



Chapter 01. EDA (Exploratory Data Analysis) - 정소진

Chapter01. Exploratory Data Analysis

- 고전적 통계학 : 추론 *inference*, 즉 적은 표본(샘플)을 가지고 더 큰 모집단에 대한 결론을 도출하기 위한 일련에 복잡한 과정에 관해 다루었음
- 존 투키 : 통계학의 개혁을 요구. 통계적 추론을 하나의 구성요소로 보는 데이터 분석이라는 새로운 과학적 학문 제언. 탐색적 데이터 분석은 투키의 책을 통해 정립되었음. 요약 통계량(평균, 중앙값, 분위수 등)과 함께 데이터 집합을 그림으로 표현하는 데 도움이 되는 간단한 도표(box plot, scatter plot 등)를 제시.

1.1 정형화된 데이터의 요소

- 통계적 개념들을 활용하기 위해서는 정형화되지 않은 폭발적인 양의 raw data를 활용 가능한 형태 정보(정형화된 형태)로 변환해야 한다.
 - 정형 데이터 중 가장 일반적인 형태는 행과 열이 있는 테이블 형태 ex) 관계형 데이터베이스, 연구용으로 수집한 데이터
- 수치형 *numeric* 데이터
 - 연속형 *continuous* 데이터 : 풍속이나 지속 시간
 - 이산 *discrete* 데이터 : 사건의 발생 빈도 ex) 제품 판매량, 불량품 개수
- 범주형 *categorical* 데이터 : 범위가 정해진 값들을 갖는 경우 ex) TV 스크린 종류, 도시명
 - 이진 *binary* 데이터 : 0과 1, 예/아니오, 참/거짓
 - 순서형 *ordinal* 데이터 : 수치로 나타낼 수 있는 평점 ex) 1, 2, 3, 4, 5
- 데이터를 분석하고 예측을 모델링할 때, 시각화, 해석, 통계 모델 결정 등에 데이터 종류가 중요한 역할을 한다.
- 또한 소프트웨어는 변수의 종류에 따라 해당 변수에 관련된 계산을 어떤 식으로 수행할지 결정한다.
- 데이터 구분의 이점

- 소프트웨어가 차트 생성이나 모델 피팅 등 통계분석을 수행하는 방식을 결정하는데 도움을 준다.
- 관계형 데이터베이스에서처럼 저장소와 인덱싱을 최적화하는 데 사용한다.
- 범주형 변수가 취할 수 있는 값들은 소프트웨어적으로 처리가 가능하다. (enum 처럼)
 - enum은 관련있는 상수들의 집합. 상수 클래스라고 생각하며 될 듯. 고정되어 있는 값(문자열, 숫자)을 enum으로 선언해서 사용. enum을 사용하지 않은 상수값들은 사용할 때 마다 메모리를 차지하는데, enum으로 선언해 놓으면 어플리케이션이 실행될 때 enum 상수값들이 미리 할당되어 계속 그 메모리 주소에 있는 상수를 사용하게 된다. 따라서 메모리에 유리.

1.2 테이블 데이터 *rectangular data*

Data Frame : 통계와 머신러닝 모델에서 가장 기본이되는 테이블 형태의 데이터 구조
 Feature : 일반적으로 테이블의 각 열이 하나의 피처를 의미. (= 특징, 속성, 열, column, 입력, 예측변수, 변수)
 Record : 일반적으로 테이블의 각 행은 하나의 레코드를 의미 (= 사건, 행, row, 사례, 예제, 관측값, 패턴, 샘플)
 Outcome 결과 : 데이터 과학 프로젝트의 목표는 대부분 어떤 결과를 예측하는 데 있음. 실험이나 연구에서 결과를 예측하기 위해 피처를 사용. (= 종속변수, 응답, 목표, 출력)

- 테이블 데이터: 각 레코드를 나타내는 행과 피처를 나타내는 열로 이루어진 이차원 행렬을 의미
 - 문자열과 같은 비정형 데이터는 테이블 데이터의 피처 형태로 표현되도록 처리해야 한다.
 - 관계형 데이터베이스에 있는 데이터를 불러올 때도 역시 하나의 테이블 형태로 변환해야 한다.
 - 경매에 경쟁이 있는지 없는지를 나타내는 이진변수가 있을 경우, 이 지표변수 *indicator variable*도 결과변수 *outcome variable*가 될 수 있다.

1.2.1 데이터 프레임과 인덱스

- 하나 혹은 그 이상의 열을 인덱스로 지정 => 데이터베이스의 쿼리 성능을 크게 향상할 수 있다.
- 용어 차이

통계학자응답변수(response variable)나 종속변수(dependent variable)를 예측하데 예측변수(predictor variable)를 사용데이터 과학자목표(target)를 예측하는데 피쳐(feature)를 사용샘플컴퓨터 과학자 : 보통 각각의 행을 칭함통계학자 : 여러 행의 집합

1.2.2 테이블 형식이 아닌 데이터 구조

- 시계열 데이터 : 동일한 변수 안에 연속적인 측정값을 갖는다.
 - 통계적 예측 기법들을 위한 원재료
 - 사물 인터넷과 같이 다양한 디바이스에서 생산되는 데이터들에서 중요한 요소
- 공간 데이터 : 지도 제작과 위치 정보 분석에 사용됨. 테이블 데이터보다 좀 더 복잡하고 다양
 - 객체 *object* 를 표현할 때 : 어떤 객체(ex. 주택)와 그것의 공간 좌표
 - 필드 *field* 정보를 표현할 때 : 공간을 나타내는 작은 단위들과 적당한 측정 기준값(ex. 픽셀의 밝기)에 중점
- 그래프(혹은 네트워크) 데이터 : 물리적 관계, 사회적 관계, 추상적인 관계들을 표현하기 위해 사용됨.
 - 페이스북, 링크드인과 같은 소셜 네트워크에서의 그래프
 - 도로로 연결된 물류 중심지를 나타낸 그래프
 - 네트워크 최적화나 추천 시스템에 유용
- 데이터 종류들은 각자 맞는 아주 특화된 방법론을 가짐. 하지만 이 책에서는 예측 모델의 기본이 되는 테이블 데이터에 대해 주로 다룸.

