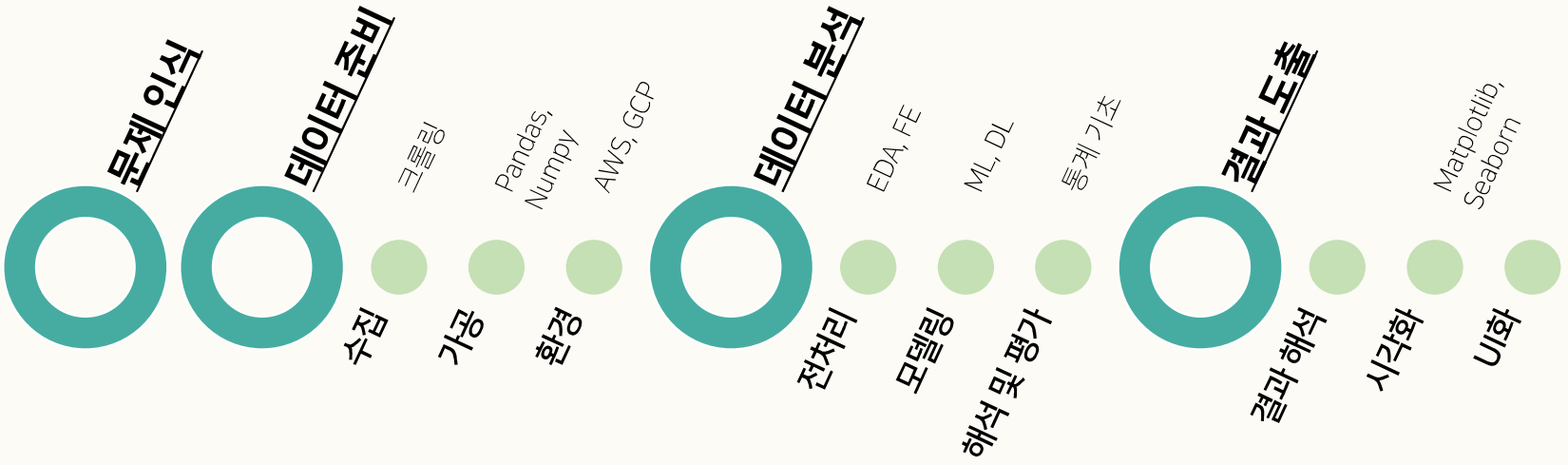
YBIGTA 18기 교육세션

EDA/시각화

Explorary Data Analysis / Data Visualization

17기 Data Analytics 팀 김수빈(경영)

1. 데이터 분석 과정


문제 인식
데이터 준비
수집
크롤링
가공
Pandas, Numpy
환경
AWS, GCP
데이터 분석
전처리
EDA, FE
모델링
ML, DL
해석 및 평가
통계 기초
결과 도출
결과 해석
시각화
Matplotlib, Seaborn
UI화

2. EDA와 FE의 차이

EDA

Exploratory Data Analysis

FE에 사용할 자료 특징 찾기

FE

Feature Engineering

전처리

3. EDA 이론

저항성

평균은 자료의 이상치나 입력오류의 영향을 많이 받음

평균보다 중앙값 선호

잔차의 해석

아웃라이어들이 왜 생겼는지 잘 파악해야 함

아웃라이어 : 이상치! 잔차가 엄청 크거나 작은 값

자료의 재표현

자료가 선형적일 수도 있지만 재표현해야 분석이 단순해질 때도 있음

로그 / 제곱근 / 역수 등

자료의 현시성

정말 간단한 자료가 아니면 숫자만 보고 이해하기는 어려움

자료를 시각화해서 보면 이해하기 편함

3. EDA 이론

일반적 EDA 과정

- 데이터 형태 파악
- 각 변수 타입 파악
- 결측치/이상치 확인
- 종속변수 분포 확인
- 다른 변수들의 분포, 종속변수와의 관계 확인

4. Titanic Data

데이터 형태 파악 | 각 변수 타입 파악 | 결측치/이상치 확인 | 종속변수 분포 확인 | 다른 변수들의 분포, 종속변수와의 관계 확인

- TIP! 데이터 제공 출처의 데이터 설명 활용

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

Detail Compact Column

PassengerId	# Survived
1.00 - 90.00 Count: 89	
1 891	0 1
1	0
2	1
3	1
4	1
5	0
6	0
7	0

4. Titanic Data

데이터 형태 파악 · 각 변수 타입 파악 · 결측치/이상치 확인 · 종속변수 분포 확인 · 다른 변수들의 분포, 종속변수와의 관계 확인

```
[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

4. Titanic Data

데이터 형태 파악 | **각 변수 타입 파악** | 결측치/이상치 확인 | 종속변수 분포 확인 | 다른 변수들의 분포, 종속변수와의 관계 확인

```
#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)']
```

- int64: 정수형 데이터
- float64: 실수형 데이터
- object: 문자형 데이터

4. Titanic Data

데이터 형태 파악 | 각 변수 타입 파악 | **결측치/이상치 확인** | 종속변수 분포 확인 | 다른 변수들의 분포, 종속변수와의 관계 확인

- 숫자로 결측치 확인

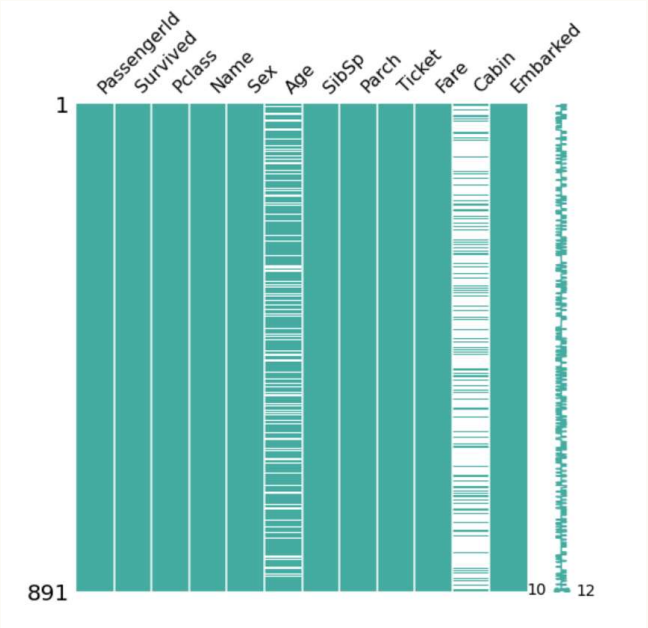
```
df.info()
```

```
<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
```

```
df.isnull().sum() / len(df) *100
#(df.isnull().sum() / len(df) *100
```

```
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
```

- 그래프로 결측치 확인



4. Titanic Data

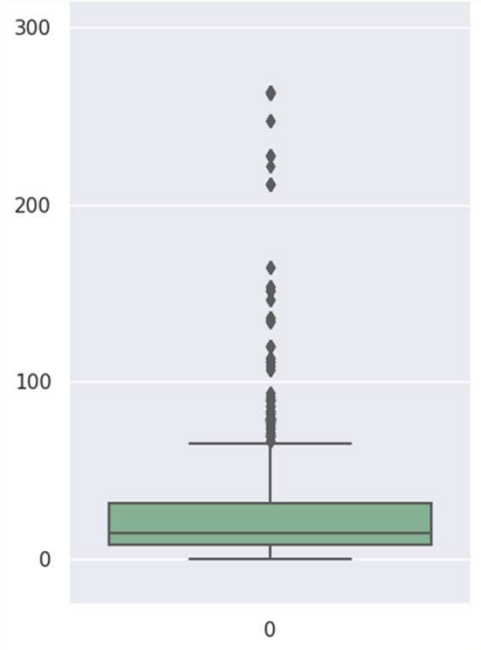
데이터 형태 파악 | 각 변수 타입 파악 | 결측치/이상치 확인 | 종속변수 분포 확인 | 다른 변수들의 분포, 종속변수와의 관계 확인

- 숫자로 이상치 확인

