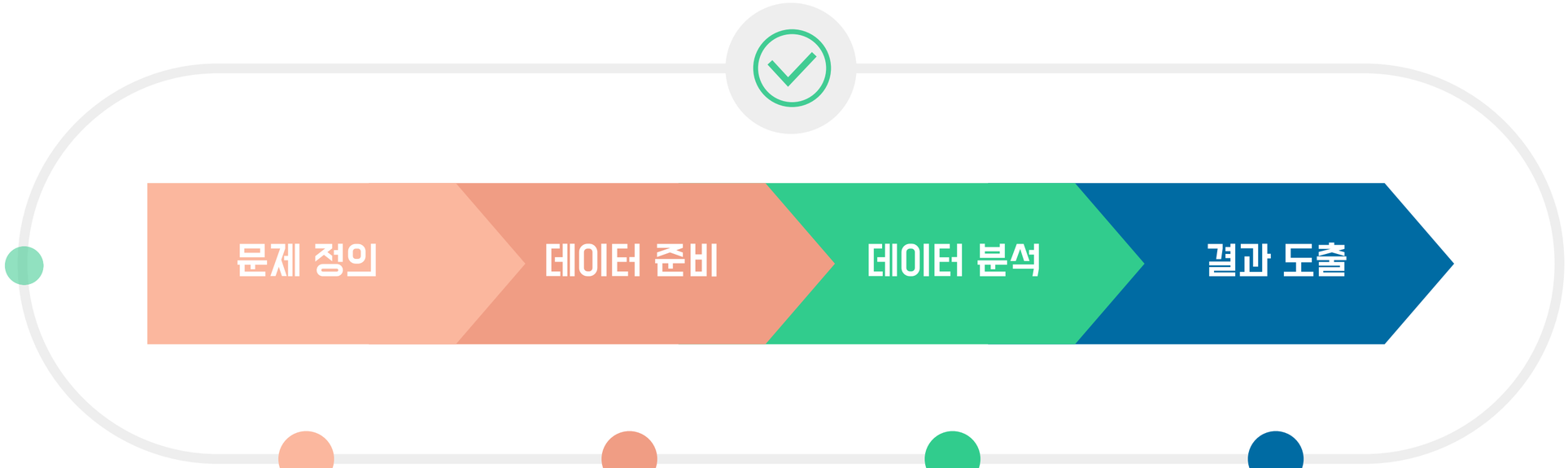
2022-1 신입교육세션

EDA&시각화

19기 DA 이은지

20기 분들 환영합니당 :)

01
데이터 분석 과정



문제 정의 및 분석 목적 설정

어떤 데이터가 필요할 지 파악

데이터 수집 및 가공

-크롤링, 공공데이터, 로그 데이터 등
- pandas, numpy 같은 패키지를 활용한 데이터 가공

전처리와 모델링

-전처리: EDA, FE
-모델링: 머신러닝, 딥러닝

결과 해석 및 활용

분석 목적에 맞추어 결과를 해석
도출된 결과를 문제 해결에 활용

02
EDA vs FE

EDA Exploratory Data Analysis 탐색적 데이터 분석

데이터를 살펴보면서 FE에서 사용할 자료의 특징을 찾는 것

FE Feature Engineering

EDA 단계에서 발견한 자료들의 특징을 이용해
ML/DL의 성능이 잘 나오도록 전처리하는 과정

Garbage in, garbage out

= EDA & FE 의 중요성

데이터를 잘 정제해야 좋은 결과를 얻을 수 있다

03
EDA의
4가지 주제

EDA시 크게 다음 네 가지에 집중

1. 저항성

: 자료의 일부가 파손되었을 때, 영향을 적게 받는 성질

ex 평균은 중앙값에 비해 자료의 이상치나 입력오류에 큰 영향을 받음
= 중앙값은 평균에 비해 저항성이 크다.

2. 잔차의 해석

잔차가 엄청 크거나 작은 값들(=아웃라이어)이 왜 생겼는지를 파악

3. 자료의 재표현

: 데이터의 분석과 해석을 단순하게 할 수 있도록 원래의 변수를 적당한 척도로 바꾸는 것
자료가 선형적일 수도 있지만 로그/제곱근/역수 등으로 바꿔야 분석이 단순해질 때도 있음.

또 변수를 적당한 척도로 변환해봄으로써 분포의 대칭성, 선형성, 분산 안정성 등을 파악해볼 수 있음.

