스터디 3주차
기술통계분석

1. 데이터란?

변수

관측값

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S

1. 데이터분석의 목적

데이터분석은 6가지 목적을 갖고 목적에 따라 과정이 달라진다.

(1) 현상을 설명 : 기술통계분석(Descriptive analysis)

- 데이터를 통해 **"지금의 상태를 설명(현상)을"** 설명 하는것이 목표
- 생존자는 얼마나 될까?, 탑승객의 나이는 어땠을까?

(2) 여러 데이터간 관계를 탐색: 탐색적 데이터분석(Exploratory Data Analysis)

- 생존과 강한 연관이 있는 요인들은 어떤 요인일까?
- 가설이라는것이 등장.

03 주차 스터디

3주차 스터디 내용

1 데이터분석에서 데이터란?

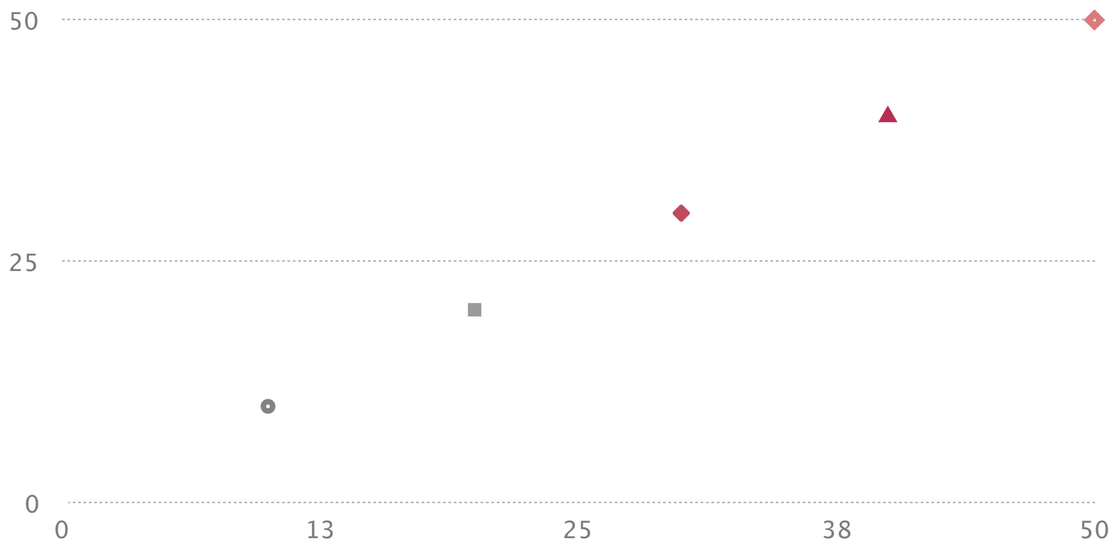
2 기술통계분석과 EDA 상세

1. 데이터분석에서 데이터란

데이터는 N차원의 벡터(점) 이다.

(1) 데이터는 N차원의 벡터(점) 이다.

- N차원이란 점을 설명하는 특성(변수)의 개수
- 점은 N개의 특성으로 표현이 가능

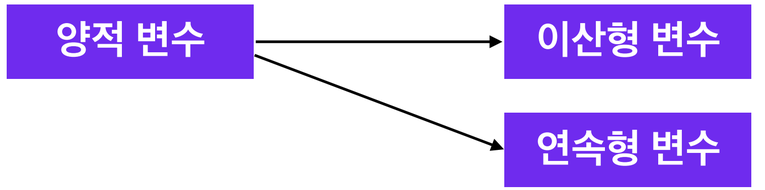


1. 변수란?

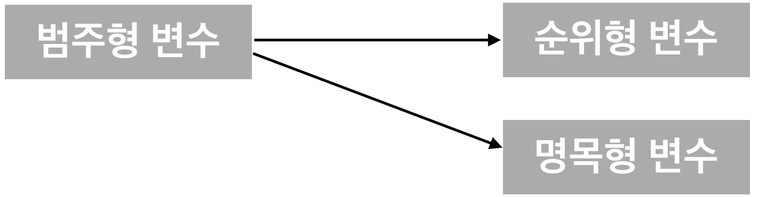
원래는 변하는 수 를 변수라고 부름. 단 점(데이터)을 설명하는 특성을 변수라고 생각하면 쉬움.