1.3 위치 추정

- 데이터를 살펴보는 가장 기초적 단계는 각 피쳐(변수)의 대표값 *typical value*를 구하는 것.
 - 중심경향성(대부분의 값이 어디쯤에 위치하는지)를 나타내는 추정값.

1.3.1 평균

- 평균: 모든 값의 총합을 값의 개수로 나눈 값

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

N: 모집단 n: 모집단에서 얻은 표본의 개수 통계에서의 구분. 데이터과학에서는 N, n은 같은 의미 (레코드나 관측값의 개수)

- 절사평균 *trimmed mean* : 정해진 개수의 극단값 *extreme value*을 제외한 나머지 값들의 평균 (= 절단평균 *truncated mean*)
 - 극단값의 영향을 제거

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

- 가중평균 *weighted mean* : 가중치를 곱한 값의 총합을 가중치의 총합으로 나눈 값
 - 어떤 값들이 본래 다른 값들에 비해 큰 변화량을 가질때, 이런 관측값에 대해 더 작은 가중치를 줄 수 있음.
 - 데이터를 수집할 때, 관심있는 서로 다른 대조군에 대해서 항상 똑같은 수가 얻어지지 않음. 이를 보정하기 위해 데이터가 부족한 소수 그룹에 더 높은 가중치를 적용하기도 함.

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

1.3.2 중간값과 로버스트 추정

- 중간값 *median* : 데이터에서 가장 가운데 위치한 값 (= 50번째 백분위수)
 - 데이터의 개수가 짝수라면, 실제 데이터 값이 아닌 가운데 있는 두값의 평균으로 한다.
 - 많은 경우, 데이터에 매우 민감한 평균보다는 중간값이 위치 추정에 더 유리하다.
- 가중중간값 : 가중 평균을 사용하는 이유와 동일. 특잇값에 로버스트하다.
 - 로버스트하다 *robust* : 극단값들에 민감하지 않다. (= 저항성 있다 *resistant*)
- 특잇값 *outlier* : 대부분의 값과 매우 다른 데이터 값 (= 극단값)
 - 중간값은 특잇값에 영향 받지 않으므로 로버스트한 위치추정 방법
 - 특잇값은 데이터 값 자체가 유효하지 않다거나 잘못되었다는 뜻이 아님.
 - 이상 검출 *anomaly detection*: 대부분 정상적인 데이터보다 예외적으로 측정된 특이값이 주 관심 대상이 된다.
 - 절사평균 역시 특잇값의 영향을 줄이기 위해 많이 사용된다. 중간값과 평균의 절충안이라고 볼 수 있다.
- 백분위수 *percentile* : 전체 데이터의 P%를 아래에 두는 값 (= 분위수)

1.4 변이 추정

- 변이 *variability* : 데이터 값이 얼마나 밀집해 있는지 혹은 퍼져 있는지를 나타내는 산포도 *dispersion*
 - 위치추정에 다양한 방법이 있는 것처럼, 변이추정에도 다양한 방법이 있음.

1.4.1 표준편차와 관련 추정값들

- 편차 *deviation* : 관측값과 위치 추정값 사이의 차이 (= 오차, 잔차)
 - 데이터가 중앙값을 주변으로 얼마나 퍼져있는지를 말해줌
 - 변이측정 방법 1) 편차들의 대푯값을 추정
 - 평균을 기준으로 편차의 합은 항상 0
 - 절댓값을 취해 음의 편차가 양의 편차를 상쇄하는 것을 막음
- 평균절대편차 *mean absolute deviation* : 평균과 편차의 절댓값의 평균 (= L1 Norm, 맨해튼 노름)

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

- 분산 *variance* : 평균과 편차를 제공한 값들의 합을 n-1로 나눈 값. n은 데이터 개수 (= 평균제곱오차)
 - 모분산 : 모집단의 분산
 - 표본분산 : 표본집단의 분산

모평균과 모분산, 표본평균과 표본분산: 모평균과 모분산은 모든 정보가 공개된 상태에서 평균과 분산을 구했을 경우. 표본평균과 표본분산은 전체 정보 중에 극히 일부만 공개되어있고, 공개가 된 일부 정보를 가지고 값을 구한 후, 다시 전체 모르는 값을 추정하겠다는 것: 모집단 대상, 표본집단 대상 분산, 평균을 구하는 방식은 같을 수 없음.: 일부 공개된 정보(표본집단)를 가지고 편차제곱합을 구했을 경우,

전체 정보(모집단)를 가지고 가지고 편차제곱합을 구했을 경우보다 작게 계산되는 경우가 있다. 그래서 n-1로 나눠줌으로써 분산값을 올려줘야하는 경우가 생긴다..!- 그렇다면 왜 표본분산은 모분산 보다 값이 작게 추정될까?:

표본분산을 구할 때, 평균에서 표본값을 빼는 과정(편차를 구할 때)에서 표본평균은 표본값들을 잘 표현하기 때문에 (전체 정보 중 일부를 이용해 표본평균을 구하기 때문에) 편차가 모분산을 구할 때에 비교해서 작을 수 밖에 없다.=> 분산 수식에 n을 분모로 사용하게 된다면, 모집단의 분산과 표준편차의 참 값을 과소평가하게 된다. 이를 편향 *biased* 추정이라고 부른다. 하지만 만약 n 대신 n-1로 나눈다면 이 분산은 비편향 *unbiased* 추정이다

된다.

- 왜 하필이면 n-1인가? => 이해 완벽히 x => 링크 다시 한번 보기: 표준편차는 표본의 평균에 따른다는 하나의 제약 조건을 가지고 있기 때문에 n-1의 자유도를 갖는다.

