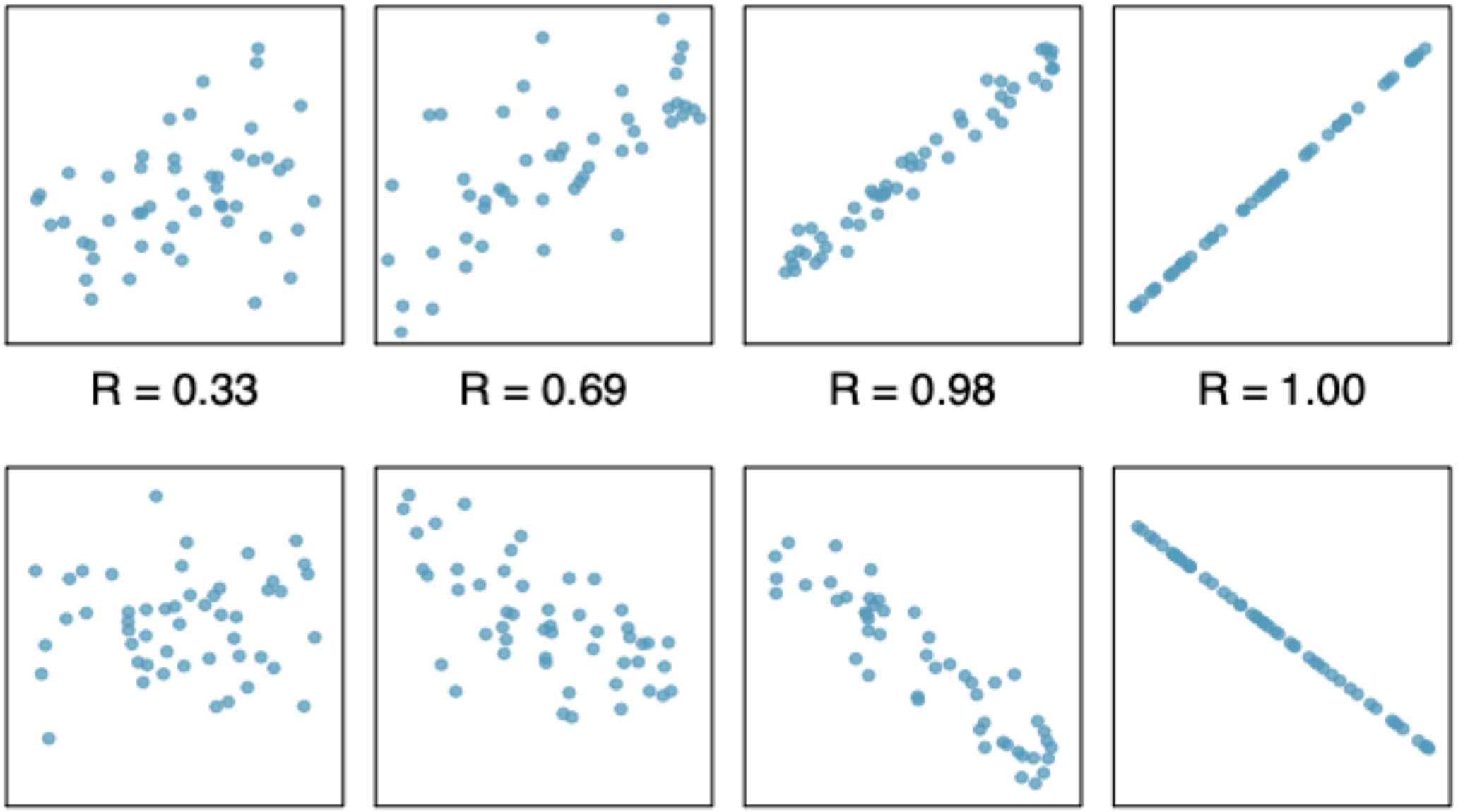
- 산점도와 더불어 두 변수사이의 관계를 하나의 요약 통계량으로 나타낼 수 있을까?
- 이 질문에 대답하기 위해 칼 피어슨은 피어슨 상관계수 (Pearson's correlation coefficient)를 제안하였다.
- 만약 우리가 n쌍의 데이터 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$를 관측한다면 피어슨 상관계수는 다음과 같이 정의된다.