4. 그래프를 통한 현시성

그래프를 통한 시각화 → 데이터를 직관적으로 파악

04
EDA의 과정

① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

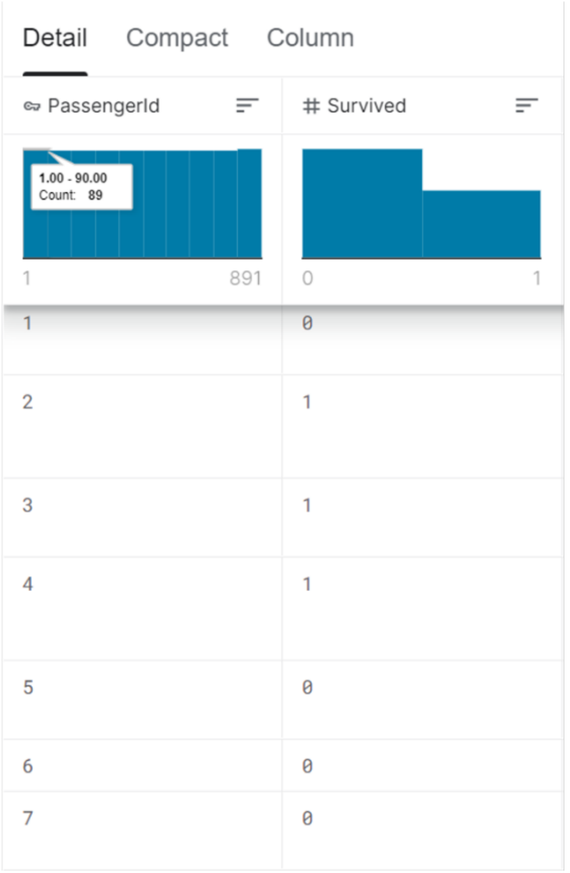
④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

좌석등급, 성별, 나이 등의 변수들을 가지고
승객의 생존 여부를 예측하는 Task

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton



05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

```
[55] #한 코드 내에서 여러 DataFrame을 보고 싶으면 꼭 display를 사용해주세요!

     display(df.head())
     display(df.tail())
```

	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

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

```
#pd.DataFrame.info() : 데이터 수, 각 column의 이름과 정상데이터 수, 데이터 타입 등 표시해줌
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

```
[192] # unique 활용
      print("---Passenger Id---₩n",df["PassengerId"].unique()[:10])
      print("₩n---Age---₩n",df["Age"].unique()[:10])
      print("₩n---Name---₩n",df["Name"].unique()[:10])
```

```
---Passenger Id---
 [ 1  2  3  4  5  6  7  8  9 10]

---Age---
 [22. 38. 26. 35. nan 54.  2. 27. 14.  4.]

---Name---
 ['Braund, Mr. Owen Harris'
 'Cumings, Mrs. John Bradley (Florence Briggs Thayer)'
 'Heikkinen, Miss. Laina' 'Futrelle, Mrs. Jacques Heath (Lily May Peel)'
 'Allen, Mr. William Henry' 'Moran, Mr. James' 'McCarthy, Mr. Timothy J'
 'Palsson, Master. Gosta Leonard'
 'Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)'
 'Nasser, Mrs. Nicholas (Adele Achem)']
```

05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

```
#결측치 비율
df.isnull().sum() / len(df) *100
#(df.isnull().sum() / len(df) *100).round(2)
```

결측치면 True 반환

```
PassengerId     0.000000
Survived        0.000000
Pclass          0.000000
Name            0.000000
Sex             0.000000
Age            19.865320
SibSp           0.000000
Parch           0.000000
Ticket          0.000000
Fare            0.000000
Cabin          77.104377
Embarked        0.224467
dtype: float64
```

05
EDA의 과정
with 타이타닉 데이터셋

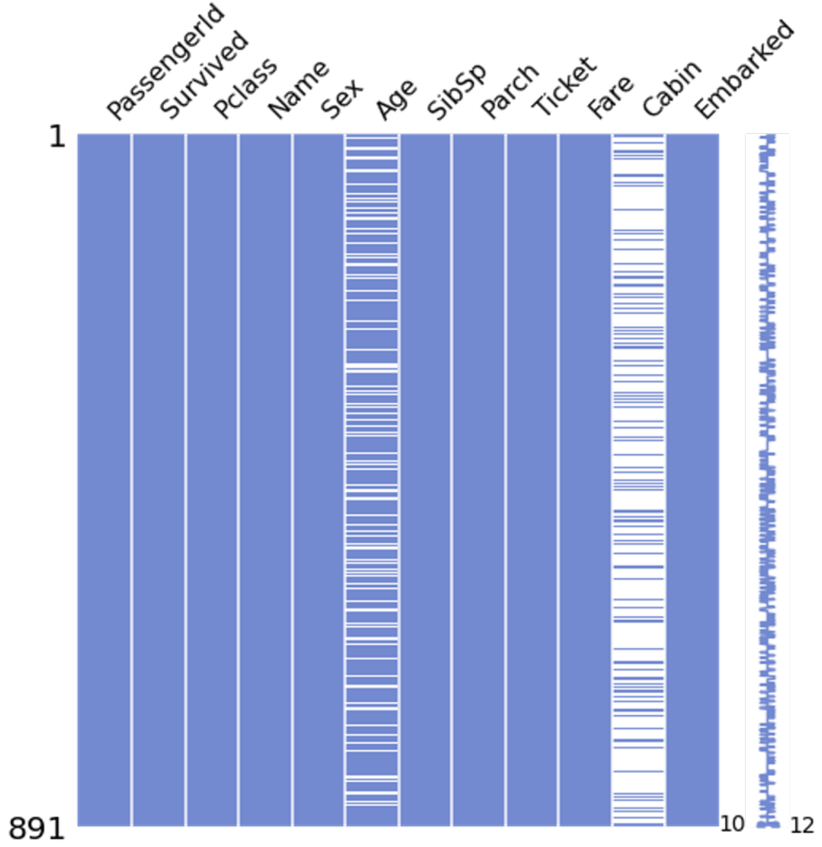
① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악



05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

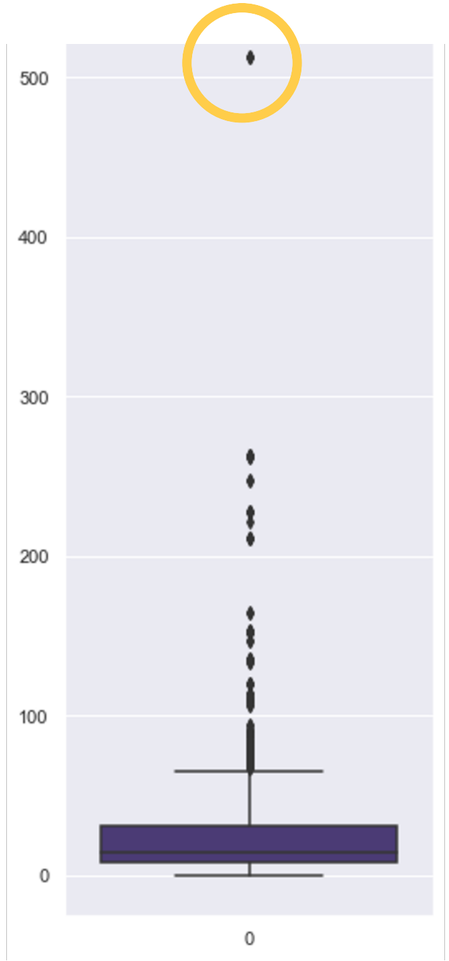
④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

```
df["Fare"].describe()
```

```
count    891.000000
mean      32.204208
std       49.693429
min        0.000000
25%        7.910400
50%       14.454200
75%       31.000000
max      512.329200
Name: Fare, dtype: float64
```

평균이 32인데 중앙값이 14
→ 이상치의 존재를 의심해볼 수 있음



05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

② 변수 타입 파악

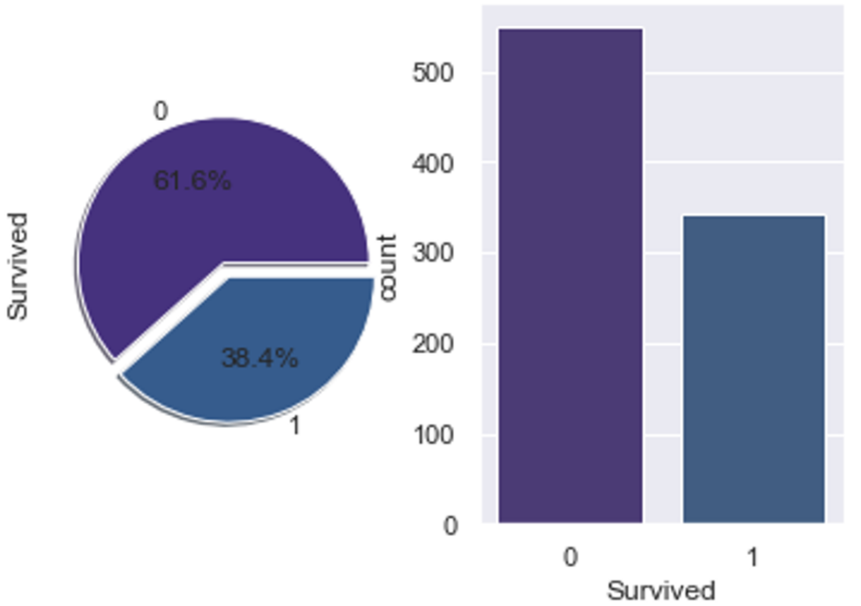
③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