- 자유도?: (계산식의 미지수 수) - (계산식의 추정치 수) 라고 이해하기

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

< 모분산 >

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

< 표본분산 >

- 표준편차 *standard deviation* : 분산의 제곱근
 - 원래 데이터와 같은 척도에 있기 때문에 분산보다 해석이 쉽다.
- 중간값의 중위절대편차 (MAD) *median absolute deviation from the median* : 중간값과의 편차의 절댓값의 중간값 => 중앙값 절대편차
 - 분산, 표준편차, 평균절대편차 모두 특잇값과 극단값에 로버스트하지 않다. 분산과 표준편차는 제곱편차를 사용하기 때문에 특히 특잇값에 민감하다.
 - 중간값의 중위절대편차는 로버스트한 변이 추정값
 - 아래 식에서 m은 데이터의 중간값을 의미한다.
 - 관측값에서 중앙값들을 뺀 값들의 중앙값을 구한다.

$$\text{중위절대편차} = \text{중간값}(|x_1 - m|, |x_2 - m|, \dots, |x_N - m|)$$

1.4.2 백분위수에 기초한 추정

: 정렬된 데이터가 얼마나 퍼져 있는지를 보는 것

- 순서통계량 *order statistics* : 최소에서 최대까지 정렬된 데이터 값에 따른 계량형 (= 순위)

- 정렬(순위) 데이터를 나타내는 통계량
- 여기서 기본이 되는 측도는 범위 *range* : 데이터의 최댓값과 최솟값의 차이
 - 최대, 최소값 자체가 특잇값을 분석하는데 큰 도움을 준다.
 - 하지만 특잇값에 매우 민감하고 데이터의 변이를 측정하는 데 그렇게 유용하지는 않다.
- 백분위수 *percentile* : P번째 백분위수는 어떤 값들의 P퍼센트가 이 값 혹은 더 작은 값을 갖고, (100-P)퍼센트가 이 값 혹은 더 큰 값을 갖도록 하는 값(= 분위수 *quantile* => 0.8분위수 = 80번째 백분위수)
- 사분위범위 (IQR) *interquartile range* : 75번째 백분위수와 25번째 백분위수 사이의 차이
 - {1, 2, 3, 3, 5, 6, 7, 9} 일때, 25번째 백분위수는 2.5, 75번째 백분위수는 6.5
 - 따라서 사분위범위는 $6.5 - 2.5 = 4$ 이다.
- 머신러닝과 통계 소프트웨어에서는 백분위수의 근삿값을 사용

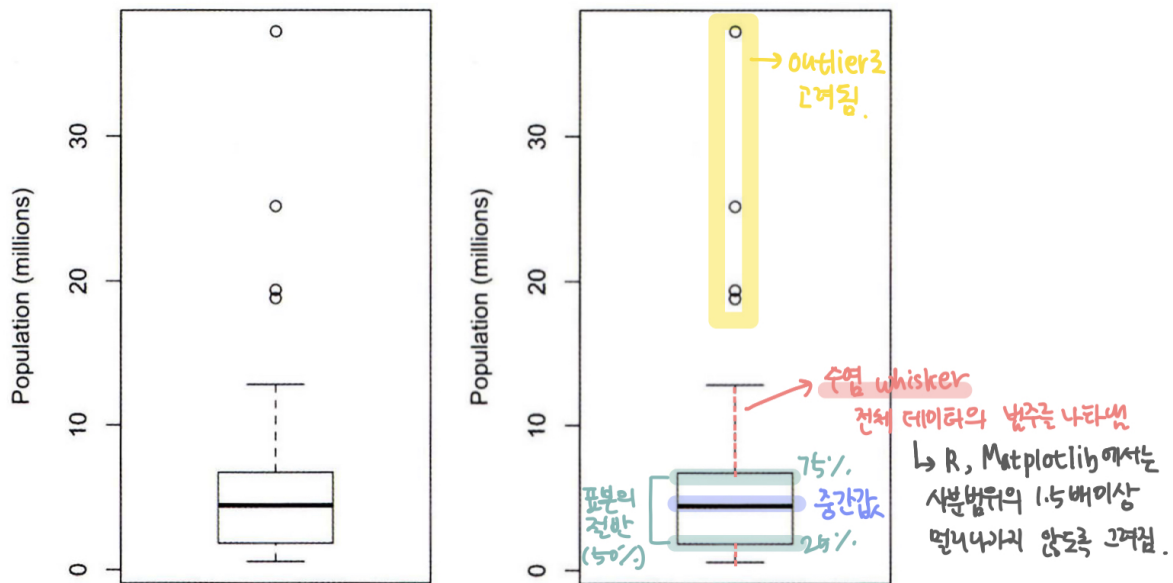
1.5. 데이터 분포 탐색하기

: 데이터가 전체적으로 어떻게 분포하는지 알아보는 것도 유용

1.5.1 백분위수와 boxplot

전체 분포를 알아보는 데에도 백분위수가 유용. 주로 사분위수 *quartile* (25, 50, 75번째 백분위수)이나 십분위수 *decile* (10, 20, ..., 90번째 백분위수)를 공식적으로 사용. 특히 백분위수는 분포의 꼬리 *tail* 부분을 묘사하는데 제격.


- 상자그림 *boxplot* : 투키가 데이터의 분포를 시각화하기 위한 간단한 방법으로 소개한 그림. 백분위수를 이용해 데이터의 분산을 손쉽게 시각화하는 방법.



1.5.2 도수분포표와 히스토그램

- 도수분포표 *frequency table* : 어떤 구간 *interval* (빈 *bin*)에 해당하는 수치 데이터 값들의 빈도를 나타내는 기록
 - 변수의 범위를 동일한 크기의 구간으로 나눈 다음, 각 구간마다 몇개의 변수값이 존재하는지를 보여주기 위해 사용된다.
 - 구간의 크기를 바꿔보며 유용한 정보를 얻을 수 있다.
 - 구간의 크기가 너무 크면, 분포를 나타내는 중요한 특징을 놓칠 수 있다.
 - 구간의 크기가 너무 작아도 결과가 너무 쪼개져있어 큰 그림을 볼 수 없게 된다.