$$R = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

- 피어슨 상관계수는 -1과 1 사이의 값을 나타내면 상관계수의 절대값이 1에 가까울수록 두 변수들이 강한 직선관계가 있음을 의미한다.

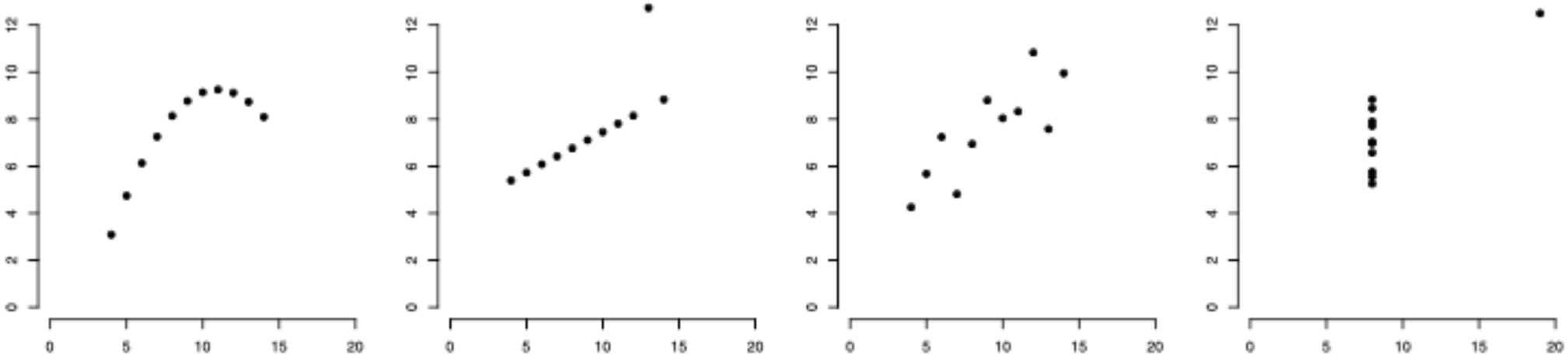
상관관계

상관계수

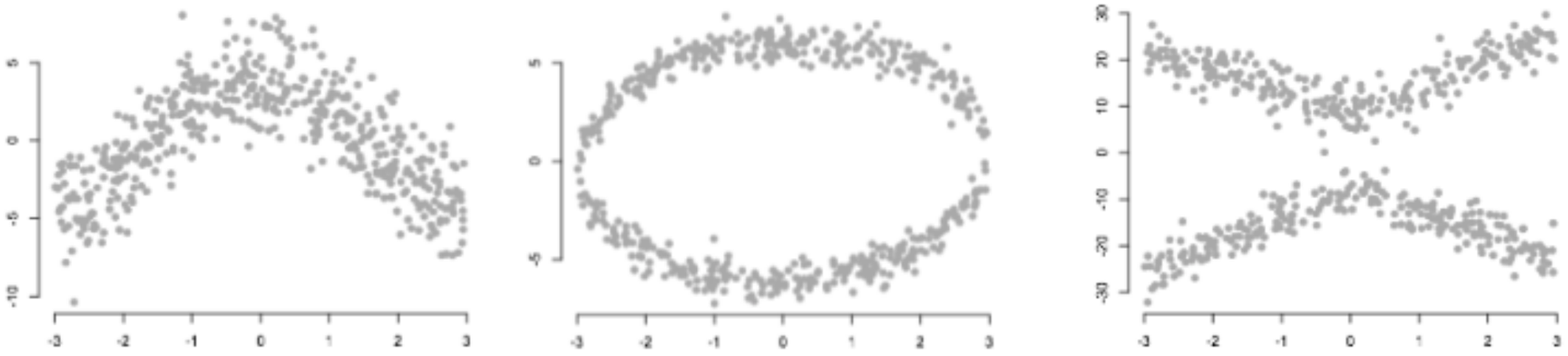


상관관계

상관계수



상관계수가 모두 0.816인 경우. “Graphs in Statistical Analysis.” American Statistician 27. p19-20

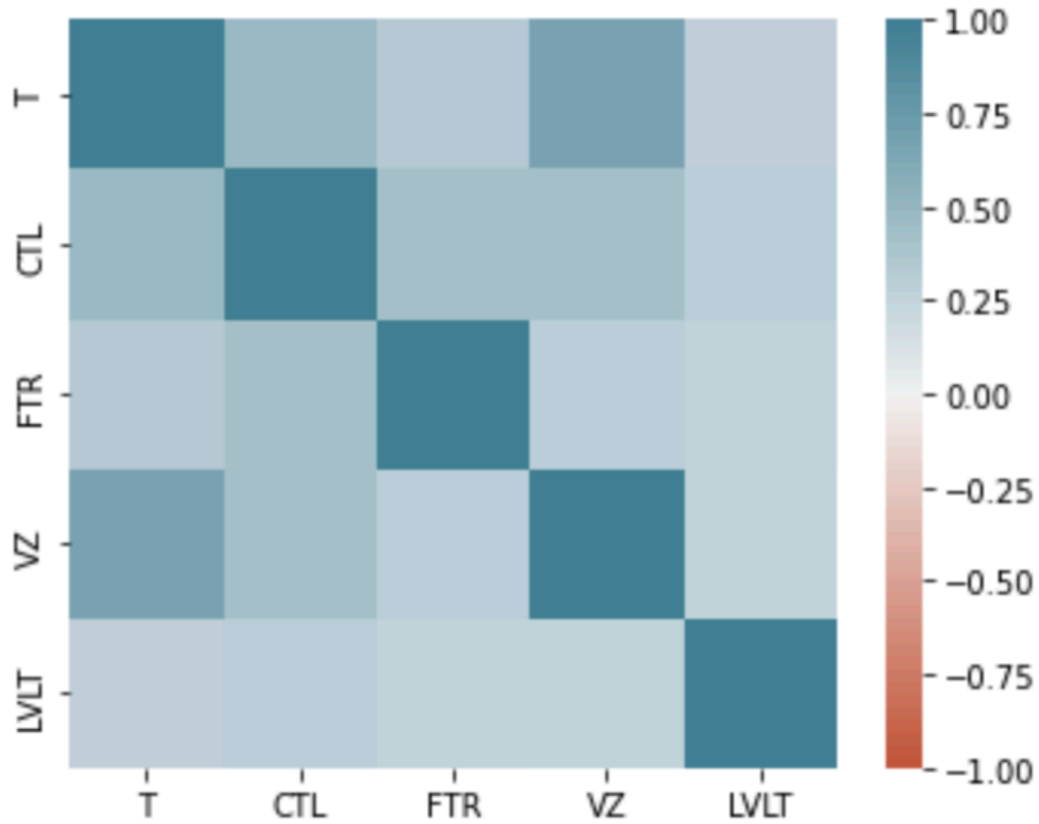


상관계수가 모두 0인 경우. Data Science. p26 https://bit.ly/34Fx73I

상관관계

상관계수

- 만약 변수가 여러개인 경우 상관관계를 비교하기 위해 heatmap을 사용하는 것을 고려할 수 있다.
- 오른쪽 그림은 여러 통신사주식의 일간 수익의 상관계수를 heatmap으로 표시한 그림이다.



