탐색적 데이터 분석

충북대학교 소프트웨어학과
류관희

목 차

01

데이터 분석

- 데이터분석 모델
- 탐색적데이터 분석
- 타이타닉 데이터 이해
- EDA 데이터 분석 사이클

02

탐색적 데이터 분석

- 타이타닉 데이터 사례 I

03

탐색적 데이터 분석

- 타이타닉 데이터 사례 II

학습목표

이번 파트에서는 탐색적 데이터 분석을 위한 다음과 과정을 공부한다.

- 데이터분석 모델
- 탐색적데이터 분석
- 타이타닉 데이터 이해
- EDA 데이터 분석 사이클

데이터 분석 모델

- 탐색적 데이터 분석
- 통계적 데이터 분석
- 머신러닝
  - Supervised 러닝: 회귀,분류
  - Unsupervised 러닝: 군집화
- 딥러닝
  - Artificial neural network(ANN): Perceptron, Multilayer perceptron
  - Convolution neural network(CNN)
  - Recurrent neural network(RNN)

탐색적 데이터 분석

- 분석 내용
  - 타이타닉에 탑승한 사람들의 신상정보를 활용하여, 승선한 사람들의 생존여부를 예측하는 모델을 생성
- 탐색적 데이터 분석
  - EDA: Exploratory data analysis
  - 여러 feature 들을 개별적으로 분석하고, feature들 간의 상관관계를 확인. 여러 시각화 툴을 사용하여 insight를 얻은 과정.

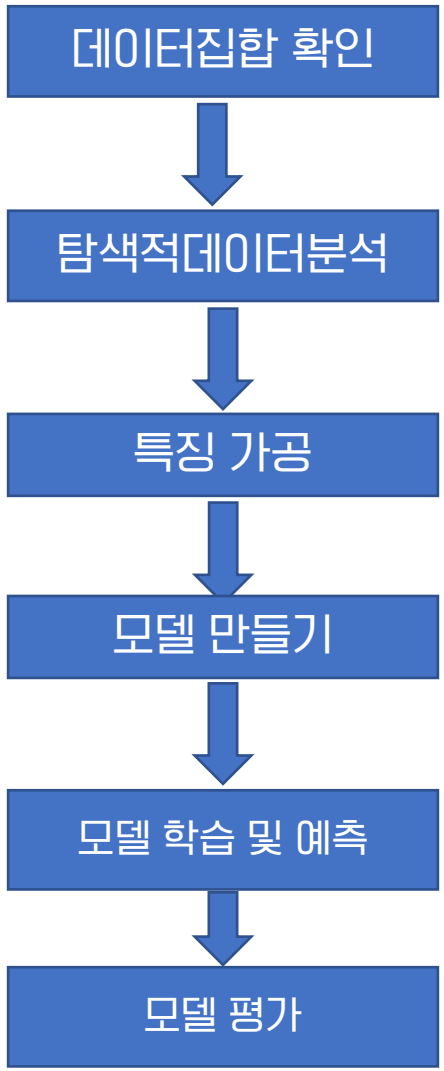
타이타닉 데이터 이해


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('fivethirtyeight')
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

```
data=pd.read_excel('titanic.xlsx')
data.head()
```

- survived : 생존 여부
- pclass : 승객의 클래스
- sex : 성별. male, female로 표기
- sibsp : 형제 혹은 자매의 수
- parch : 부모 혹은 자녀의 수
- fare : 탑승 요금
- embarked : 출발지의 고유 이니셜
- class : 선실의 클래스
- who : male, female을 man, woman으로 표기
- adult_male : 성인 남성 인지 아닌지 여부
- deck : 선실 고유 번호의 가장 앞자리 알파벳(A ~ G)
- embark_town : 출발지
- alive : 생존 여부 데이터를 yes 혹은 no로 표기
- alone : 가족이 없는 경우 True