표 1-5 주별 인구 도수분포표

BinNumber	BinRange	Count	States
1	563,626-4,232,658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE, WV,NM,NV,UT,KS,AR,MS,IA,CT,OK,OR
2	4,232,659-7,901,691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA, WA
3	7,901,692-11,570,724	6	VA,NJ,NC,GA,MI,OH
4	11,570,725-15,239,757	2	PA,IL
5	15,239,758-18,908,790	1	FL
6	18,908,791-22,577,823	1	NY
7	22,577,824-26,246,856	1	TX
8	26,246,857-29,915,889	0	 → 빈칸이 존재! ⇒ 중요한 정보를 줌.
9	29,915,890-33,584,922	0	
10	33,584,923-37,253,956	1	CA

- 히스토그램 *histogram* : x축은 구간들을, y축은 빈도수를 나타내는 도수 테이블의 그림. 막대그래프와 시각적으로 비슷하지만 헛갈리면 안됨
 - 도수분포표를 시각화한 방법
 - 히스토그램이 담고있는 정보
 - 그래프에 빈 구간들이 있을 수 있다.
 - 구간은 동일한 크기를 갖는다.
 - 구간의 수(혹은 구간의 크기)는 사용자가 결정할 수 있다.
 - 빈 구간이 있지 않은 이상, 막대 사이는 공간없이 서로 붙어 있다.

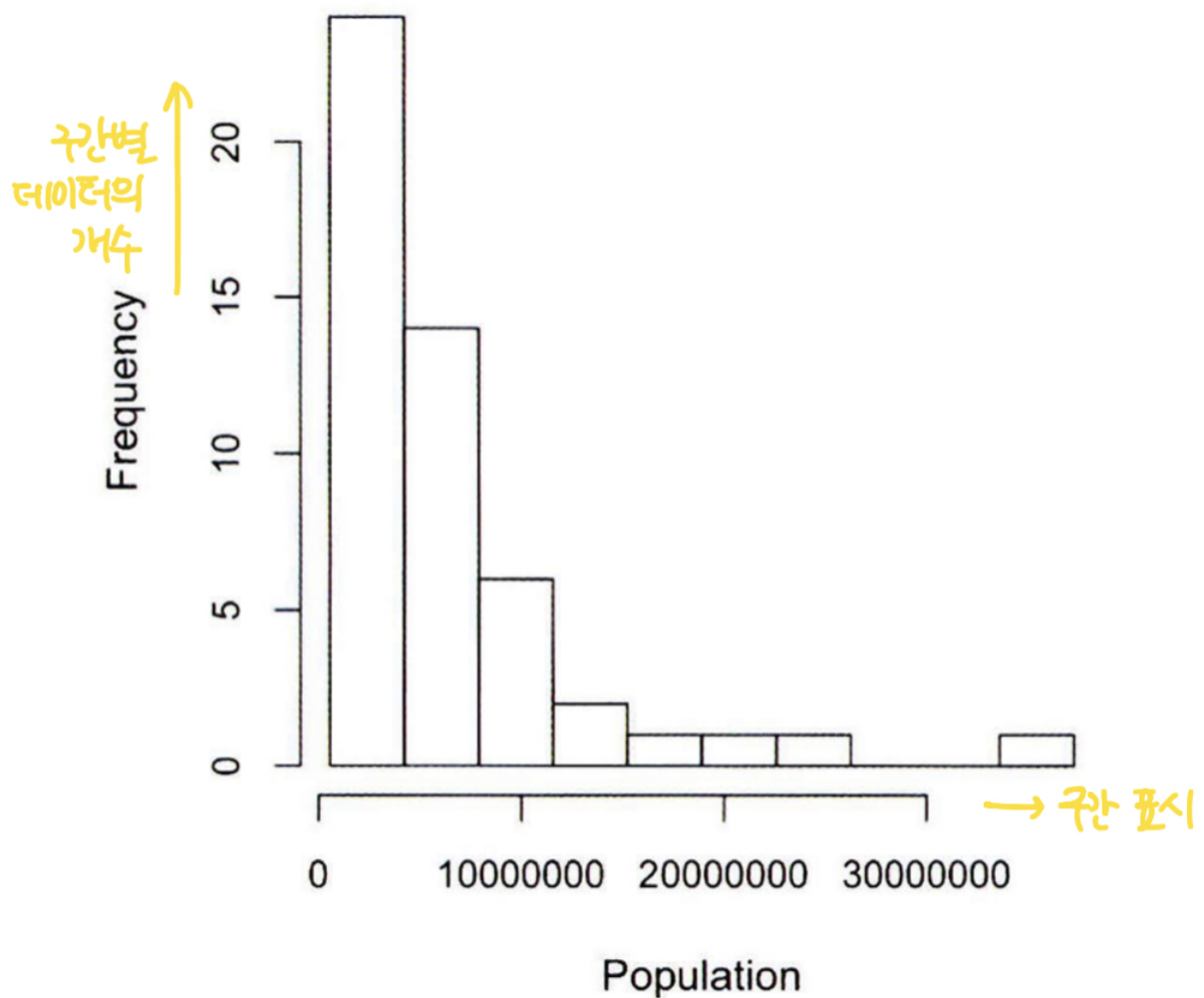


그림 1-3 주별 인구의 히스토그램

- 통계학에서의 모멘트 *moment* (혹은 적률)
 - 왜도 *skewness* : 삼차 모멘트. 데이터가 큰 값이나 작은 값으로 얼마나 비스듬히 쏠려 있는지를 나타냄
 - 첨도 *kurtosis* : 사차 모멘트. 데이터가 극단 값을 가는 경향성.
 - 박스 플롯 또는 히스토그램과 같이 시각화하여 모멘트 값들을 확인할 수 있다.

1.5.3 밀도 그림과 추정

- 밀도 그림 *density plot* : 히스토그램을 부드러운 곡선으로 나타낸 그림.
 - 데이터로부터 밀도 그림을 얻기 위해서는 밀도 추정 함수를 이용해야 하는데, 주로 커널밀도추정 *kernel density estimation*가 사용된다.
 - 데이터의 분포를 연속된 선으로 보여준다.

- 히스토그램과 가장 큰 차이는 y축 값의 단위, 밀도그림에서는 개수가 아닌 비율을 표시.
- 밀도추정 density estimation
 - 관측된 데이터들의 분포로부터 원래 변수의 (확률) 분포 특성을 추정하고자 하는 것
 - 어떤 변수가 가질 수 있는 값 및 그 값을 가질 가능성의 정도를 추정하는 것

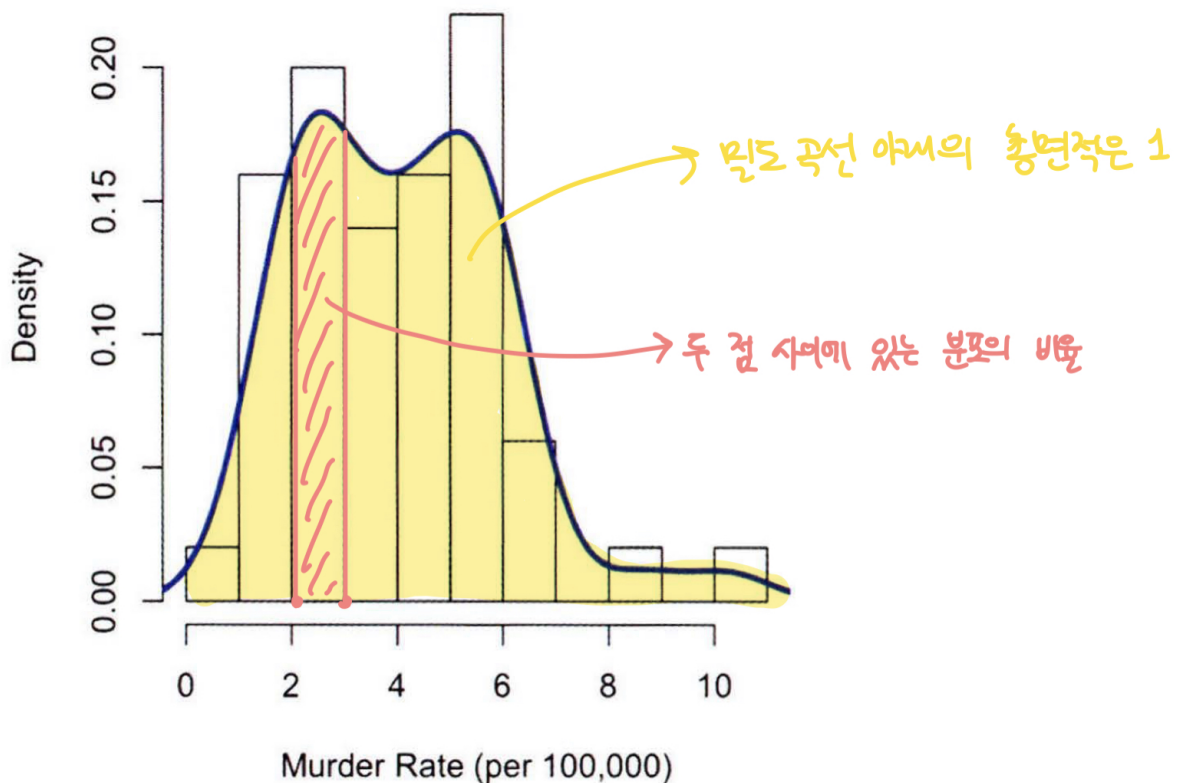


그림 1-4 주별 살인율 밀도

- 그래프 해석 (틀릴 수 있음..)
- 위 그림에서 밀도 함수와 x축이 그리는 면적(노란색)은 1 => 모든 데이터들은 x축에 나타난 구간 안에 분포 한다.
- 빨간색 빗금친 영역 : 주별 살인율이 10만명 당 2~3%(or 건)일 확률을 나타낸다.

1.6 이진 데이터와 범주 데이터 탐색하기

범주형 데이터에서는 간단한 비율이나 퍼센트를 이용해 탐색할 수 있다.

