데이터프레임 복습

1.환경준비

(1) 라이브러리 불러오기

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

(2) 데이터 로딩

1) titanic

- url : 'https://raw.githubusercontent.com/DA4BAM/dataset/master/titanic_simple.csv'

[titanic_simple 데이터 셋 정보]

- PassengerId : 승객번호
- Survived : 생존여부(1:생존, 0:사망)
- Pclass : 객실등급(1:1등급, 2:2등급, 3:3등급)
- Name : 승객이름
- Sex : 성별(male, female)
- Age : 나이
- Fare : 운임($)
- Embarked : 승선지역(Southampton, Cherbourg, Queenstown)

In [3]:
```
path = 'https://raw.githubusercontent.com/DA4BAM/dataset/master/titanic_simple.csv'
titanic = pd.read_csv(path)
titanic.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	7.2500	Southampton
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	71.2833	Cherbourg
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	7.9250	Southampton
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	53.1000	Southampton
4	5	0	3	Allen, Mr. William Henry	male	35.0	8.0500	Southampton

2) New York Air Quality

- url : https://raw.githubusercontent.com/DA4BAM/dataset/master/air2.csv

[airquality 데이터 셋 정보]

- Ozone: 오존 농도
- Solar.R: 태양복사량
- Wind: 풍속
- Temp: 기온
- Date : 연,월,일

In [4]:
```
path = 'https://raw.githubusercontent.com/DA4BAM/dataset/master/air2.csv'
air = pd.read_csv(path)
air.head()
```

Out[4]:

	Ozone	Solar.R	Wind	Temp	Date
0	41	190.0	7.4	67	1973-05-01
1	36	118.0	8.0	72	1973-05-02
2	12	149.0	12.6	74	1973-05-03
3	18	313.0	11.5	62	1973-05-04
4	19	NaN	14.3	56	1973-05-05

2.데이터프레임 정보 조회하기

- 두 데이터프레임에 대해서 다음의 정보를 조회해 봅시다.

(1) 상위 5개 행 조회

In [5]:
```
titanic.head()
```

Out[5]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	7.2500	Southampton
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	71.2833	Cherbourg
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	7.9250	Southampton
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	53.1000	Southampton
4	5	0	3	Allen, Mr. William Henry	male	35.0	8.0500	Southampton

In [6]:
```
air.head()
```

Out[6]:

	Ozone	Solar.R	Wind	Temp	Date
0	41	190.0	7.4	67	1973-05-01
1	36	118.0	8.0	72	1973-05-02
2	12	149.0	12.6	74	1973-05-03
3	18	313.0	11.5	62	1973-05-04
4	19	NaN	14.3	56	1973-05-05

(2) 행과 열의 수

In [7]:
```
titanic.shape
```

Out[7]:
```
(891, 8)
```

In [8]:
```
air.shape
```

Out[8]:
```
(153, 5)
```

(3) 칼럼 정보

In [9]:
```
titanic.columns
```

Out[9]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'Fare',
       'Embarked'],
      dtype='object')
```

In [10]:
```
air.columns
```

Out[10]:
```
Index(['Ozone', 'Solar.R', 'Wind', 'Temp', 'Date'], dtype='object')
```

(4) 칼럼 이름만 리스트에 담아 조회

In [11]:
```
titanic_columns_names = titanic.columns
titanic_columns_names
```

Out[11]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'Fare',
       'Embarked'],
      dtype='object')
```

In [12]:
```
air_columns_names = air.columns
air_columns_names
```

Out[12]:
```
Index(['Ozone', 'Solar.R', 'Wind', 'Temp', 'Date'], dtype='object')
```

(5) 데이터프레임 각 열 들의 기초통계량 조회

In [13]:
```
titanic.describe().T
```

Out[13]:

	count	mean	std	min	25%	50%	75%	max
PassengerId	891.0	446.000000	257.353842	1.00	223.5000	446.0000	668.5	891.0000
Survived	891.0	0.383838	0.486592	0.00	0.0000	0.0000	1.0	1.0000
Pclass	891.0	2.308642	0.836071	1.00	2.0000	3.0000	3.0	3.0000
Age	714.0	29.699118	14.526497	0.42	20.1250	28.0000	38.0	80.0000
Fare	891.0	32.204208	49.693429	0.00	7.9104	14.4542	31.0	512.3292

In [14]:
```
air.describe().T
```

Out[14]:

	count	mean	std	min	25%	50%	75%	max
Ozone	153.0	42.052288	30.156127	1.0	20.00	34.0	59.00	168.0
Solar.R	146.0	185.931507	90.058422	7.0	115.75	205.0	258.75	334.0
Wind	153.0	9.957516	3.523001	1.7	7.40	9.7	11.50	20.7
Temp	153.0	77.882353	9.465270	56.0	72.00	79.0	85.00	97.0

(6) 데이터프레임에 NaN이 존재하는지 확인

In [15]:
```
titanic.isna().sum()
```

Out[15]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
Fare             0
Embarked         2
dtype: int64
```