EDA 기반 데이터 분석 사이클



모델의 성능을 높일 수 있도록 특징을 가공
one-hot encoding, class로 나누기,
구간으로 나누기, 텍스트 데이터 처리 등

데이터시각화

타이타닉 데이터 이해


```
1 data=pd.read_excel('titanic.xlsx')
2 data
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S	Second	man	True	NaN	Southampton	no	True
887	1	1	female	19.0	0	0	30.0000	S	First	woman	False	B	Southampton	yes	True
888	0	3	female	NaN	1	2	23.4500	S	Third	woman	False	NaN	Southampton	no	False
889	1	1	male	26.0	0	0	30.0000	C	First	man	True	C	Cherbourg	yes	True
890	0	3	male	32.0	0	0	7.7500	Q	Third	man	True	NaN	Queenstown	no	True

891 rows × 15 columns

데이터 변수 유형 확인

변수 (feature, variable)	정의	설명	타입
survival	생존여부	target label 임. 1, 0 으로 표현됨	integer
Pclass	티켓의 클래스	1 = 1st, 2 = 2nd, 3 = 3rd 클래스로 나뉘며 categorical feature	integer
sex	성별	male, female 로 구분되며 binary	string
Age	나이	continuous	integer
sibSp	함께 탑승한 형제와 배우자의 수	quantitative	integer
parch	함께 탑승한 부모, 아이의 수	quantitative	integer
ticket	티켓 번호	alphabat + integer	string
fare	탑승료	continuous	float
cabin	객실 번호	alphabat + integer	string
embared	탑승 항구	C = Cherbourg, Q = Queenstown, S = Southampton	string

NULL 데이터 점검

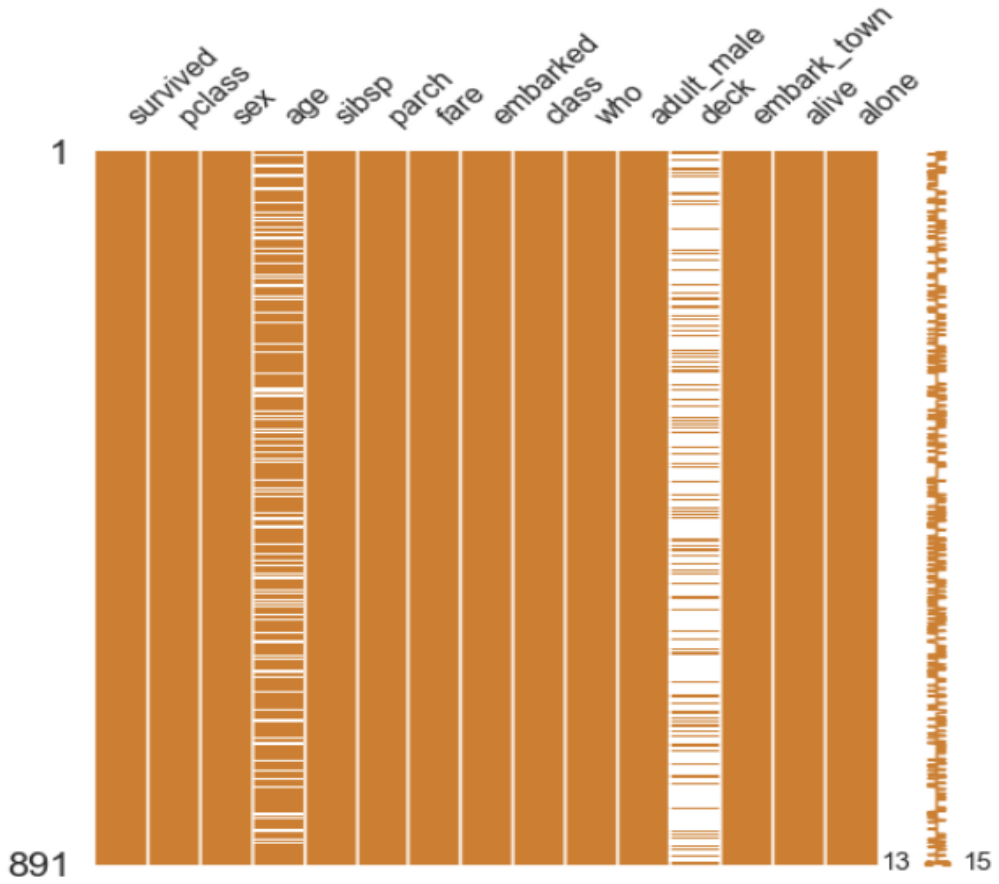

```python
for col in titanic_data.columns:
    msg = 'column: {:>10}₩t Percent of NaN value: {:.2f}%'.format(col,
                100 * (titanic_data[col].isnull().sum() / titanic_data[col].shape[0]))
    print(msg)
```

```
column:   survived       Percent of NaN value: 0.00%
column:     pclass       Percent of NaN value: 0.00%
column:        sex       Percent of NaN value: 0.00%
column:        age       Percent of NaN value: 19.87%
column:      sibsp       Percent of NaN value: 0.00%
column:      parch       Percent of NaN value: 0.00%
column:       fare       Percent of NaN value: 0.00%
column:   embarked       Percent of NaN value: 0.22%
column:      class       Percent of NaN value: 0.00%
column:        who       Percent of NaN value: 0.00%
column: adult_male       Percent of NaN value: 0.00%
column:       deck       Percent of NaN value: 77.22%
column: embark_town      Percent of NaN value: 0.22%
column:      alive       Percent of NaN value: 0.00%
column:      alone       Percent of NaN value: 0.00%
```

NULL 데이터 확인

```
1  msno.matrix(df=titanic_data.iloc[:, :], figsize=(8, 8), color=(0.8, 0.5, 0.2))
```

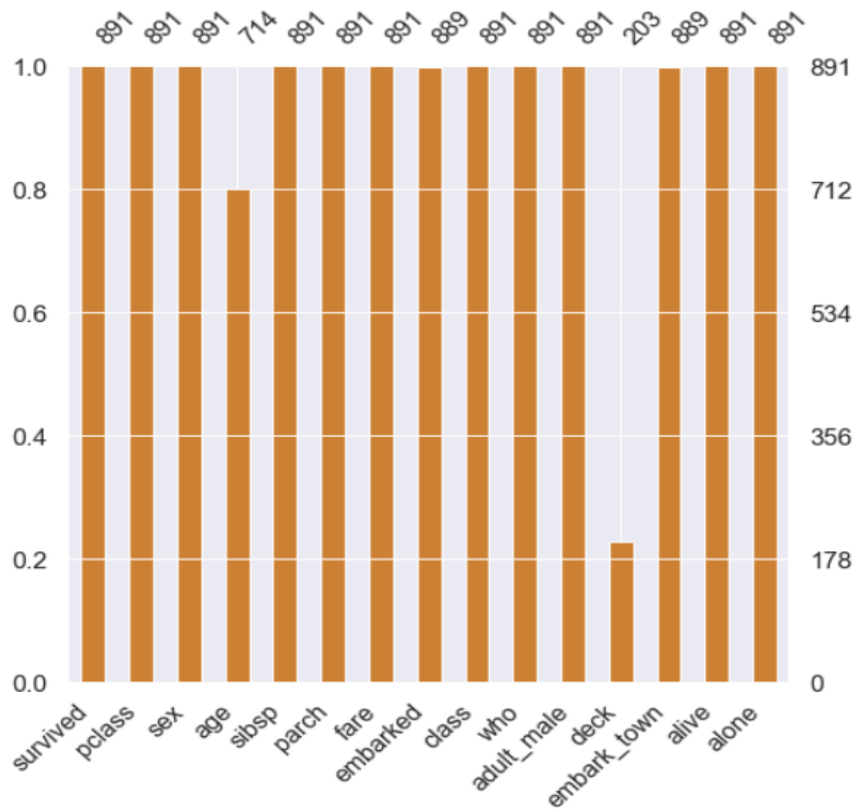
```
<AxesSubplot:>
```



NULL 데이터 확인

```
1  msno.bar(df=titanic_data.iloc[:, :], figsize=(8, 8), color=(0.8, 0.5, 0.2))
```

```
<AxesSubplot:>
```



문제풀이

- 탐색적 데이터 분석 사이클을 설명하시오.

- 탐색적 데이터 분석에서 우선적으로 꼭 점검해야 하는 사항을 설명하시오.

요약

- 데이터 분석 과정에서 탐색적 데이터 분석의 중요성과 과정을 공부하였음

- 탐색적 데이터 분석에서 우선적으로 꼭 점검해야 하는 NULL 데이터를 처리하는 방법을 공부하였음.

01

데이터 분석

- 데이터분석 모델
- 탐색적데이터 분석
- 타이타닉 데이터 이해
- EDA 데이터 분석 사이클

02

탐색적 데이터 분석

- 타이타닉 데이터 사례 I

03

탐색적 데이터 분석

- 타이타닉 데이터 사례 II

학습목표

충북대학교
SW중심대학사업단

이번 파트에서는 탐색적 데이터 분석 사례 공부한다.

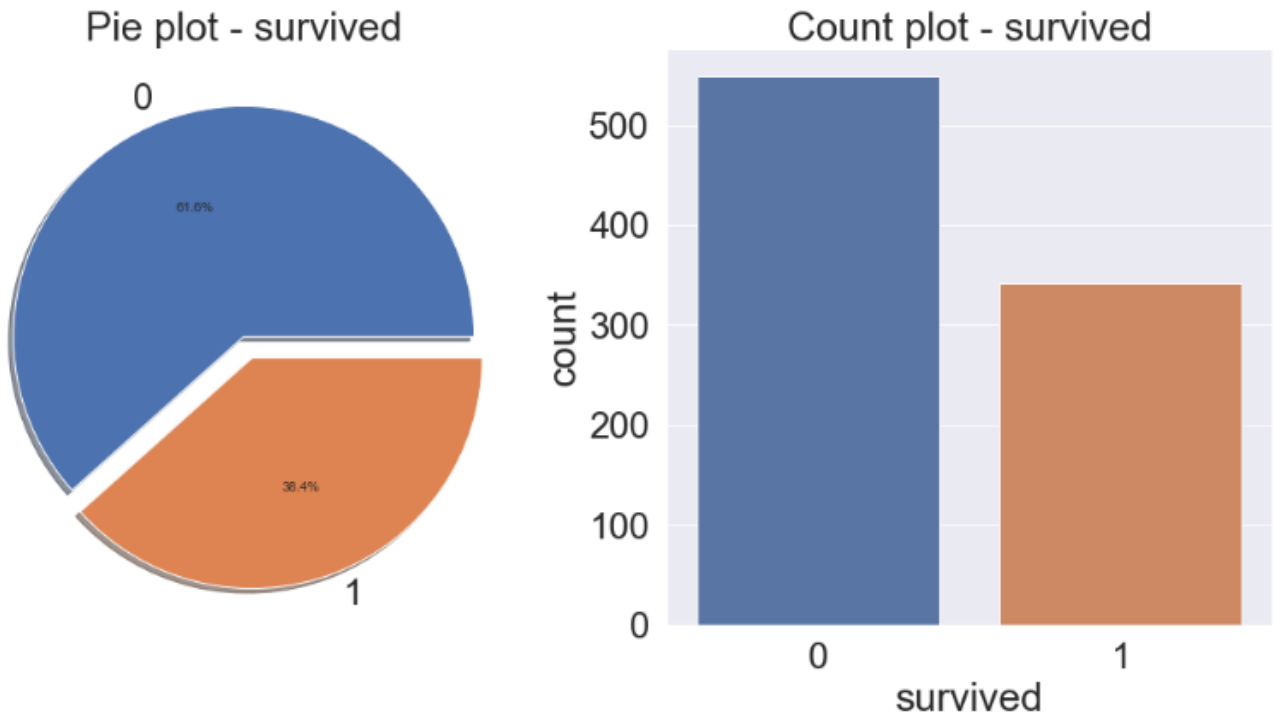
- 타이타닉 데이터 사례 I

EDA-분석목표(생존여부) 확인

```
f, ax = plt.subplots(1, 2, figsize=(18, 8))

titanic_data['survived'].value_counts().plot.pie(explode=[0, 0.1],
                                                 autopct='%1.1f%%', ax=ax[0], shadow=True)
ax[0].set_title('Pie plot - survived')
ax[0].set_ylabel('')
sns.countplot('survived', data=titanic_data, ax=ax[1])
ax[1].set_title('Count plot - survived')

plt.show()
```



EDA-pclass에 대한 분석

- pclass 분석
  - 승객의 클래스: ordinal, 서수형 데이터.
  - 카테고리이면서, 순서가 있는 데이터 타입

- 분석 목표와 pclass와의 관계 확인

```
titanic_data[['pclass', 'survived']].groupby(['pclass'],
                                             as_index=True).count()
```

	survived
pclass	
1	216
2	184
3	491

EDA-pclass에 대한 분석

```
pd.crosstab(titanic_data['pclass'], titanic_data['survived'],
            margins=True).style.background_gradient(cmap='summer_r')
```

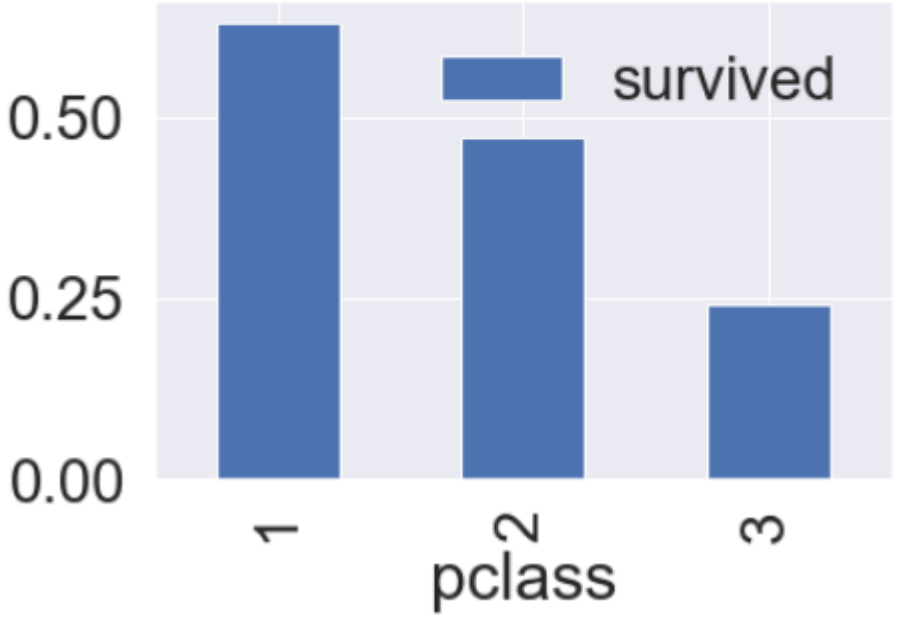
survived	0	1	All
pclass			
1	80	136	216
2	97	87	184
3	372	119	491
All	549	342	891

EDA-pclass에 대한 분석

- 각 pclass 별 생존율을 확인

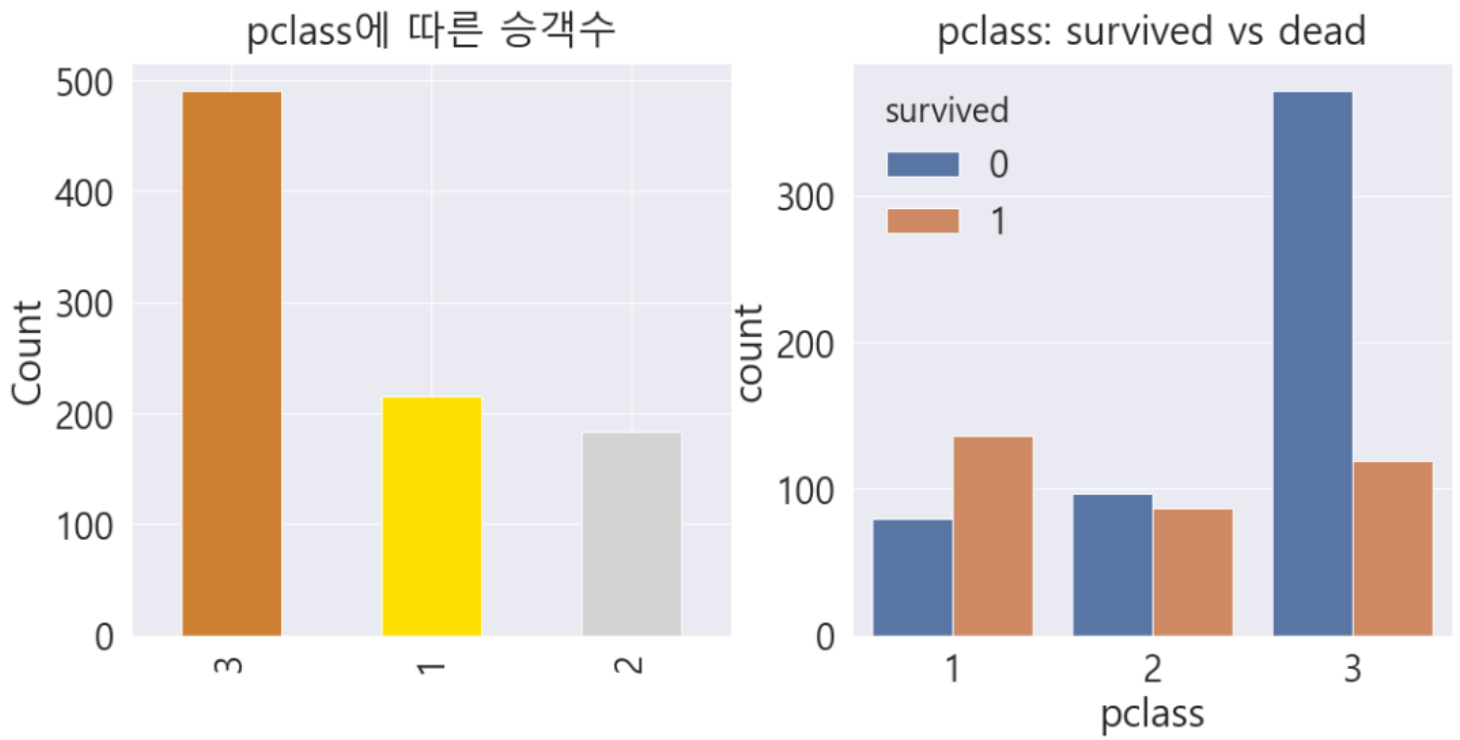
```
titanic_data[['pclass', 'survived']].groupby(['pclass'],
            as_index=True).mean().sort_values(by='survived',
            ascending=False).plot.bar()
```

```
<AxesSubplot:xlabel='pclass'>
```



EDA-pclass에 대한 분석

```
y_position = 1.02
f, ax = plt.subplots(1, 2, figsize=(18, 8))
titanic_data['pclass'].value_counts().plot.bar(color=['#CD7F32','#FFDF00','#D3D3D3'], ax=ax[0])
ax[0].set_title('pclass에 따른 승객수', y=y_position)
ax[0].set_ylabel('Count')
sns.countplot('pclass', hue='survived', data=titanic_data, ax=ax[1])
ax[1].set_title('pclass: survived vs dead', y=y_position)
plt.show()
```



한글 깨지는 문제 해결