```
df["Survived"].value_counts()
```

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



05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

② 변수 타입 파악

③ 결측치, 이상치 확인

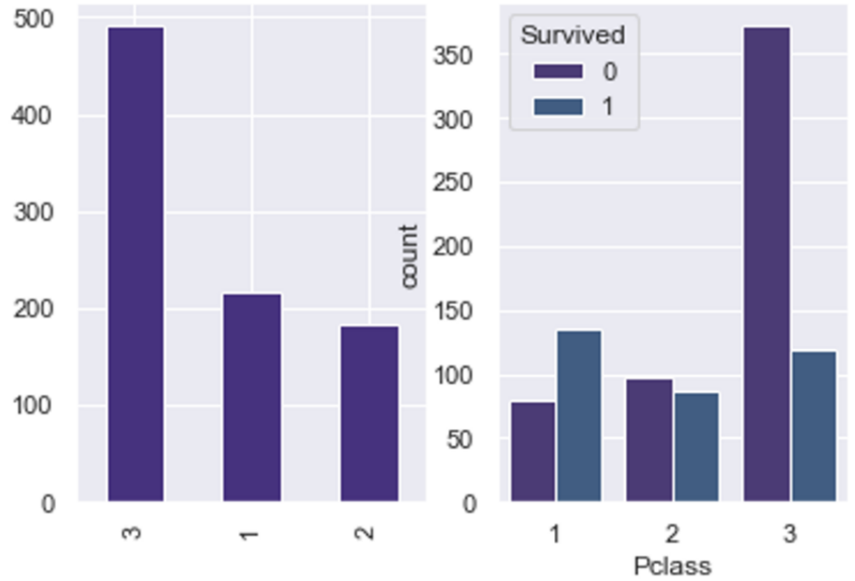
④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

Survived

Pclass	
1	136
2	87
3	119



barplot, countplot