(**1) 변수의 성질에 따른 종류** (내포된 의미에 맞게 가변적임)

- **양적 변수** : 금액, 몸무게 처럼 정보의 양을 표현하기 위한 변수
  - 이산형 변수, 연속형 변수로 구분할 수 있음.
  - 단 관념적으로 구분할 수 있으나, 구분은 분석가의 재량



- **범주형 변수(질적 변수) :** 성별, 선호도와 같이 카테고리로 구분 하는 변수
  - 순위형 변수 : 카테고리 내 순서가 있는 경우 ex) 선호도, 등급
  - 명목형 변수 : 순위가 없는 경우 ex) 성별, 연령대 등



1. 타이타닉 데이터 변수 분류해보기

PassengerId : 식별자 (유저 마다 고유한 값 ex 주민등록번호, 사회보장번호) –> 파악할 필요없음.
Survived : 0과 1로만 구분되어 있음. (0 = No, 1 = Yes)
Pclass : 티켓 등급(사회적 등급을 나타냄) (1 = 1st, 2 = 2nd, 3 = 3rd) [1등급일수록 고급등급]
Name : 이름
Sex : 성별
Age : 나이
SibSp : 형제/자매, 배우자(약혼자)
Parch : 자녀
Ticket : 티켓
Fare : 티켓요금
Cabin : 짐칸
Embarked : 승선장 (C = Cherbourg, Q = Queenstown, S = Southampton)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

2. 기술통계 분석

기술통계 분석의 목적은 데이터, 정확하게는 변수를 이해하고 파악하기 위함.

* 기술(Description) 은 대상이나 과정의 내용과 특징을 있는 열거, 서술하다.

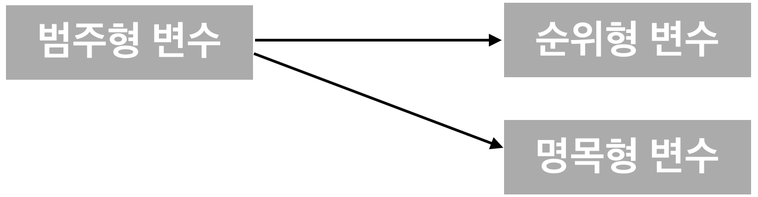
* 기술통계분석은 자료를 조직하고 요약하는 한 방법으로, 실험자로부터 얻은 데이터를 자료의 특성들을 이해하기 쉽게 기술해 주는 수치로 표현하는 방법.

* 기술통계치 : 실험자로부터 얻은 데이터를 자료의 특성을 이해하기 쉽게 기술한 내용. -> 변수마다 다름

변수 타입에 따라 분석하는 방법론이 달리지기 때문에 정확하게 파악(or 설계) 하는 것이 중요하다.
뭘 봐야하지? 는 변수의 종류를 파악하는 것으로 부터 시작함.

2. 기술통계 분석

기술통계 분석의 목적은 데이터, 정확하게는 변수를 이해하고 파악하기 위함.



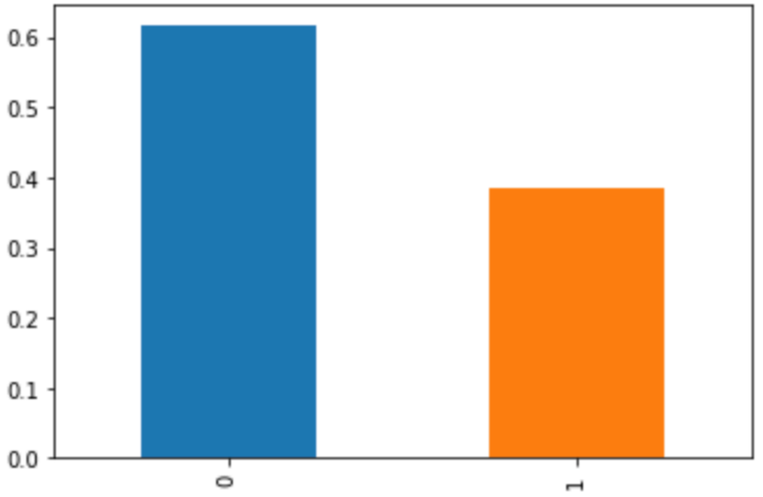
범주형 변수의 기술통계 분석

- 기술통계치 : 피봇테이블을 활용하여, 갯수(count), 횟수(Frecuncy) 나, **비율(Ratio)**
- 그래프 : 막대그래프(Bar Plot) 를 통해 살펴본다.
  HTTPS://PUBLIC.TABLEAU.COM/PROFILE/BO.MCCREADY8742#!/

```
# Survived 칼럼
#value_counts() 메소드에 관한 설명은 201 p에서 참조가능.

train.Survived.value_counts()
```

```
0    549
1    342
Name: Survived, dtype: int64
```



2. 기술통계 분석

기술통계 분석의 목적은 데이터, 정확하게는 변수를 이해하고 파악하기 위함.

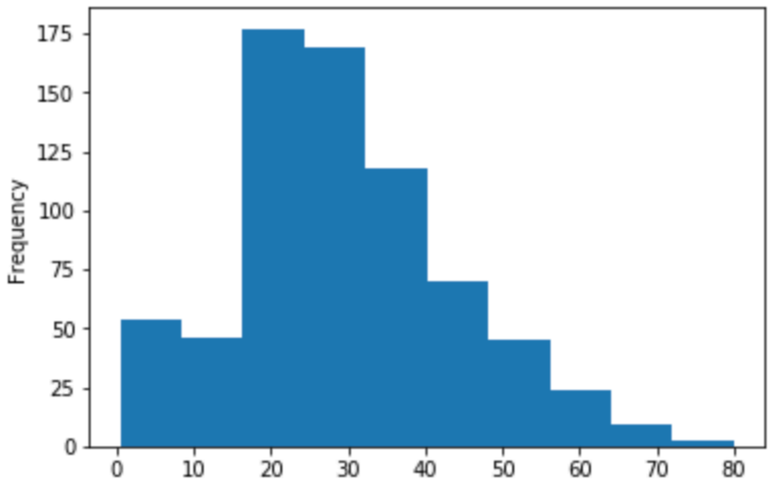
양적 변수 → 이산형 변수
양적 변수 → 연속형 변수

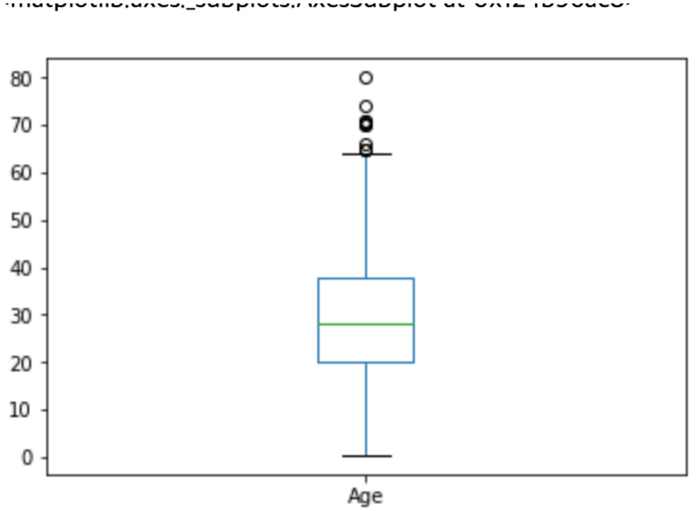
수치형 변수의 기술통계 분석

- 기술통계치 : 평균, 중앙값, 표준편차(분산), 왜도, 첨도, 4분위수, 10분위수

- 그래프 : 히스토그램(혹은 밀도그래프), 상자그래프 누적분포 그래프 등을 사용하여 분포를 요약함.
  HTTPS://PUBLIC.TABLEAU.COM/PROFILE/BO.MCCREADY8742#!/

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200



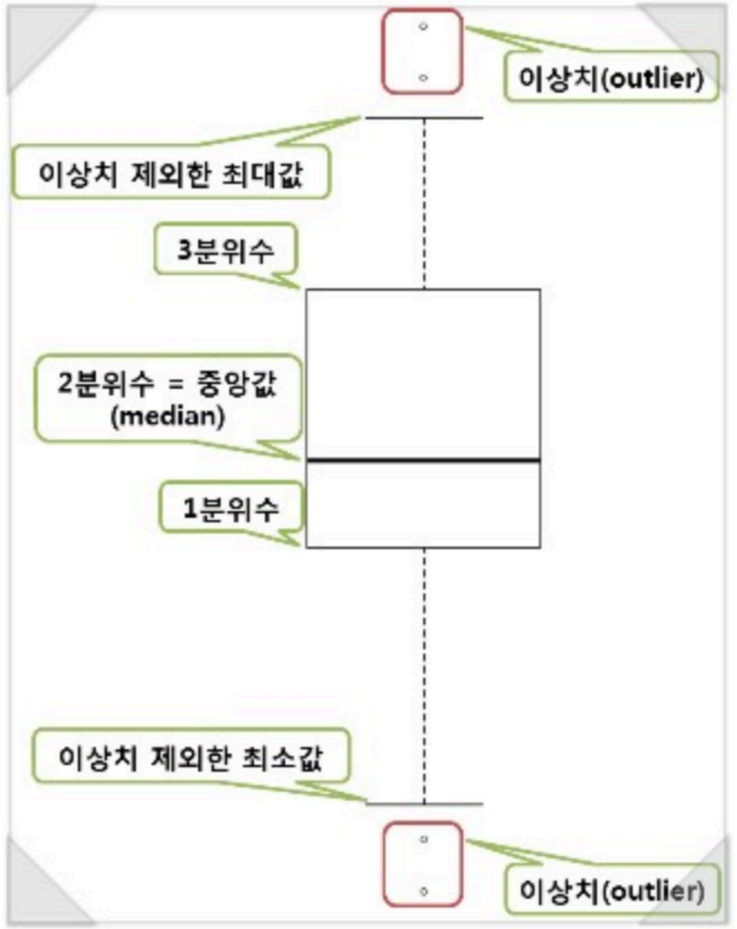


2. 기술통계 분석

기술통계 분석의 목적은 데이터, 정확하게는 변수를 이해하고 파악하기 위함.

상자 그림 해석하는 법

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200



2. 탐색적 데이터 분석

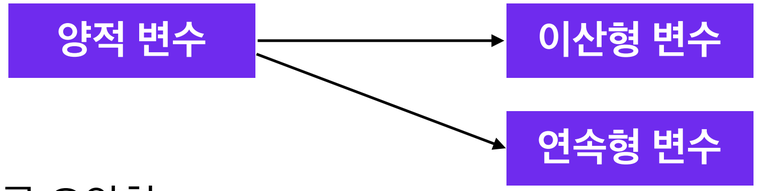
타겟변수(종속변수)와 다른 변수(독립변수)간의 관계를 탐색

타겟이 범주형 변수고 독립변수도 범주형 변수일때

- 기술통계치 : 그룹화된 피봇테이블을 활용하여, 갯수(count), 횟수(Frecuncy) 나, **비율(Ratio)**
- 그래프 : 그룹화된 막대그래프(Bar Plot) 를 통해 살펴본다.

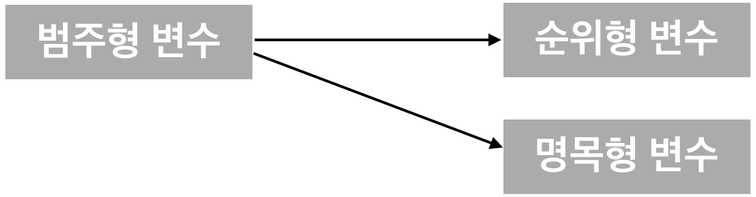
타겟이 범주형 변수고 독립변수가 양적 변수일때(범주형이 항상그룹)

- 기술통계치 : 그룹 별 평균, 중앙값, 표준편차(분산), 왜도, 첨도, 4분위수, 10분위수
- 그래프 : 히스토그램(혹은 밀도그래프), 상자그래프 누적분포 그래프 등을 사용하여 분포를 요약함.



타겟이 양적 변수고 독립변수가 범주형 변수일때(범주형이 항상그룹)

- 기술통계치 : 피봇테이블을 활용하여, 갯수(count), 횟수(Frecuncy) 나, **비율(Ratio)**
- 그래프 : 막대그래프(Bar Plot) 를 통해 살펴본다.



타겟이 양적 변수고 독립변수가 양적 변수일때

- 기술통계치 : 공분산(Corvariation), 상관계수(Correlation)
- 그래프 : 히트맵(Heatmap), 산점도(scatter Plot)

Q&A

실습 시작

그리고 과제