```
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
```

- 그래프로 이상치 확인



4. Titanic Data

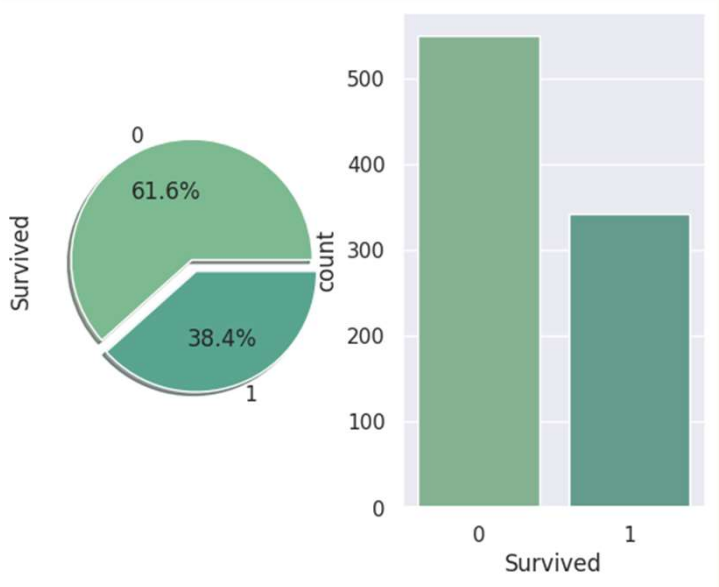
데이터 형태 파악 | 각 변수 타입 파악 | 결측치/이상치 확인 | **종속변수 분포 확인** | 다른 변수들의 분포, 종속변수와의 관계 확인

- 숫자로 종속변수 분포 확인

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

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

- 그래프로 종속변수 분포 확인