Pclass(좌석 등급)에 따라 몇 명이 생존했는지
이렇게 숫자만 보더라도 1등급 사람들의 생존 비율이 더 높아요

05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

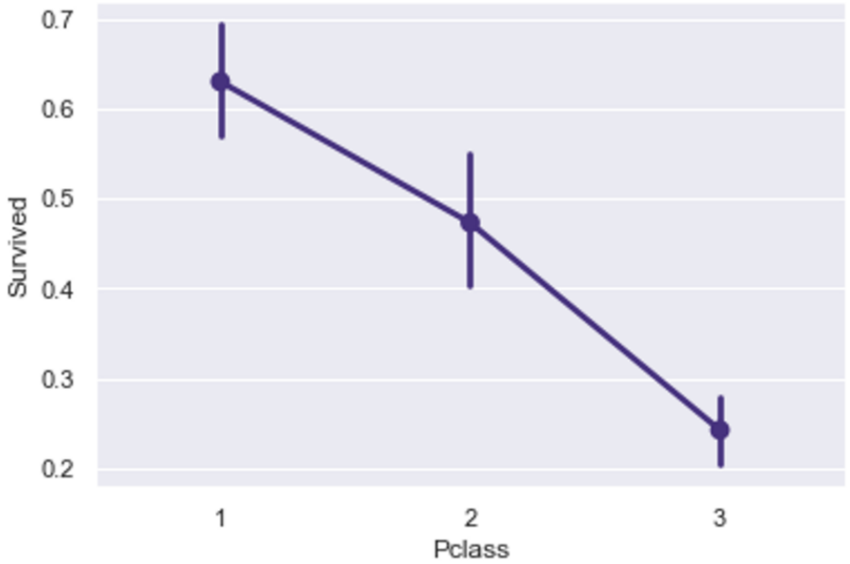
② 변수 타입 파악

③ 결측치, 이상치 확인

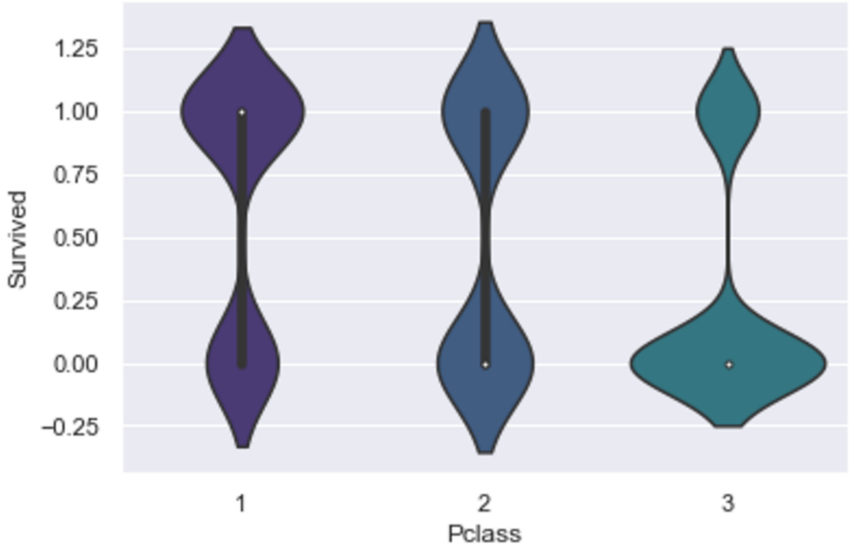
④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

데이터 타입/ 보고싶은 정보에 따라 다양한 플롯 활용 가능



Point Plot



Violin Plot

05
EDA의 과정
with 타이타닉 데이터셋

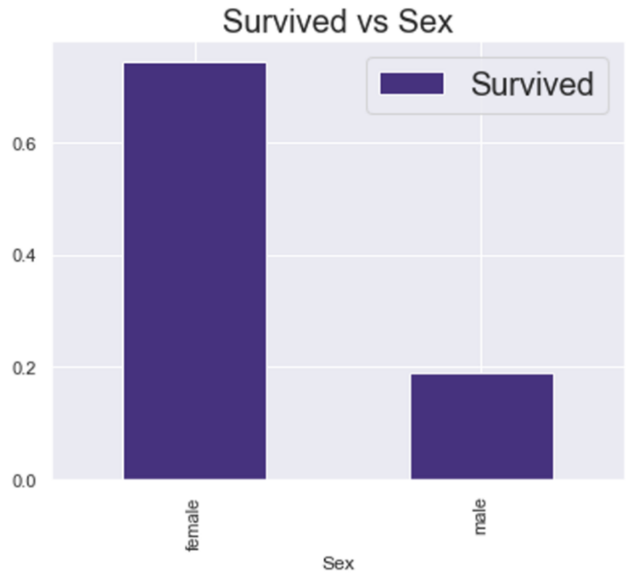
① 데이터 형태 파악

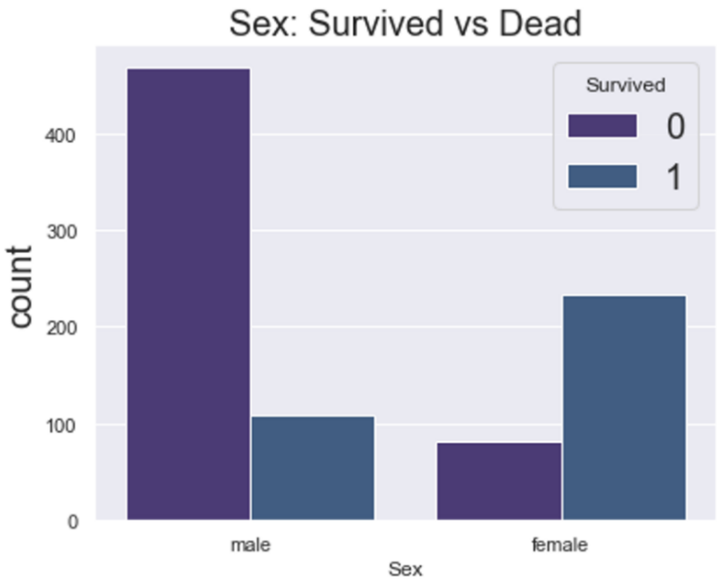
② 변수 타입 파악

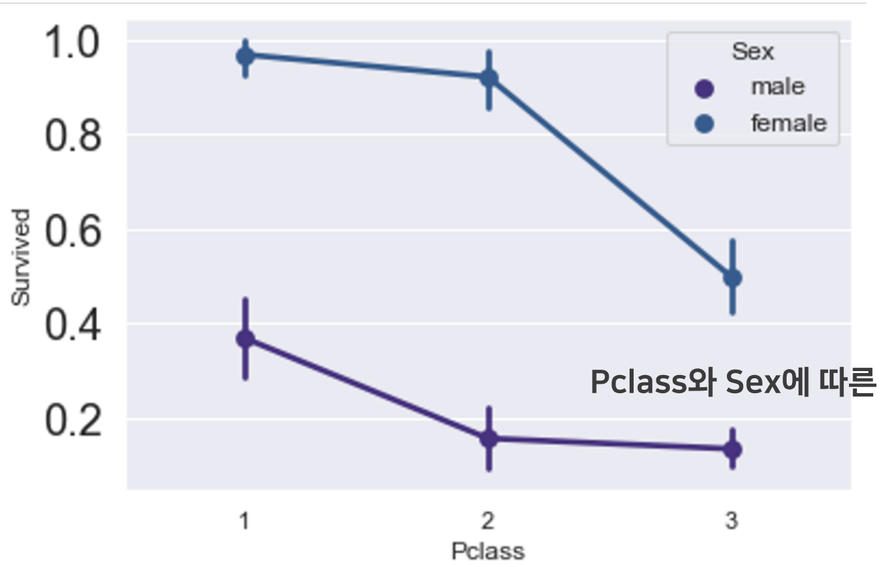
③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악







Pclass와 Sex에 따른 생존율을 동시에 확인 가능

05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

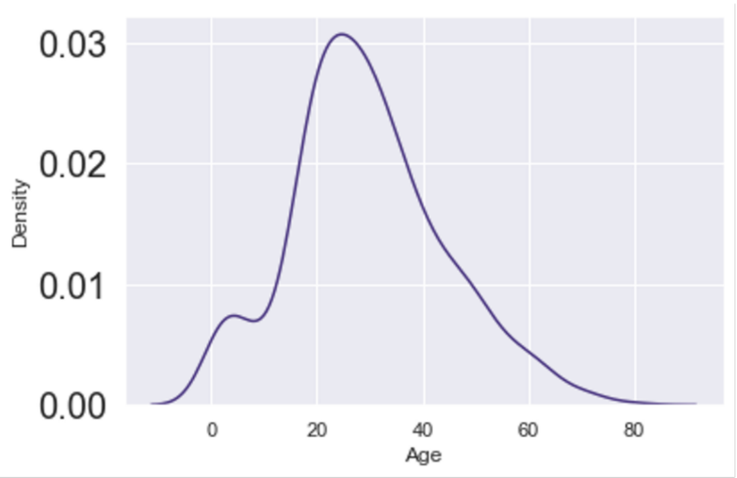
② 변수 타입 파악

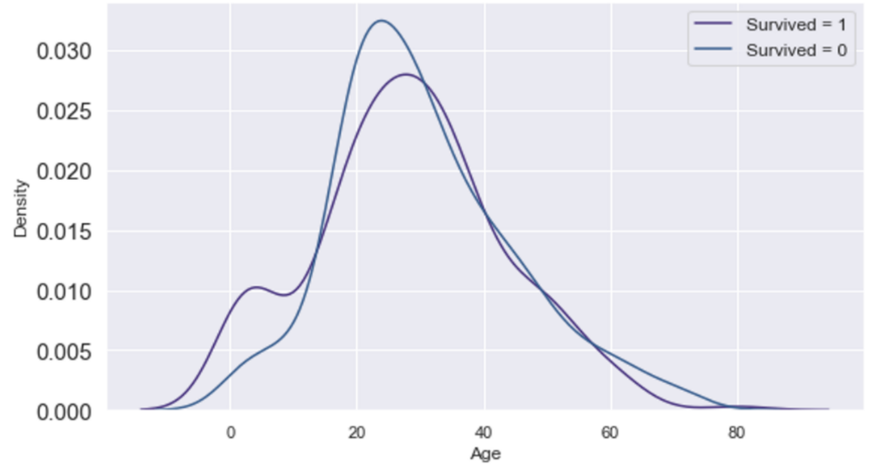
③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

실수형 변수인 age(나이)의 분포
- KDE Plot(확률 밀도 함수)





05
EDA의 과정
with 타이타닉 데이터셋

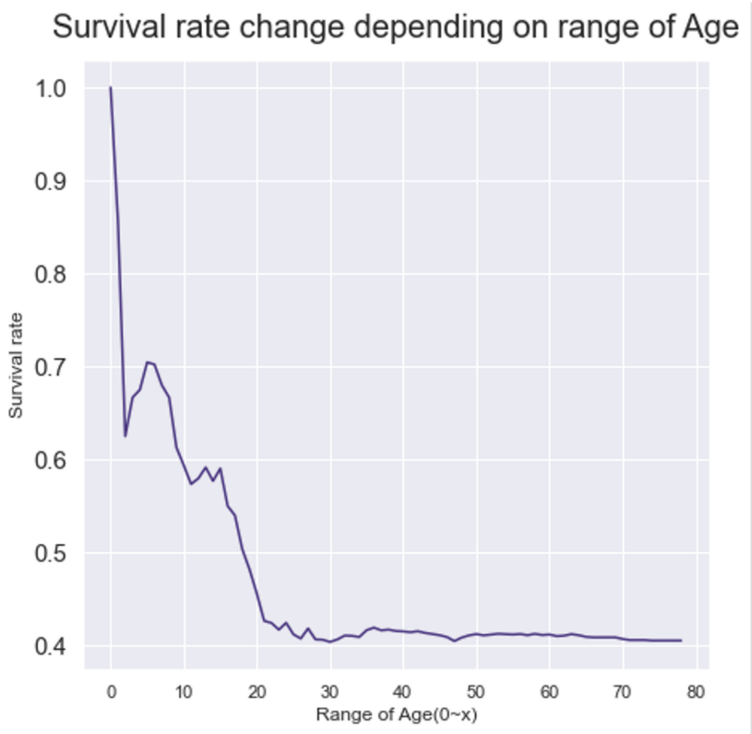
① 데이터 형태 파악

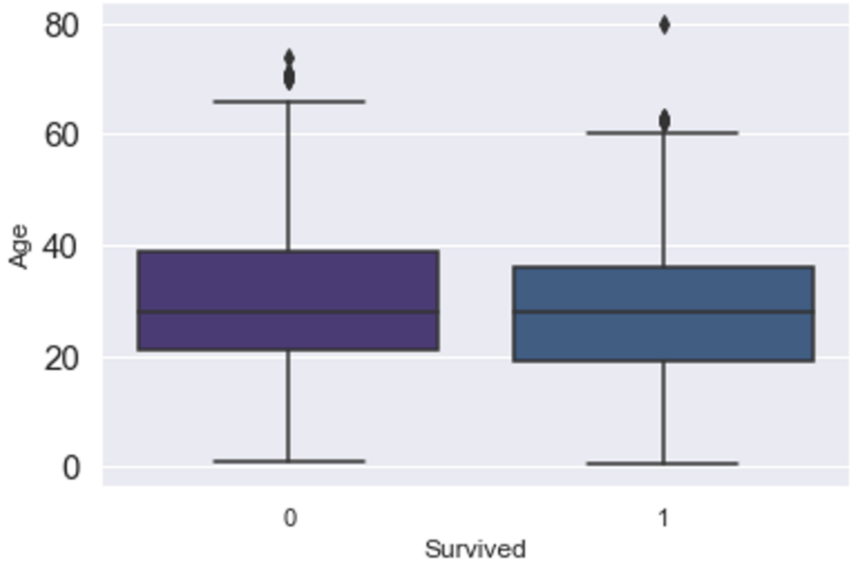
② 변수 타입 파악

③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악





05
EDA의 과정
with 타이타닉 데이터셋

① 데이터 형태 파악

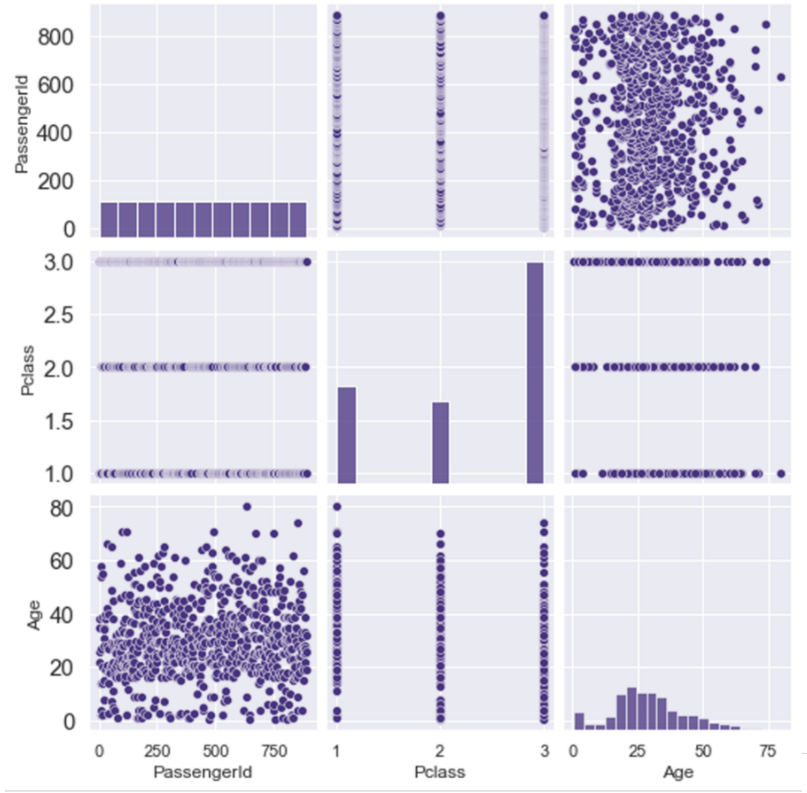
② 변수 타입 파악

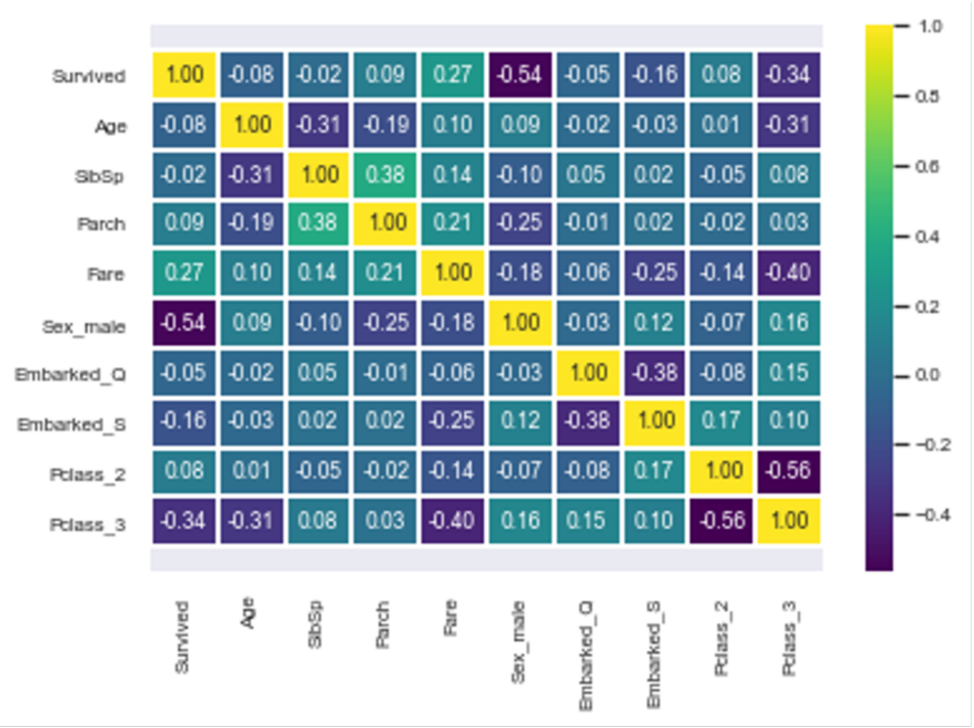
③ 결측치, 이상치 확인

④ 종속변수의 분포 확인

⑤ 변수들 간의 분포 &
변수-종속변수 간 관계 파악

변수들 간의 관계를 한 눈에 보기 좋은 pair plot, heatmap





2022-1 신입교육세션

수고하셨습니다

DA 화이팅 :>

친바하게 되면

만나요 우리 >_<