두개 이상의 변수 탐색하기

- 2개 이상의 변수의 자료의 정리와 시각화를 위해서 다음과 같은 방법을 사용한다.
    - 육각형 구간과 등고선
    - 분할표 (범주형 vs 범주형)
    - Side-By-Side Boxplot (범주형 vs 연속형)
    - 다변수 시각화

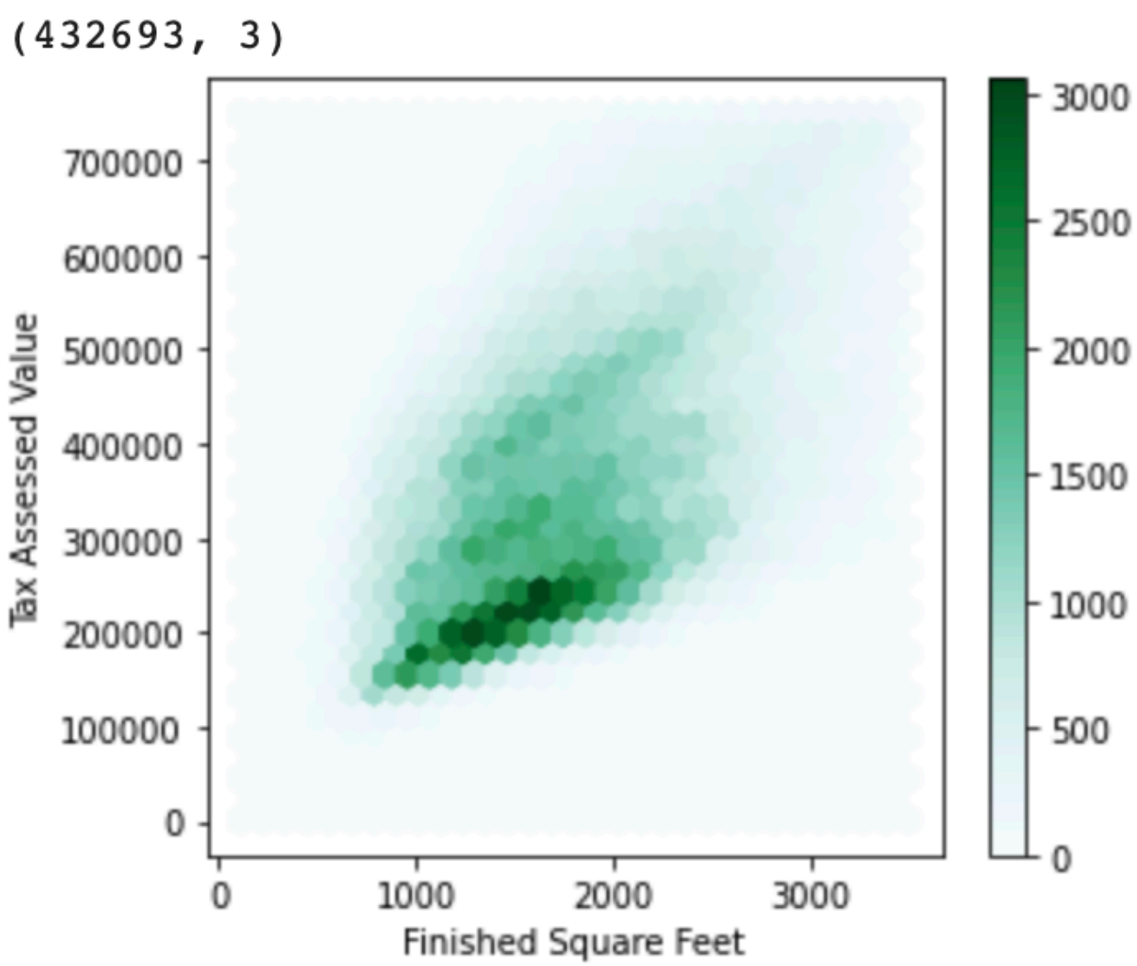
두개 이상의 변수 탐색하기

육각형 구간과 등고선

- 데이터가 겹쳐있는 경우 산점도를 사용한다면 정확한 분포를 알아볼 수 없다.

- 워싱턴 주의 킹 카운티의 주택세 기준 가격에 대해 시각화를 해보자.

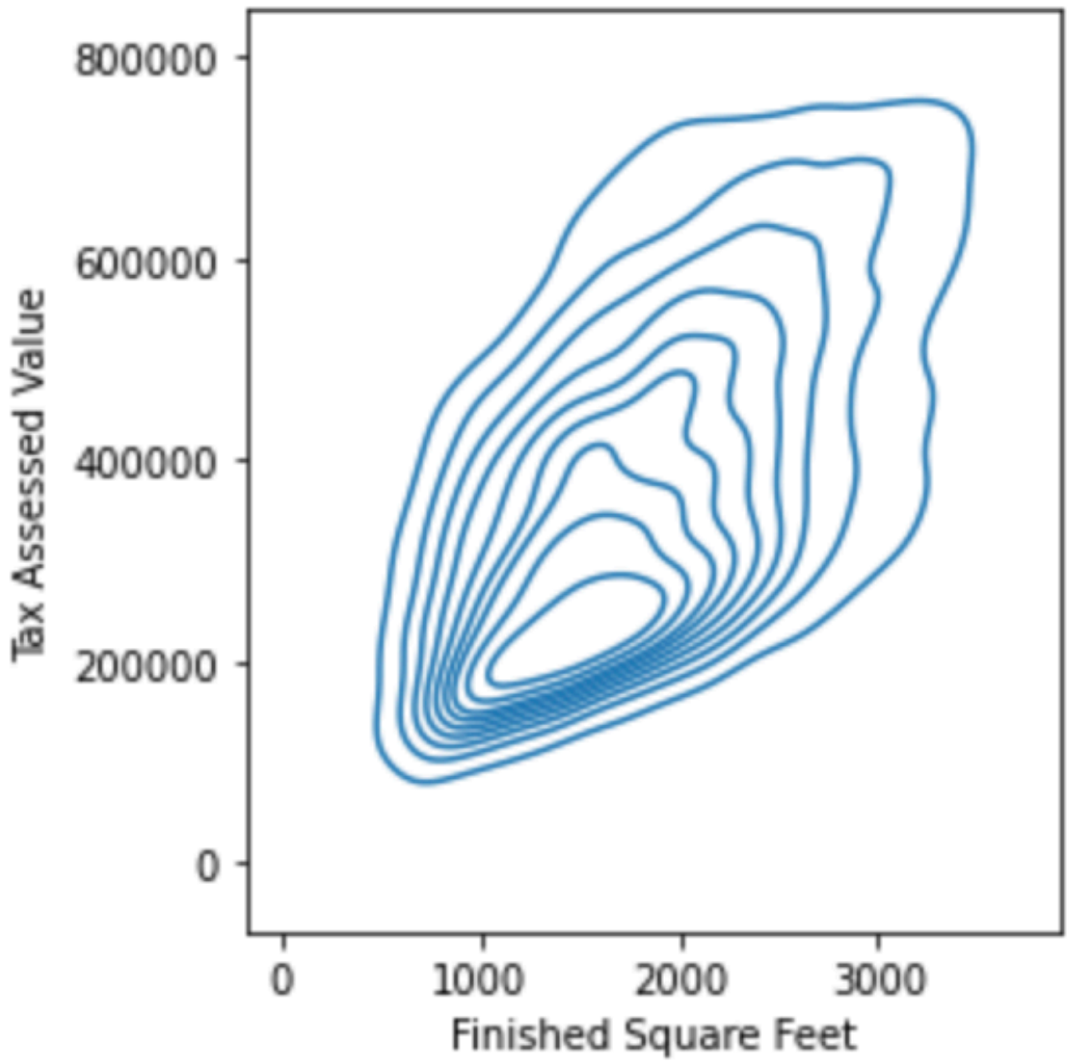
- 오른쪽 그림에서는 개별 데이터를 표시하는대신 전체 기록값의 범위를 육각형 모양 구간을 나눈 후 각 구간에 포함되는 데이터 개수를 색상을 이용하여 표현한다.