4. Titanic Data

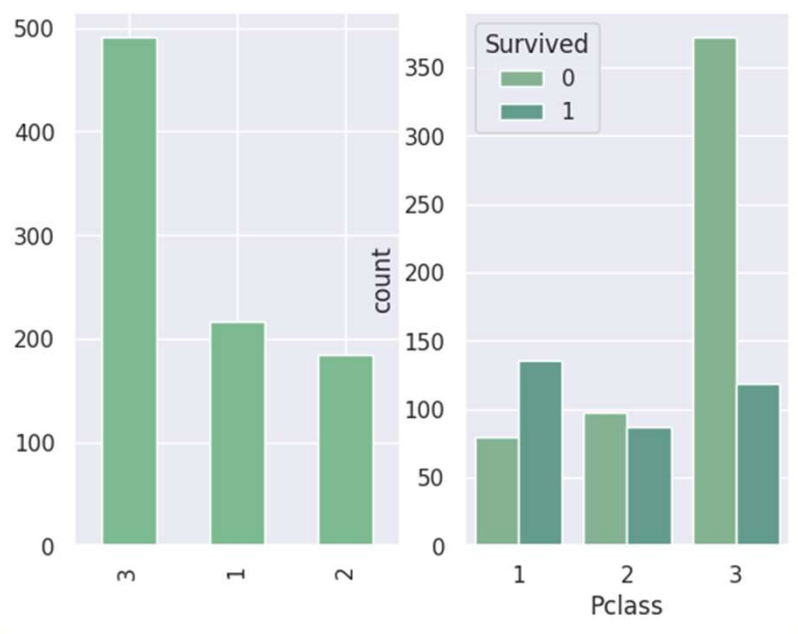
데이터 형태 파악 · 각 변수 타입 파악 · 결측치/이상치 확인 · 종속변수 분포 확인 · 다른 변수들의 분포, 종속변수와의 관계 확인

- 숫자로 종속변수와의 분포 확인

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

Pclass	Survived
1	136
2	87
3	119

- 그래프로 종속변수와의 분포 확인

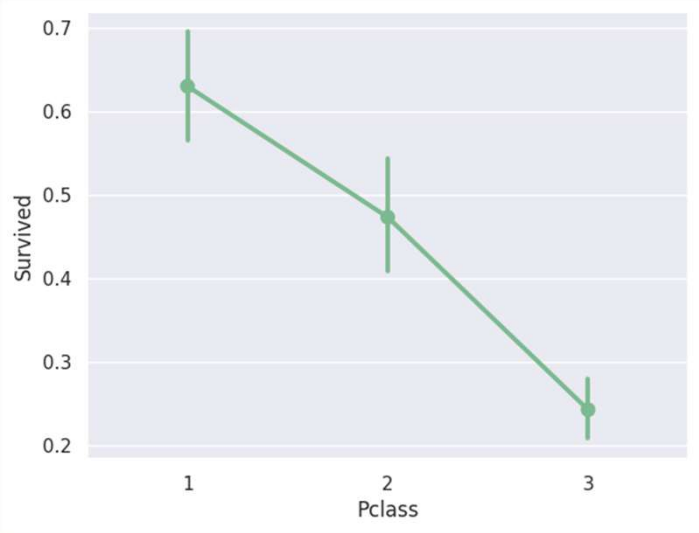


4. Titanic Data

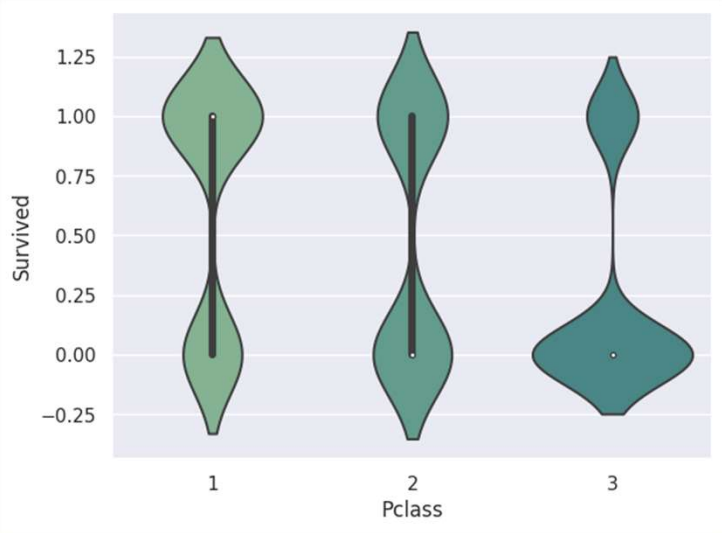