- 막대도표 *bar chart* : 각 범주의 빈도수 혹은 비율을 막대로 나타낸 그림

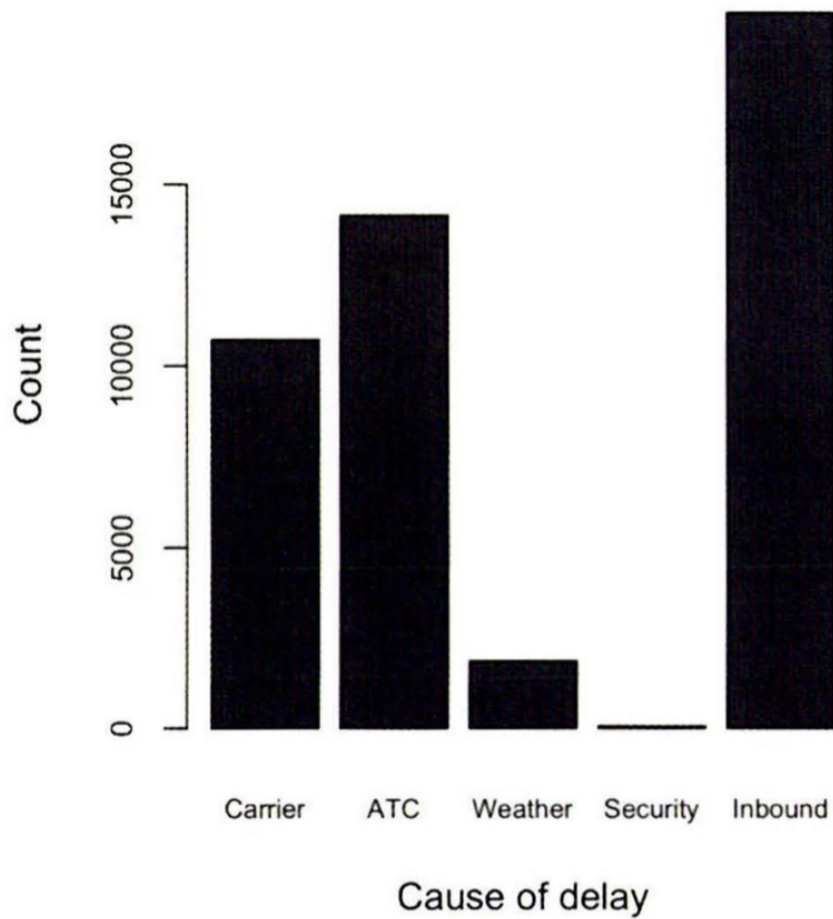
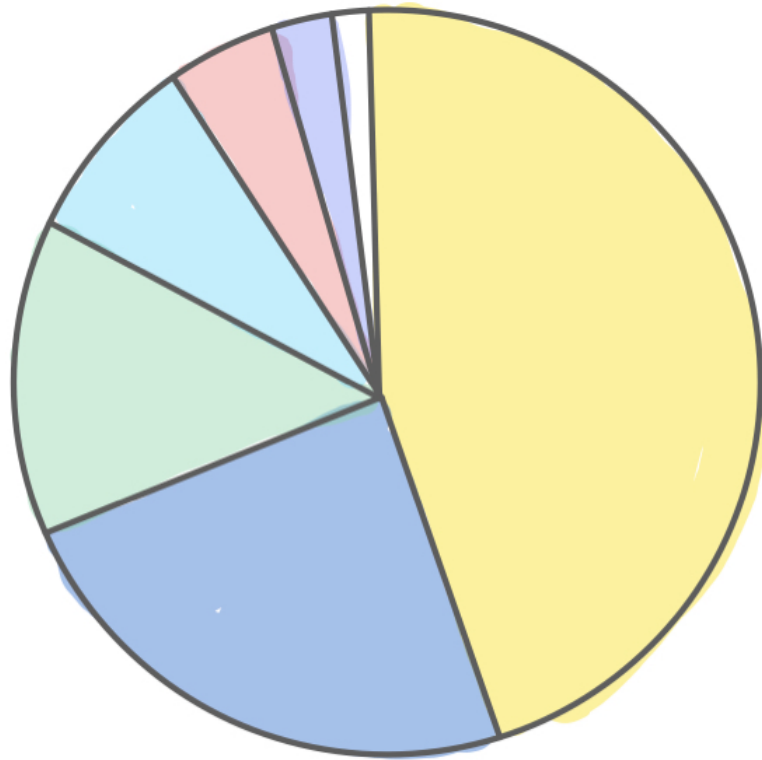


그림 1-5 델러스-포트워스 공항의 항공기 운행 지연 요인을 보여주는 막대도표

- 파이그림 *pie chart* : 각 범주의 빈도수 혹은 비율을 원의 부채꼴 모양으로 나타낸 그림



pie chart

- 범주형 데이터로서의 수치 데이터
 - 수치 데이터를 구간별로 나누어서 데이터의 분포를 나타낸 도수분포표는 수치형 데이터를 범주형으로 변환한 형태라고 볼 수 있다. 이는 데이터의 복잡도(그리고 크기)를 줄여준다는 점에서 중요하고 실제로 많이 사용된다. 특히 데이터 분석을 시작하는 단계에서 피쳐 간의 관계를 알아보기 위해 사용된다.

1.6.1 최빈값

- 최빈값 *mode* : 데이터에서 가장 자주 등장하는 범주 혹은 값
 - 범주형 데이터를 분석하는데 간단히 사용되지만, 수치 데이터에서는 잘 사용하지 않음

1.6.2 기댓값

- 기댓값 *expected value* : 범주에 해당하는 어떤 수치가 있을 때, 범주의 출현 확률에 따른 평균
 - 가중평균과 같은 꼴. 보통 주관적인 평가에 따른 미래의 기댓값과 각 확률 가중치만큼을 모두 더한 값.

- 실제 사업 평가나 자본 예상에 가장 근본적인 토대가 된다.
- 예
 - 두 가지 서비스 제공 시,
 - 웨비나 참석자의 5%는 30만원짜리 상품에, 15%는 5만원 상품에 가입하고, 나머지 80%는 어느 것도 가입하지 않을 것이라는 설문
 - 기댓값 = $(0.05)(300) + (0.15)(50) + (0.8)(0) = 22.5$
 - 웨비나 참석자들의 가댓값은 매달 22,500원

1.6.3 확률

- 여기서 사건이 발생할 확률이란, 상황이 수없이 반복될 경우 사건이 발생할 비율을 의미한다.

1.7 상관관계



상관계수(correlation coefficient) : (범위는 -1 에서 1 까지) | 숫자(수치) 변수들 사이에 어떤 관계가 있는지를 나타내기 위해 사용되는 측정량

상관 행렬 : 행과 열 \Rightarrow 변수들을 의미하는 표 / 각 셀은 그 행과 열에 해당하는 변수들 간의 상관관계를 의미한다.

• 상관관계 (피어슨 상관계수)

- 변수들이 선형관계가 아닌경우 더 이상 유용한 측정 지표가 아님
- 두 변수사이 무슨 관계가 있는지 측정

$-1.0 \leq r \leq -0.7$: 매우 강한 음(-) 의 상관 관계

$-0.7 < r \leq -0.3$: 강한 음(-) 의 상관 관계

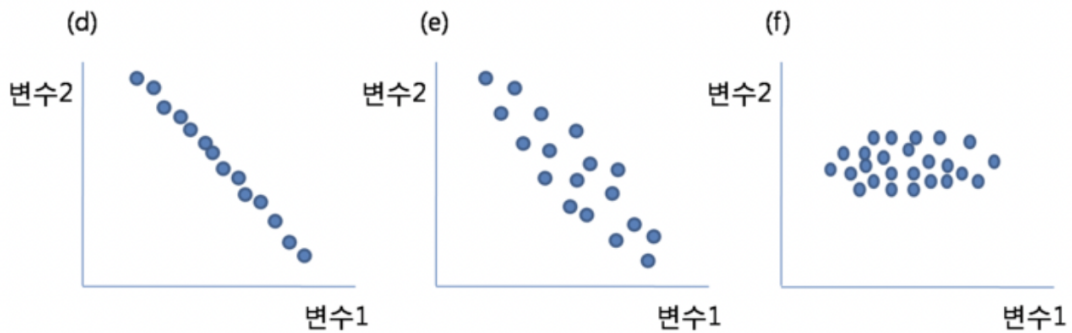
$-0.3 < r \leq -0.1$: 약한 음(-) 의 상관 관계

$-0.1 < r \leq 0.1$: 상관 관계 없음

$0.1 < r \leq 0.3$: 약한 양(+) 의 상관 관계

$0.3 < r \leq 0.7$: 강한 양(+) 의 상관 관계

$0.7 < r \leq 1.0$: 매우 강한 양(+) 의 상관 관계



한편, 이러한 두 변수간 상관 관계는 말 그대로 서로 **선형적인 증가, 감소와 관련된 상호 관계**만을 나타낼 뿐이지 한 변수가 다른 변수에 **영향을 주는 관계**를 의미하지는 않습니다. 만일 한 변수가 다른 변수에 영향을 주어 두 변수간 서로 관련성을 보일 경우에는 상관 관계가 아니라 **인과 관계(causation)**를 가진다고 합니다.

예를 들어, 어떤 고등학교 3학년 학생들의 모의고사 성적을 각 과목별로 등수로 매겼을 때 언어 영역의 등수와 수리 영역의 등수가 서로 상관 관계가 있는지는 스피어만 상관 계수로 알아볼 수 있습니다.

- **스피어만 상관계수 Spearman correlation coefficient**

상관 관계를 분석하고자 하는 두 연속형 변수의 분포가 심각하게 **정규 분포(normal distribution)**를 벗어나거나 또는 두 변수가 **순위 척도(ordinal scale)** 자료일 때 사용하는 값입니다. 연속형 자료일 때는 각 측정값을 순위 척도 자료로 변환시켜 계산합니다.

스피어만 상관 계수는 피어슨 상관 계수와는 달리 **선형적인 상관 관계를 나타내지 않고 단순히 한 변수가 증가할 때 다른 변수가 증가하는지 감소하는지에 대한 관계**만을 나타내는데, 켄달의 상관 계수(Kendall's correlation coefficient)와 함께 대표적인 **비모수적(non-parametric)** 상관 계수입니다.

- **산점도**

- 두 변수 사이의 관계 시각화
- plot.scatter 사용

1.8 두 개 이상의 변수 탐색하기

- 단어체크

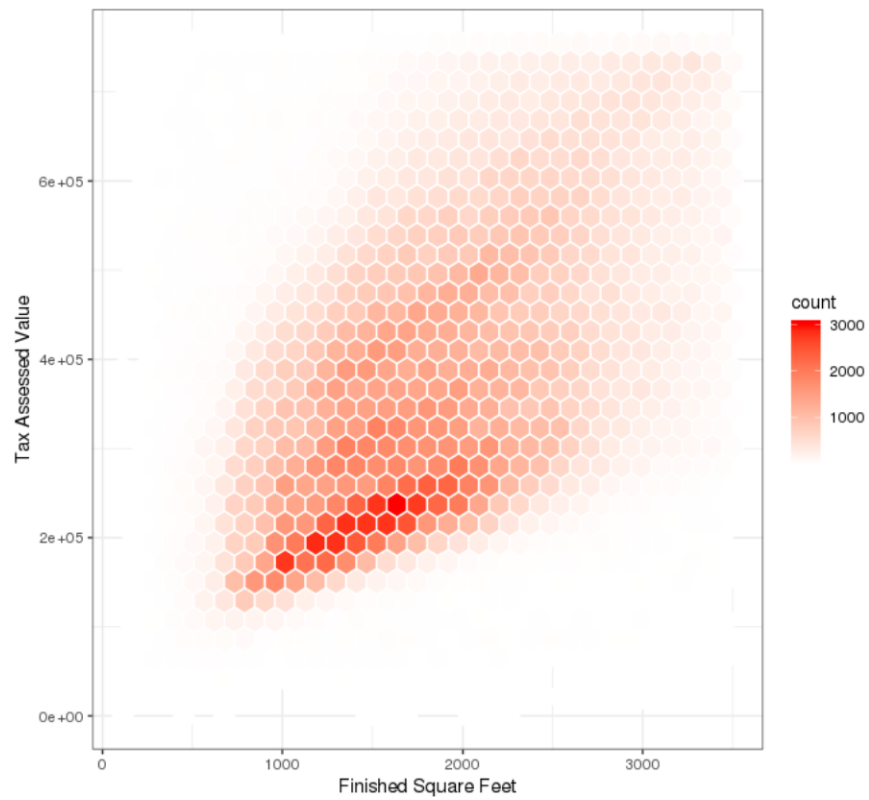
용어정리

≡ 한국어	Aa 영어	≡ 설명
단순	<u>simple</u>	독립변수(X)가 1개
다중	<u>multiple</u>	독립변수(X)가 2개 이상
<u>일변량</u>	<u>univariate</u>	종속변수(Y)가 1개
<u>이변량</u>	<u>bivariate</u>	종속변수(Y)가 2개
<u>다변량</u>	<u>multivariate</u>	종속변수(Y)가 2개 이상

- 1.8 -1 육각형 구간과 등고선

육각형 구간

- 데이터가 많아지면 → 스캐터 플롯 사용 못함
- Python에서는 hexbin을 사용해서 ⇒ plot.hexbin ⇒ 육각형 구간 도표를 쉽게 사용할 수 있다.