두개 이상의 변수 탐색하기

육각형 구간과 등고선

- 오른쪽 그림은 등고선을 보여주고 있다.
- 이차원상의 밀도를 잘 보여주고 있으며 1500sq 정도 크기에 20만불가격이 봉우리가 있음을 알 수 있다.



두개 이상의 변수 탐색하기

분할표

- 두개의 범주형 자료를 정리하는데 사용한다.

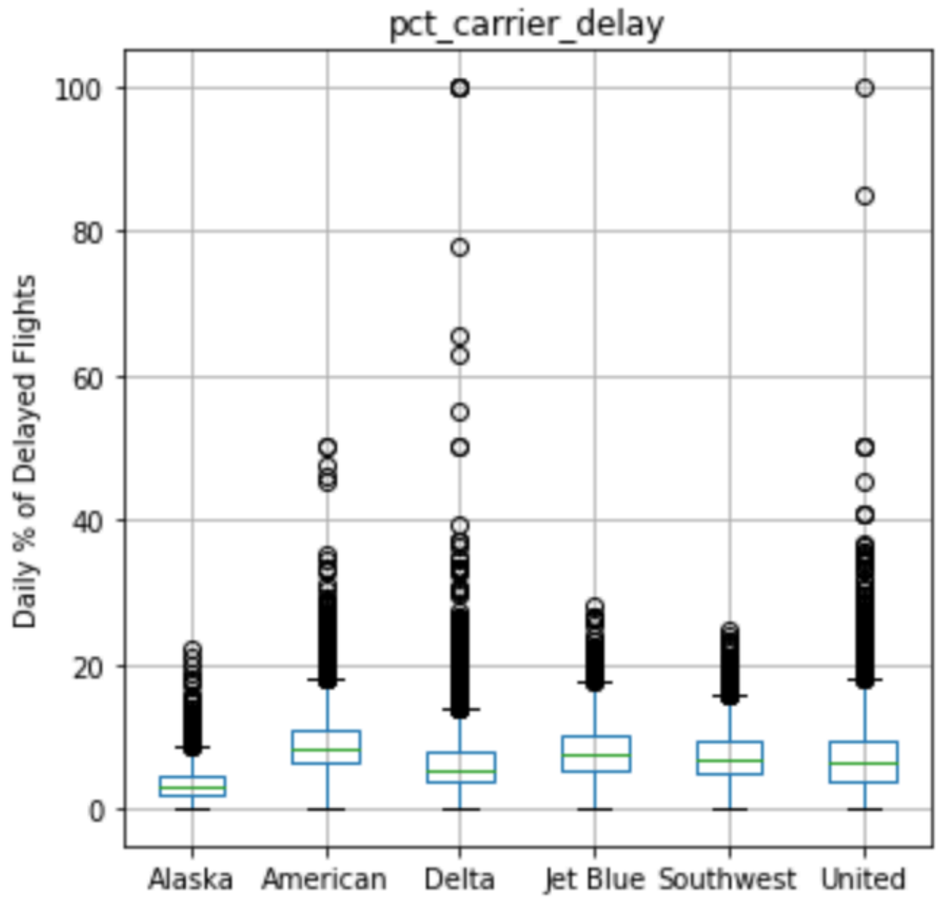
- 여자친구 팬클럽 버디 회원중 300명을 무작위로 추출하여 여자친구 멤버들에 대한 선호도 조사를 하였다. 회원들의 성별에 따라 멤버들의 선호도 차이가 있는지 여부를 알고 싶다. 선호도 조사 결과는 다음과 같다.