데이터 형태 파악 · 각 변수 타입 파악 · 결측치/이상치 확인 · 종속변수 분포 확인 · 다른 변수들의 분포, 종속변수와의 관계 확인

- 데이터 타입/보고싶은 정보에 따라 다양한 플롯 활용 - 정수형

✓ point plot



✓ violin plot

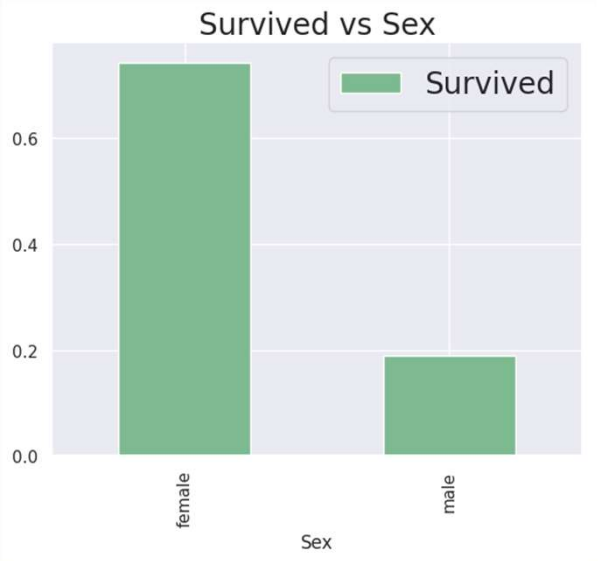


4. Titanic Data

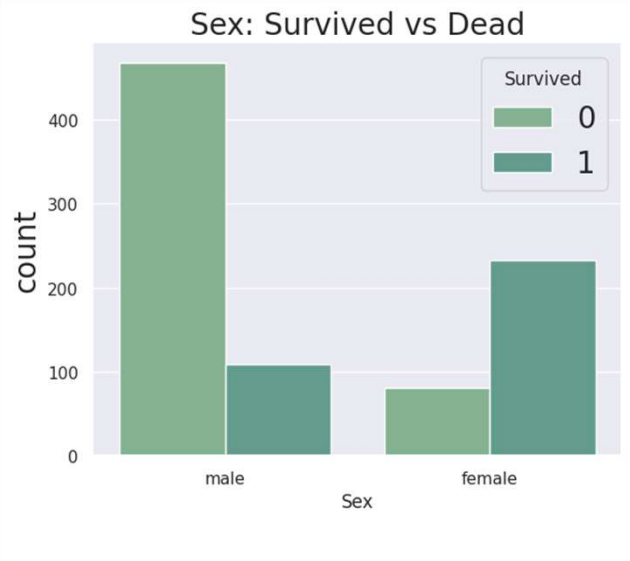
데이터 형태 파악 → 각 변수 타입 파악 → 결측치/이상치 확인 → 종속변수 분포 확인 → 다른 변수들의 분포, 종속변수와의 관계 확인

- 데이터 타입/보고싶은 정보에 따라 다양한 플롯 활용 - 문자형

✓ bar plot



✓ count plot

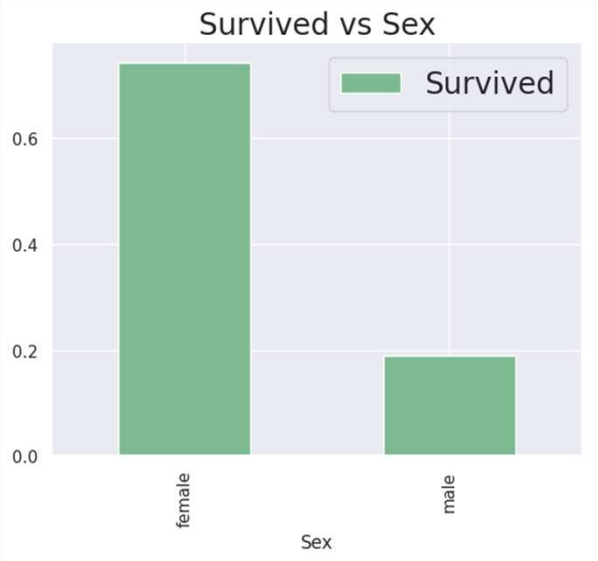


4. Titanic Data

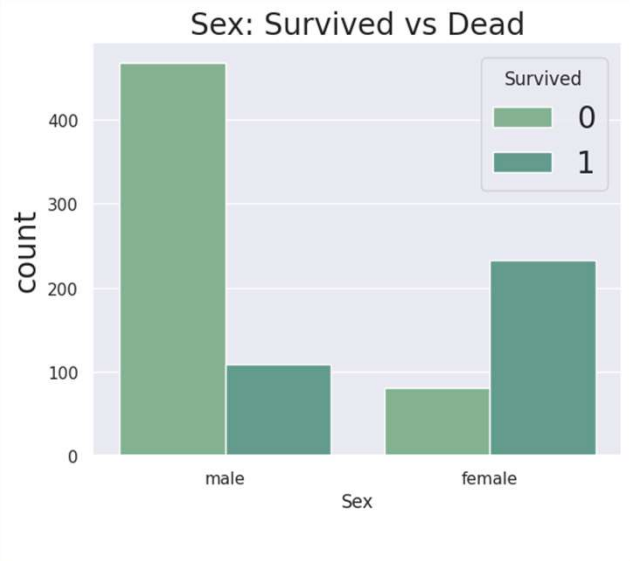
데이터 형태 파악 → 각 변수 타입 파악 → 결측치/이상치 확인 → 종속변수 분포 확인 → 다른 변수들의 분포, 종속변수와의 관계 확인

- 데이터 타입/보고싶은 정보에 따라 다양한 플롯 활용 - 문자형

✓ bar plot



✓ count plot

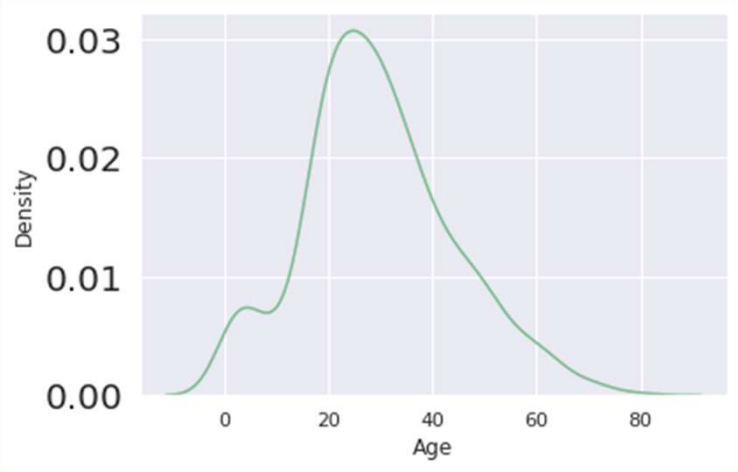


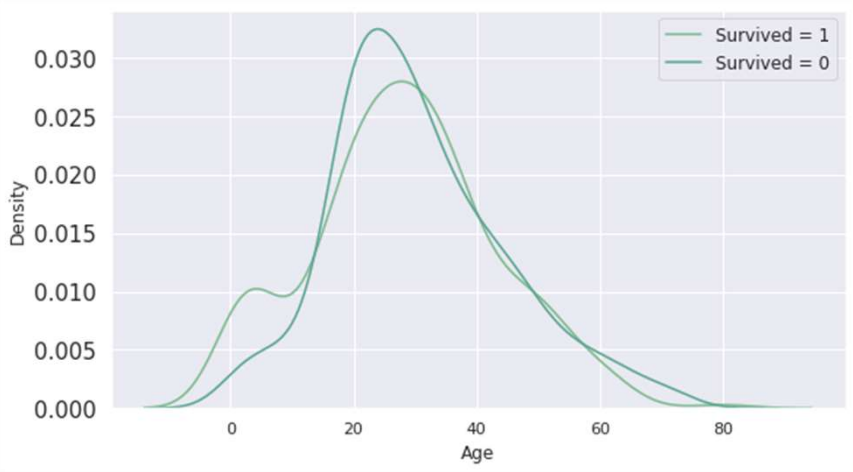
4. Titanic Data

데이터 형태 파악 · 각 변수 타입 파악 · 결측치/이상치 확인 · 종속변수 분포 확인 · **다른 변수들의 분포, 종속변수와의 관계 확인**

- **데이터 타입/보고싶은 정보에 따라 다양한 플롯 활용 - 실수형**

  ✓ kde plot

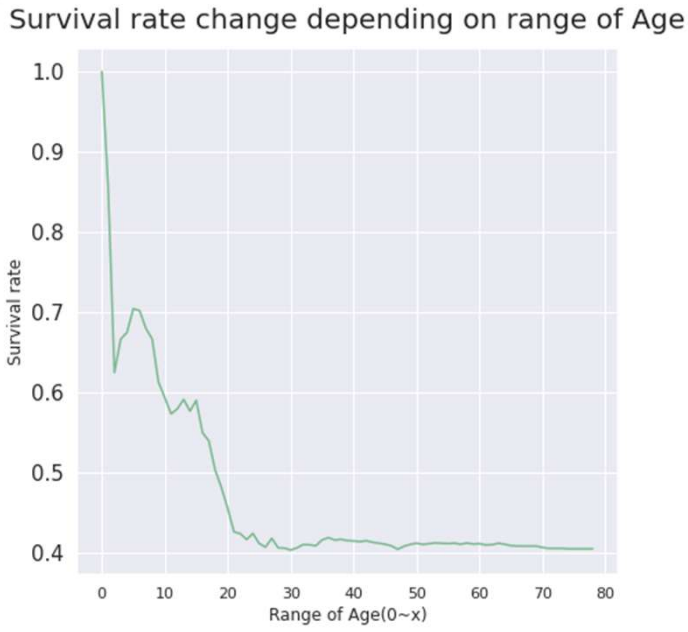


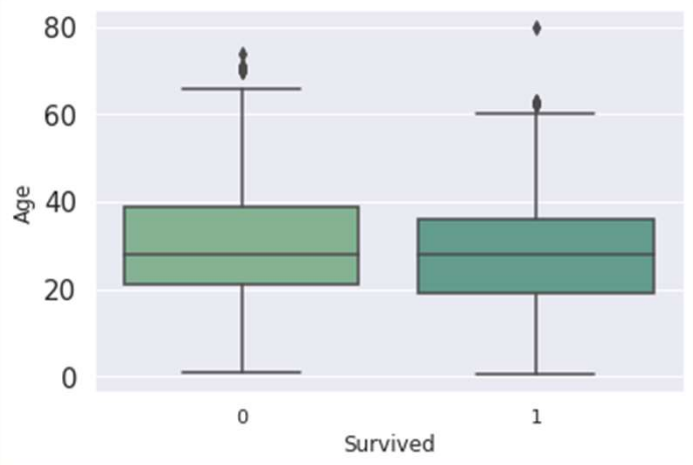


4. Titanic Data

데이터 형태 파악 각 변수 타입 파악 결측치/이상치 확인 종속변수 분포 확인 다른 변수들의 분포, 종속변수와의 관계 확인

- 데이터 타입/보고싶은 정보에 따라 다양한 플롯 활용 - 실수형



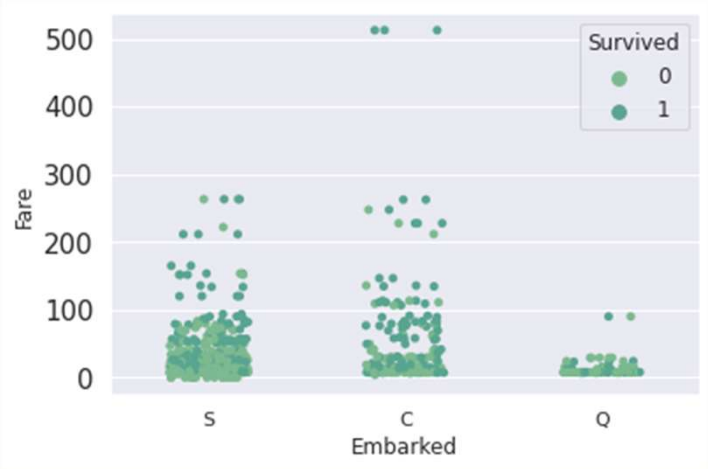


4. Titanic Data

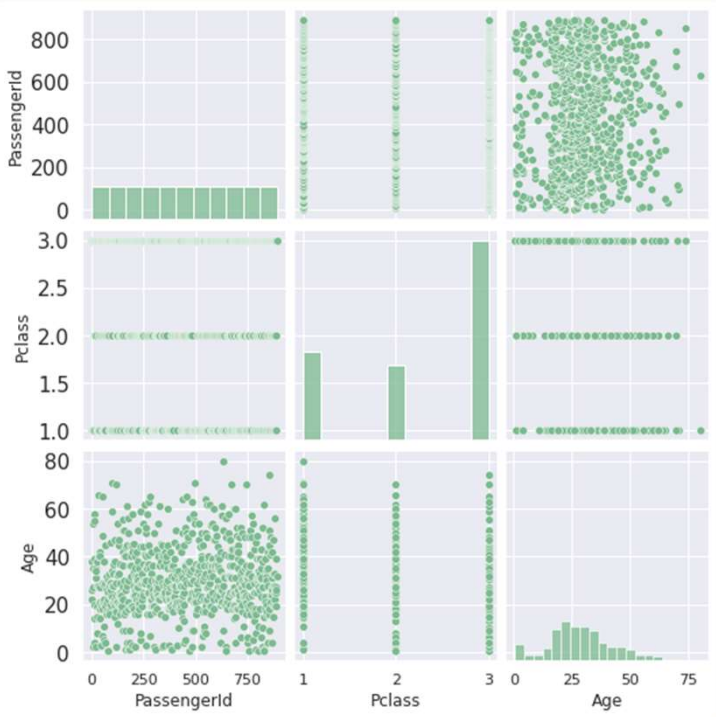
데이터 형태 파악 → 각 변수 타입 파악 → 결측치/이상치 확인 → 종속변수 분포 확인 → **다른 변수들의 분포, 종속변수와의 관계 확인**

- **여러 변수간의 관계**

✓ strip plot



✓ pair plot

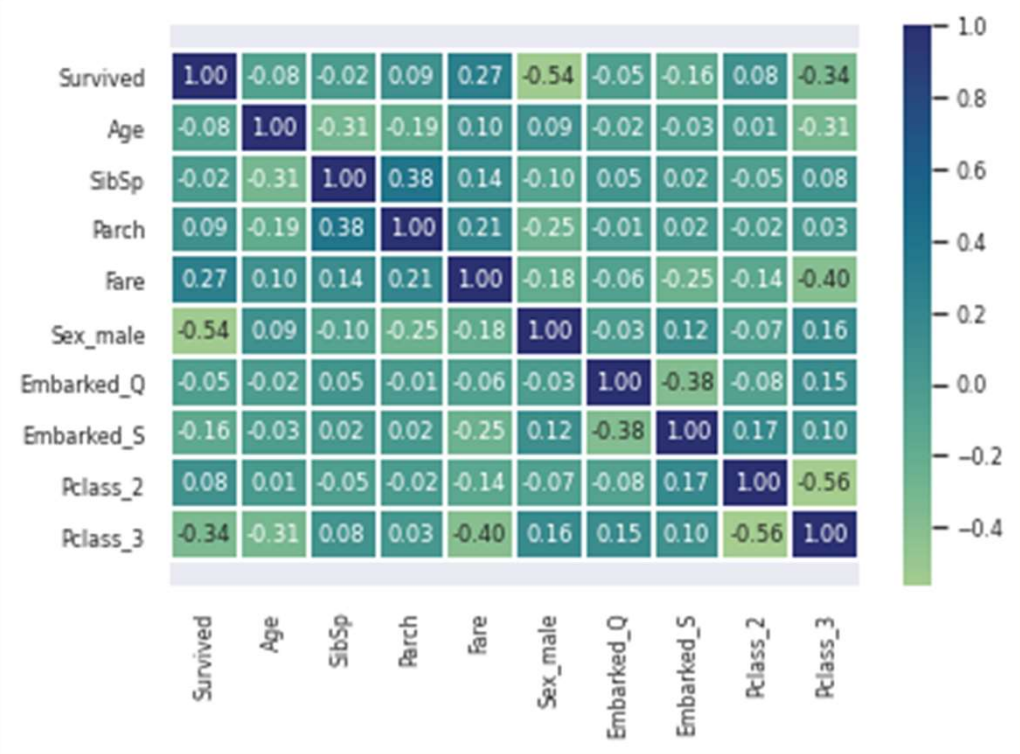


4. Titanic Data

데이터 형태 파악 각 변수 타입 파악 결측치/이상치 확인 종속변수 분포 확인 다른 변수들의 분포, 종속변수와의 관계 확인

- 여러 변수간의 관계

✓ heatmap



감사합니다

DA팀 오세요