```
import matplotlib as mpl
import matplotlib.font_manager as fm
# 그래프에서 마이너스 폰트 깨지는 문제에 대한 대처
mpl.rcParams['axes.unicode_minus'] = False
```

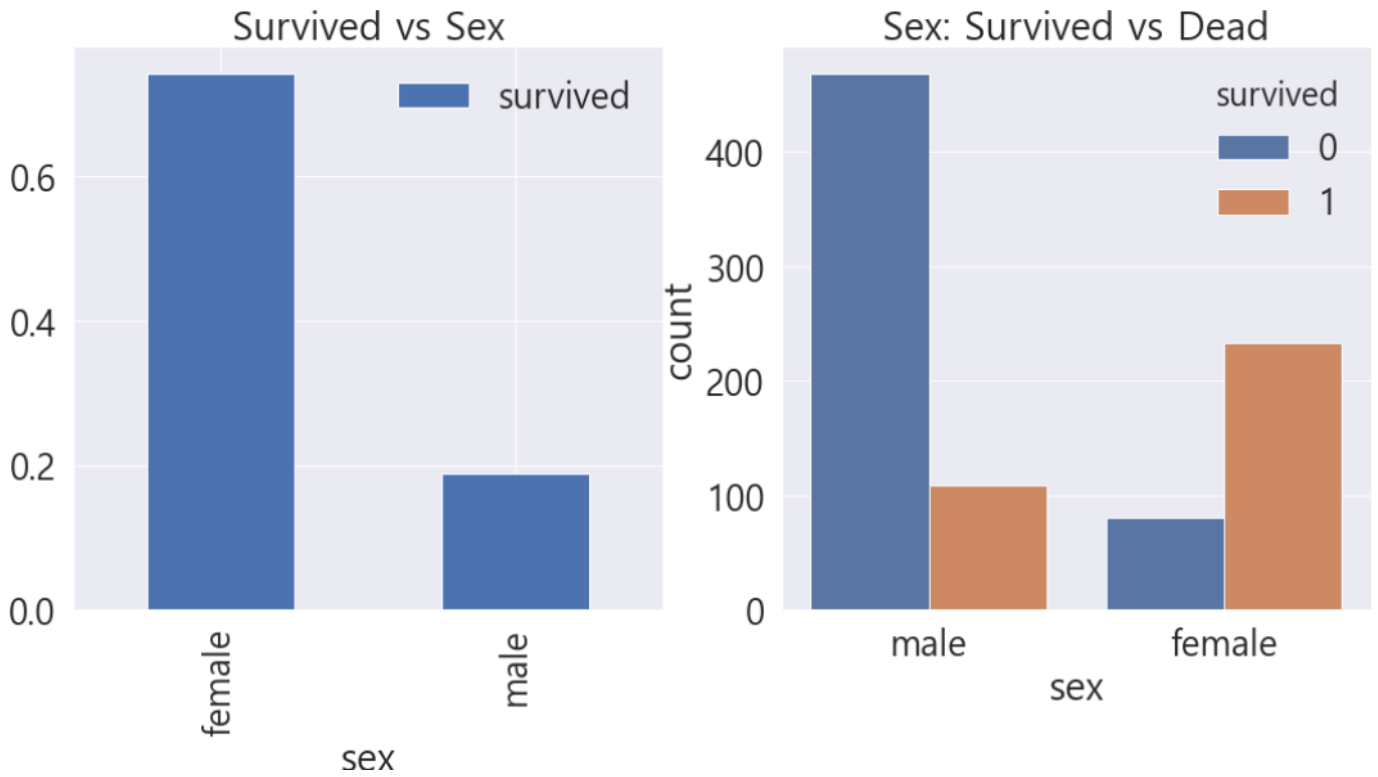
```
#plt.rcParams["font.family"] = 'Nanum Brush Script OTF'
#plt.rcParams["font.size"] = 20
plt.rcParams["figure.figsize"] = (20,10)

plt.rc('font', family='NanumGothic') # For Windows
plt.rcParams['font.family'] = 'Malgun Gothic'
plt.rcParams.update({'font.size': 15})
print(plt.rcParams['font.family'])
```

```
['Malgun Gothic']
```

EDA-sex 데이터에 대한 분석

```
f, ax = plt.subplots(1, 2, figsize=(18, 8))
titanic_data[['sex', 'survived']].groupby(['sex'], as_index=True).mean().plot.bar(ax=ax[0])
ax[0].set_title('Survived vs Sex')
sns.countplot('sex', hue='survived', data=titanic_data, ax=ax[1])
ax[1].set_title('Sex: Survived vs Dead')
plt.show()
```

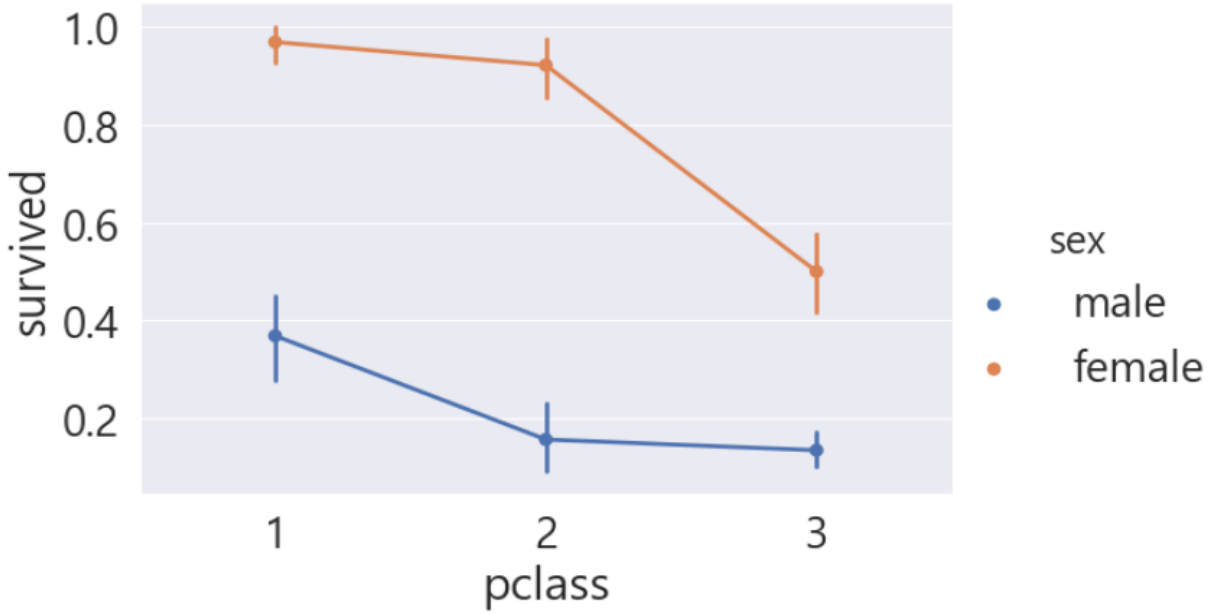


EDA-pclass, sex 데이터 분석

- 모든 클래스에서 female 이 살 확률이 male 보다 높음.
- 또한 남자, 여자 상관없이 클래스가 높을 수록 살 확률 높음.

```
1 sns.factorplot('pclass', 'survived', hue='sex',
2                data=titanic_data, size=6, aspect=1.5)
```

```
<seaborn.axisgrid.FacetGrid at 0x23f9343ce48>
```



https://kaggle-kr.tistory.com/17

문제풀이

- 탐색적 데이터 분석에서 가장 중요한 초점이 무엇인가요?
- 타이타닉 데이터를 가지고 분석하는 목적을 설명하시오.

요약

- 타이타닉 데이터를 가지고 탐색적 데이터 분석 사례를 구체적으로 공부하였음
  - 목적 기반의 데이터 분석

01

데이터 분석

- 데이터분석 모델
- 탐색적데이터 분석
- 타이타닉 데이터 이해
- EDA 데이터 분석 사이클

02

탐색적 데이터 분석

- 타이타닉 데이터 사례 I

03

탐색적 데이터 분석

- 타이타닉 데이터 사례 II

학습목표

충북대학교
SW중심대학사업단

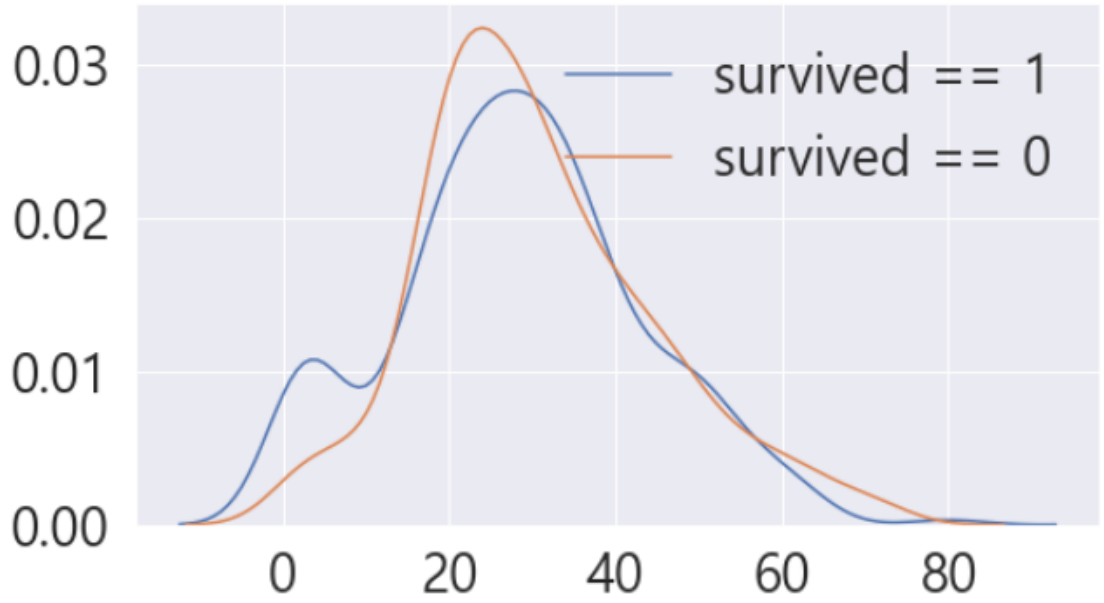
이번 파트에서는 탐색적 데이터 분석 사례 공부한다.

- 타이타닉 데이터 사례 II

EDA-age 데이터 분석

- 생존여부에 따른 나이 분포

```
fig, ax = plt.subplots(1, 1, figsize=(9, 5))
sns.kdeplot(titanic_data[titanic_data['survived'] == 1]['age'], ax=ax)
sns.kdeplot(titanic_data[titanic_data['survived'] == 0]['age'], ax=ax)
plt.legend(['survived == 1', 'survived == 0'])
plt.show()
```



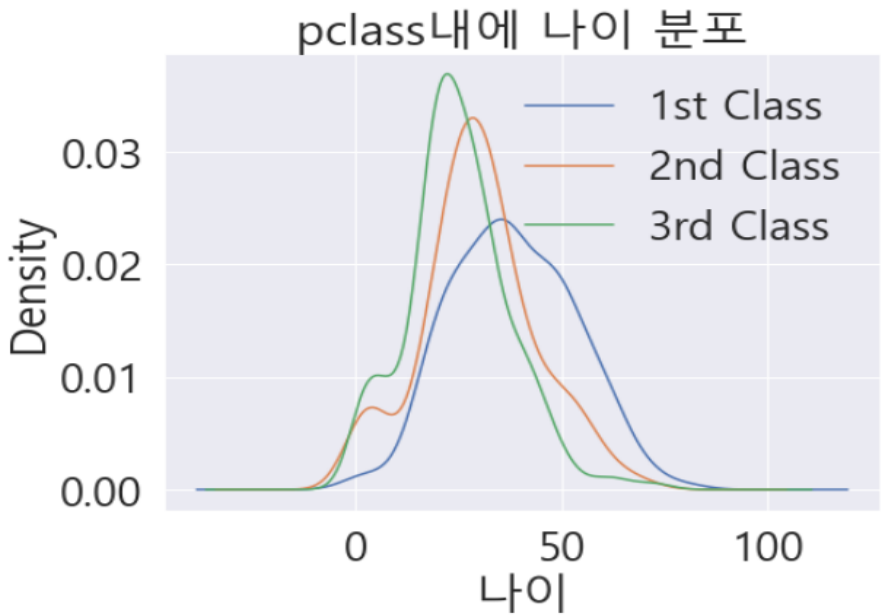
EDA-age 데이터 분석

- pclass 내에 나이 분포

```
# Age distribution withing classes
plt.figure(figsize=(8, 6))
titanic_data['age'][titanic_data['pclass'] == 1].plot(kind='kde')
titanic_data['age'][titanic_data['pclass'] == 2].plot(kind='kde')
titanic_data['age'][titanic_data['pclass'] == 3].plot(kind='kde')

plt.xlabel('나이')
plt.title('pclass내에 나이 분포')
plt.legend(['1st Class', '2nd Class', '3rd Class'])
```

```
<matplotlib.legend.Legend at 0x211d09e3710>
```

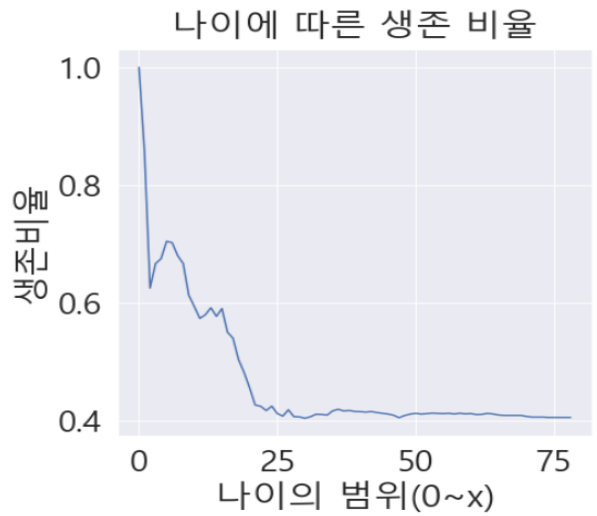


EDA-age 데이터 분석

- 나이범위에 따른 생존비율

```
cummulate_survival_ratio = []
for i in range(1, 80):
    cummulate_survival_ratio.append(
        titanic_data[titanic_data['age'] < i]['survived'].sum()
        / len(titanic_data[titanic_data['age'] < i]['survived']))

plt.figure(figsize=(7, 7))
plt.plot(cummulate_survival_ratio)
plt.title('나이에 따른 생존 비율', y=1.02)
plt.ylabel('생존비율')
plt.xlabel('나이의 범위(0~x)')
plt.show()
```



EDA-pclass, sex, age 데이터 분석

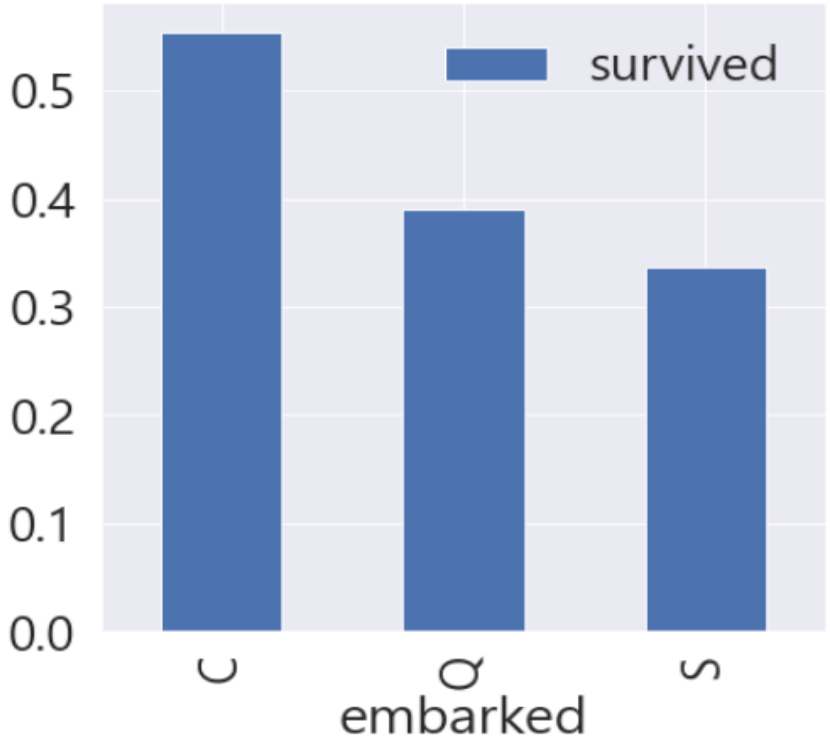
```
f,ax=plt.subplots(1,2,figsize=(18,8))
sns.violinplot("pclass","age", hue="survived",
               data=titanic_data, scale='count', split=True,ax=ax[0])
ax[0].set_title('pclass and age vs survived')
ax[0].set_yticks(range(0,110,10))
sns.violinplot("sex","age", hue="survived",
               data=titanic_data, scale='count', split=True,ax=ax[1])
ax[1].set_title('sex and age vs survived')
ax[1].set_yticks(range(0,110,10))
plt.show()
```

EDA-emarked 데이터 분석

- 탑승한 항구에 따른 생존비율

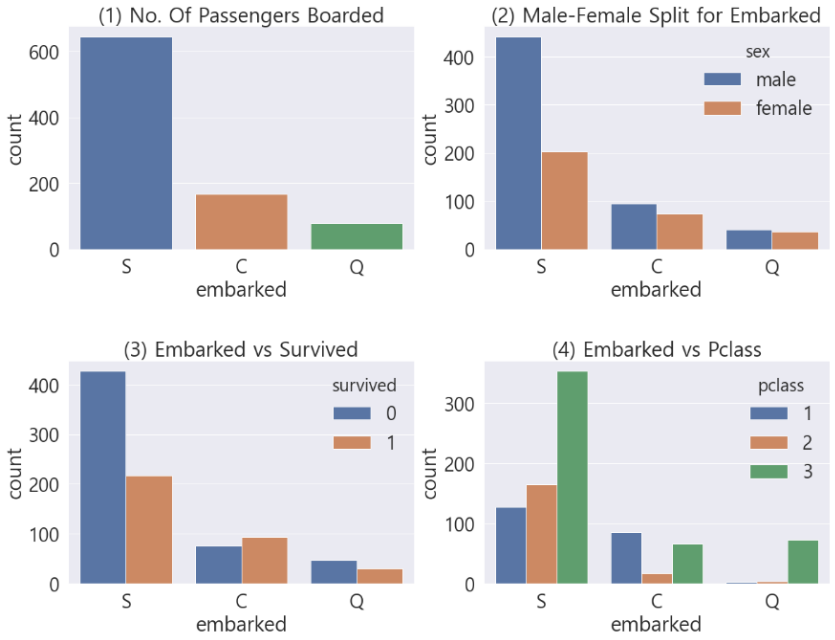
```
1  f, ax = plt.subplots(1, 1, figsize=(7, 7))
2  titanic_data[['embarked', 'survived']].groupby(['embarked'],
3          as_index=True).mean().sort_values(by='survived',
4          ascending=False).plot.bar(ax=ax)
```

```
<AxesSubplot:xlabel='embarked'>
```



EDA-emarked 데이터 분석

```
f,ax=plt.subplots(2, 2, figsize=(20,15))
sns.countplot('embarked', data=titanic_data, ax=ax[0,0])
ax[0,0].set_title('(1) No. Of Passengers Boarded')
sns.countplot('embarked', hue='sex', data=titanic_data, ax=ax[0,1])
ax[0,1].set_title('(2) Male-Female Split for Embarked')
sns.countplot('embarked', hue='survived', data=titanic_data, ax=ax[1,0])
ax[1,0].set_title('(3) Embarked vs Survived')
sns.countplot('embarked', hue='pclass', data=titanic_data, ax=ax[1,1])
ax[1,1].set_title('(4) Embarked vs Pclass')
plt.subplots_adjust(wspace=0.2, hspace=0.5)
plt.show()
```



EDA-sibsp, parch데이터 분석

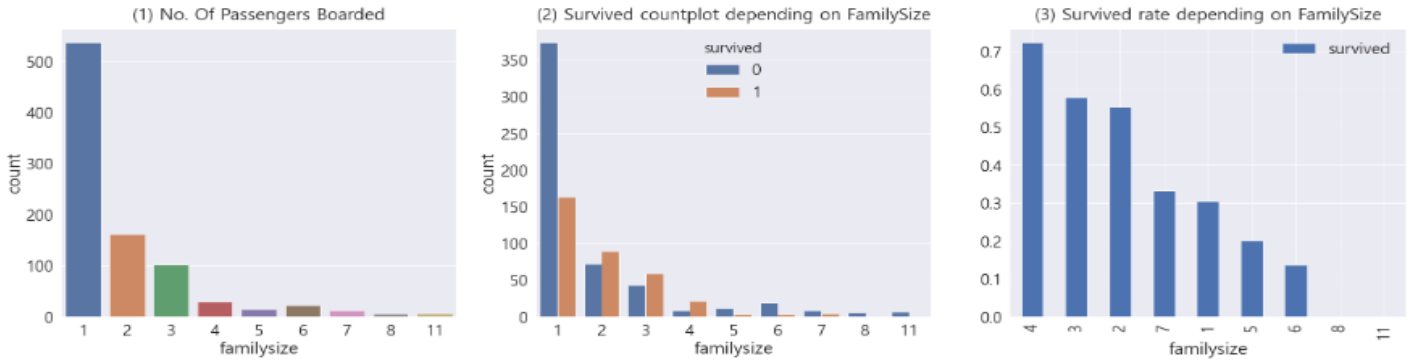
titanic_data['familysize'] = titanic_data['sibsp'] + titanic_data['parch'] + 1

```
f,ax=plt.subplots(1, 3, figsize=(40,10))
sns.countplot('familysize', data=titanic_data, ax=ax[0])
ax[0].set_title('(1) No. Of Passengers Boarded', y=1.02)

sns.countplot('familysize', hue='survived', data=titanic_data, ax=ax[1])
ax[1].set_title('(2) Survived countplot depending on FamilySize',  y=1.02)

titanic_data[['familysize', 'survived']].groupby(['familysize'],
            as_index=True).mean().sort_values(by='survived', ascending=False).plot.bar(ax=ax[2])
ax[2].set_title('(3) Survived rate depending on FamilySize',  y=1.02)

plt.subplots_adjust(wspace=0.2, hspace=0.5)
plt.show()
```



문제풀이

- 이번 파트에서 배운 탐색적 데이터 분석에서 가장 중요한 초점이 무엇인가요?

- 이번 파트에서 타이타닉 데이터를 가지고 탐색적 데이터 분석 항목을 설명하시오.

요약

- 타이타닉 데이터를 가지고 탐색적 데이터 분석 사례를 구체적으로 공부하였음
  - 생존(종속변수): 다양한 데이터(독립변수)