성별	소원	예린	은하	유주	신비	엄지	합계
남	20	33	30	22	22	23	150
여	19	34	27	22	32	16	150
합계	39	67	57	44	54	39	150

두개 이상의 변수 탐색하기

다변수 시각화하기

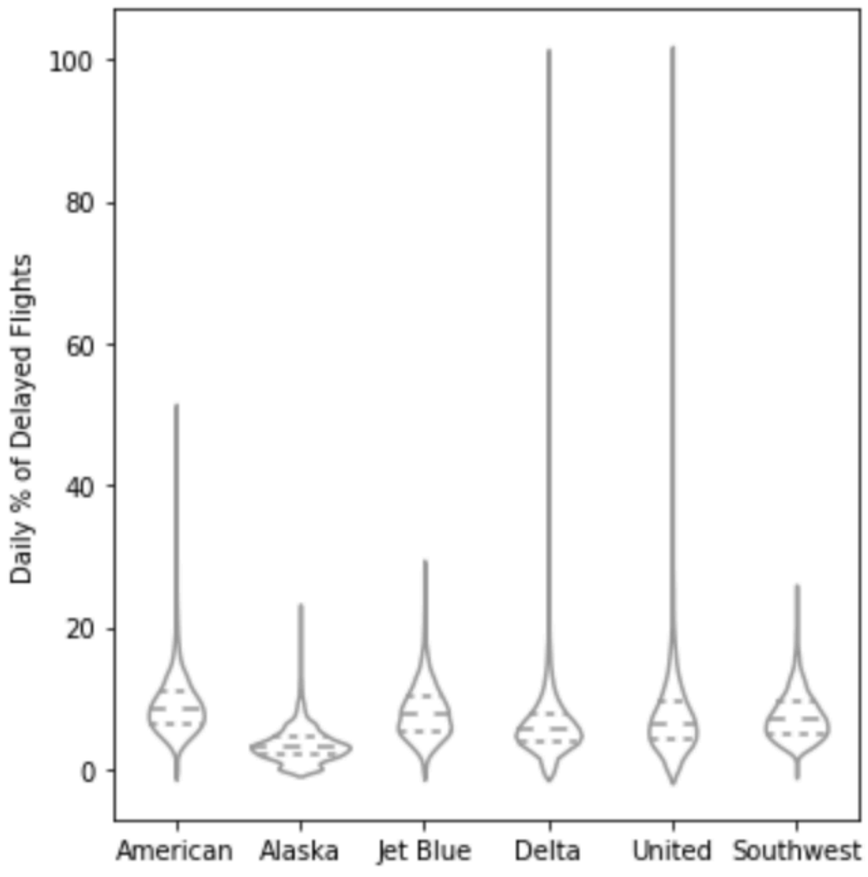
- 항공사별 비행기 지연 정도를 비교하고자 한다.
- 왼쪽 그림은 side-by-side boxplot을 보여준다.
- 알래스카 항공이 지연이 제일 적은 것으로 보인다.



두개 이상의 변수 탐색하기

다변수 시각화하기

- Boxplot에서는 개별데이터의 분포를 보기 어렵다.
- 이 점을 보완하기 위해 오른쪽 바이올린 도표가 제안되었다.
- y축의 데이터를 밀도분포 그림으로 나타낸후 좌우대칭이 되도록 표시한다
- 이 경우 특이점은 파악하기 힘들다.



두개 이상의 변수 탐색하기

다변수 시각화하기

- 2개이상의 변수에 관한 시각화를 할 경우 생각해 보자.
- 주택세 기준세금액자료의 경우 지역별 분석을 위해 우편번호를 새로운 변수로 추가했다면 우편번호별로 육각형 구간을 그릴 수 있다.
- 지역별로 집의 크기와 가격이 차이가 나는 것을 볼 수 있다.