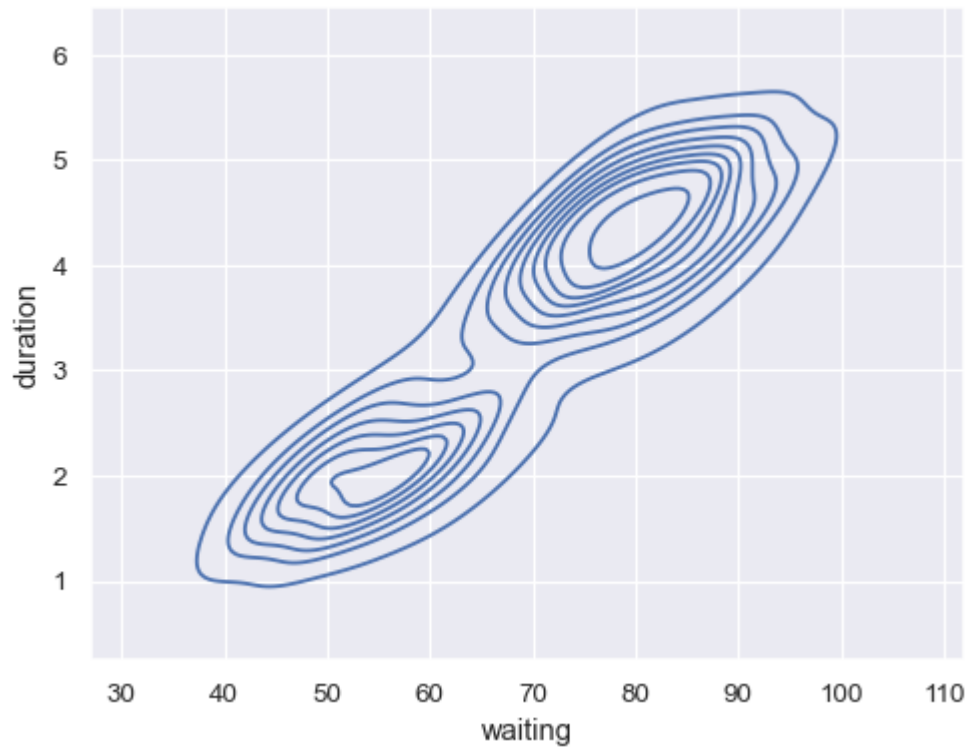


이거슨 예시 ^^

- 기록된 갯수 (count)에 따라서 표시를 한다.

등고선

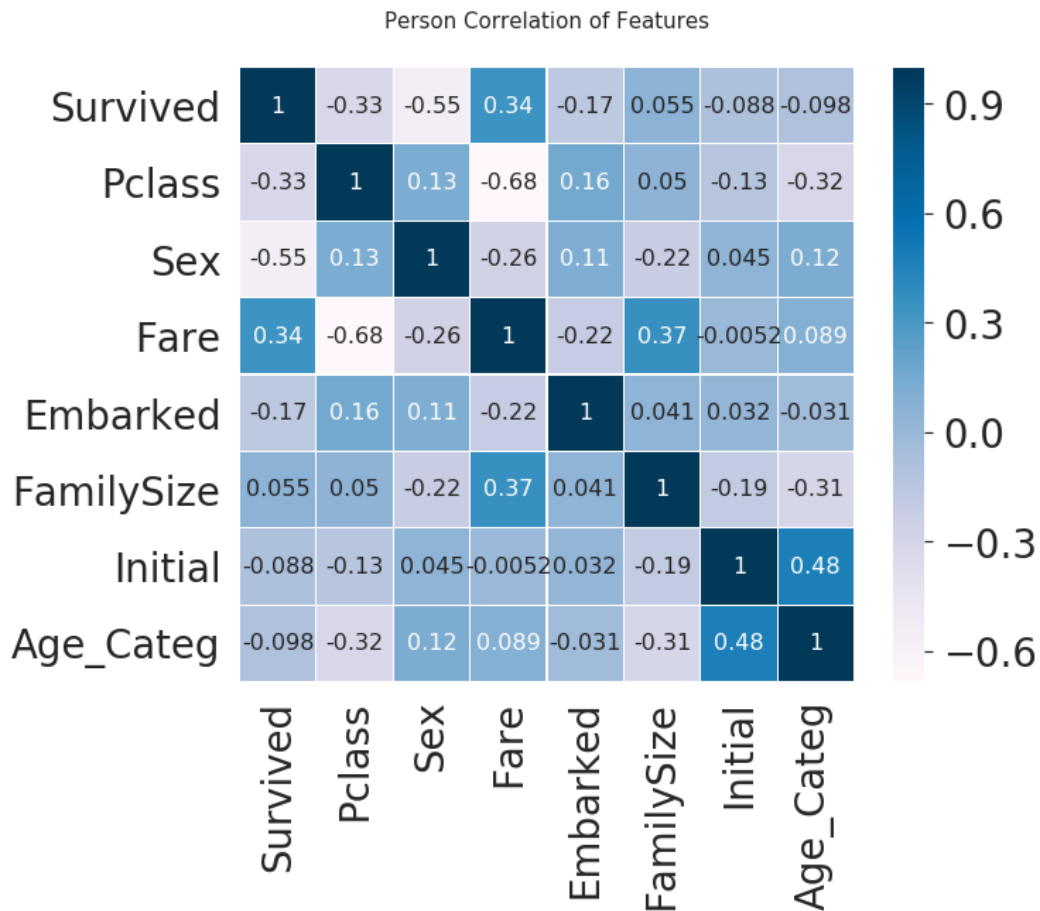
- seaborn의 `kedplot`
- 꼭대기 쪽으로 갈수록 밀도가 높아지고
- 같은 등고선은 밀도가 같다.



히트맵

- 아 이건 따로 설명 X 모두 알테니

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f46a6709a58>
```



1.8.2 범주형 변수 대 범주형 변수

분할표

- 두 범주형 변수를 요약하는데 효과적인 방법
- 파이썬의 Pivot_table ⇒ aggfunc 인수를 사용하면 횡수 정보 취득 가능

1.8.2 범주형 변수 대 수치형 변수

상자그림(박스플롯)

- stats.boxplot(by = ' ')
- by 인수로 데이터 집합을 그룹별로 분할함

바이올린 도표

- 상자그림 보완
- y축을 따라 밀도 추정 결과를 동시에 시각화
- 밀도 분포 모양을 대칭으로 겹쳐놓고 ⇒ 바이올린 모양
- 데이터의 분포를 볼 수 있다는 장점이 있다.

1.8.4 다변수 시각화하기

-
- 조건화 - 복잡하다.
- 포인트는 그냥 그래프를 쪼개서 나눠서 EDA한다는 뜻