In [16]:
```
air.isna().sum()
```

Out[16]:
```
Ozone      0
Solar.R    7
Wind       0
Temp       0
Date       0
dtype: int64
```

3.데이터프레임 조건 조회(.loc)

- 다음 질문에 맞는 조회를 수행하시오.

(1) [titanic] 객실 등급(Pclass) 1등급, 나이(Age) 10살 이하 탑승객 조회

In [19]:
```
titanic.head()
```

Out[19]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	7.2500	Southampton
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	71.2833	Cherbourg
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	7.9250	Southampton
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	53.1000	Southampton
4	5	0	3	Allen, Mr. William Henry	male	35.0	8.0500	Southampton

In [42]:
```
titanic.loc[(titanic['Pclass'] == 1) & (titanic['Age'] <= 10)]
```

Out[42]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked
297	298	0	1	Allison, Miss. Helen Loraine	female	2.00	151.5500	Southampton
305	306	1	1	Allison, Master. Hudson Trevor	male	0.92	151.5500	Southampton
445	446	1	1	Dodge, Master. Washington	male	4.00	81.8583	Southampton

(2) [titanic] 객실 등급(Pclass)별 탑승객 수

In [23]:
```
titanic.groupby(by='Pclass', as_index=False)['PassengerId'].sum()
```

Out[23]:

	Pclass	PassengerId
0	1	99705
1	2	82056
2	3	215625

(3) [titanic] 성별(Sex)이 남자인 탑승객과 여자인 탑승객의 나이를 각각 저장하시오.

In [40]:
```
titanic.loc[(titanic['Sex'] == 'male')]['Age']
```

Out[40]:
```
0      22.0
4      35.0
5       NaN
6      54.0
7       2.0
       ...
883    28.0
884    25.0
886    27.0
889    26.0
890    32.0
Name: Age, Length: 577, dtype: float64
```

In [41]:
```
titanic.loc[(titanic['Sex'] == 'female')]['Age']
```

Out[41]:
```
1      38.0
2      26.0
3      35.0
8      27.0
9      14.0
       ...
880    25.0
882    22.0
885    39.0
887    19.0
888     NaN
Name: Age, Length: 314, dtype: float64
```

(4) [titanic] 나이(Age)에 NaN이 아닌 탑승객을 조회하시오.

In [46]:
```
titanic.loc[(titanic['Age'].notna())]
```

Out[46]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	7.2500	Southampton
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	71.2833	Cherbourg
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	7.9250	Southampton
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	53.1000	Southampton
4	5	0	3	Allen, Mr. William Henry	male	35.0	8.0500	Southampton
...	...	...	...	...	...	...	...	...
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	29.1250	Queenstown
886	887	0	2	Montvila, Rev. Juozas	male	27.0	13.0000	Southampton
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	30.0000	Southampton
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	30.0000	Cherbourg
890	891	0	3	Dooley, Mr. Patrick	male	32.0	7.7500	Queenstown

714 rows × 8 columns

(5) [air quality] 오존 농도 10~20 사이의 데이터를 조회하시오.

In [36]:
```
air.head()
```

Out[36]:

	Ozone	Solar.R	Wind	Temp	Date
0	41	190.0	7.4	67	1973-05-01
1	36	118.0	8.0	72	1973-05-02
2	12	149.0	12.6	74	1973-05-03
3	18	313.0	11.5	62	1973-05-04
4	19	NaN	14.3	56	1973-05-05

In [38]:
```
air.loc[air['Ozone'].between(10, 20)]
```

Out[38]:

	Ozone	Solar.R	Wind	Temp	Date
2	12	149.0	12.6	74	1973-05-03
3	18	313.0	11.5	62	1973-05-04
4	19	NaN	14.3	56	1973-05-05
7	19	99.0	13.8	59	1973-05-08
9	20	194.0	8.6	69	1973-05-10
11	16	256.0	9.7	69	1973-05-12
12	11	290.0	9.2	66	1973-05-13
13	14	274.0	10.9	68	1973-05-14
14	18	65.0	13.2	58	1973-05-15
15	14	334.0	11.5	64	1973-05-16
19	11	44.0	9.7	62	1973-05-20
21	11	320.0	16.6	73	1973-05-22
24	17	66.0	16.6	57	1973-05-25
25	18	266.0	14.9	58	1973-05-26
26	15	NaN	8.0	57	1973-05-27
33	18	242.0	16.1	67	1973-06-03
48	20	37.0	9.2	65	1973-06-18
49	12	120.0	11.5	73	1973-06-19
50	13	137.0	10.3	76	1973-06-20
72	10	264.0	14.3	73	1973-07-12
81	16	7.0	6.9	74	1973-07-21
86	20	81.0	8.6	82	1973-07-26
94	16	77.0	7.4	82	1973-08-03
129	20	252.0	10.9	80	1973-09-07
137	13	112.0	11.5	71	1973-09-15
139	18	224.0	13.8	67	1973-09-17
140	13	27.0	10.3	76	1973-09-18
142	16	201.0	8.0	82	1973-09-20
143	13	238.0	12.6	64	1973-09-21
147	14	20.0	16.6	63	1973-09-25
150	14	191.0	14.3	75	1973-09-28
151	18	131.0	8.0	76	1973-09-29
152	20	223.0	11.5	68	1973-09-30

(6) [air quality] 날짜(Date) 1973-05-01, 1973-06-01, 1973-07-01 , 1973-08-01 을 조회하시오.

In [47]:
```
air.head()
```

Out[47]:

	Ozone	Solar.R	Wind	Temp	Date
0	41	190.0	7.4	67	1973-05-01
1	36	118.0	8.0	72	1973-05-02
2	12	149.0	12.6	74	1973-05-03
3	18	313.0	11.5	62	1973-05-04
4	19	NaN	14.3	56	1973-05-05

In [48]:
```
air.loc[air['Date'].isin(['1973-05-01', '1973-06-01', '1973-08-01'])]
```

Out[48]:

	Ozone	Solar.R	Wind	Temp	Date
0	41	190.0	7.4	67	1973-05-01
31	34	286.0	8.6	78	1973-06-01
92	39	83.0	6.9	81	1973-08-01

4.데이터프레임 값 변경

(1) [titanic] 승선지역(Embarked)을 변경하시오.(.map)

- Southamton --> S
- Cherbourg --> C
- Queenstown --> Q

In [49]:
```
titanic['Embarked'] = titanic['Embarked'].map({'Southamton': 'S', 'Cherbourg': 'C', 'Queenstow
```

In [51]:
```
titanic['Embarked'].value_counts()
```

Out[51]:
```
Embarked
C    168
Q     77
Name: count, dtype: int64
```

(2) [titanic] 운임(Fare)을 다음과 같이 변경하시오.(pd.cut)

- <= 30 ==> 'L'
- <= 100 ==> 'M'
- 100 < ==> 'H'

In [52]:
```
titanic.head()
```

Out[52]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	7.2500	NaN
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	7.9250	NaN
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	53.1000	NaN
4	5	0	3	Allen, Mr. William Henry	male	35.0	8.0500	NaN

In [55]:
```
bin = [-np.inf, 30, 100, np.inf]
label = list('LMH')
titanic['Fare'] = pd.cut(titanic['Fare'], bins=bin, labels=label)
```

In [56]:
```
titanic.head()
```

Out[56]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	L	NaN
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	M	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	L	NaN
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	M	NaN
4	5	0	3	Allen, Mr. William Henry	male	35.0	L	NaN

(3) [titanic] 성별(Sex)을 다음과 같이 변경하시오.(np.where)

- female ==> 0
- male ==> 1

In [63]:
```
titanic['Age2'] = np.where(titanic['Sex'] == 'female', 0, 1)
```

In [65]:
```
titanic.head()
```

Out[65]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Fare	Embarked	Age2
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	L	NaN	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	M	C	0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	L	NaN	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	M	NaN	0
4	5	0	3	Allen, Mr. William Henry	male	35.0	L	NaN	1

In [ ]: