csv 파일 읽기

In [21]:
```
# 관련 라이브러리 불러오기
import pandas as pd            #판다스
import matplotlib as mpl        #맷플롭립
import matplotlib.pylab as plt    #맷플롭립 파이랩
import seaborn as sns           #시본
import numpy as np             #넘파이

#파일경로를 찾고 변수 file_path에 저장(상대경로)
file_path= './산림청 국립자연휴양림관리소_국립자연휴양림 예약 정보_20220316 (3).csv'

#read csv()함수로 데이터프레임 변환
df = pd.read_csv(file_path, encoding='utf-8')
df
```

Out[21]:

	기관아이디	기관이름	상품이름	숙박일자	상태
0	141	덕유산 자연휴양림	야영데크(127)	2021-09-26	체크아웃
1	187	희리산 자연휴양림	607호 수선화	2021-10-04	체크아웃
2	191	오서산 자연휴양림	402호 복숭아꽃	2021-09-24	체크아웃
3	224	운악산 자연휴양림	[A동]꾀꼬리	2021-09-30	체크아웃
4	106	청태산 자연휴양림	분비나무	2021-09-29	체크아웃
...	...	...	...	...	...
142659	220	용현 자연휴양림	206호 창출	2022-03-14	체크아웃
142660	105	신불산 자연휴양림	하단2관 단풍취	2022-03-14	체크아웃
142661	105	신불산 자연휴양림	하단2관 고사리	2022-03-14	체크아웃
142662	300	상당산성 자연휴양림	1휴_207호금낭화	2022-03-14	체크아웃
142663	181	방장산 자연휴양림	오동나무(2층)	2022-03-14	체크아웃

142664 rows × 5 columns

데이터 확인하기(Viewing Data)

In [25]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142664 entries, 0 to 142663
Data columns (total 5 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   기관아이디   142664 non-null  int64 
 1   기관이름    142664 non-null  object
 2   상품이름    142664 non-null  object
 3   숙박일자    142664 non-null  object
 4   상태      142664 non-null  object
dtypes: int64(1), object(4)
memory usage: 5.4+ MB
```

In [14]:
```
df.shape # 행, 열이 총 몇갠지 출력
```

Out[14]:
```
(142664, 5)
```

In [18]:
```
df.head(n=5) #위에서 5개 출력
```

Out[18]:

	기관아이디	기관이름	상품이름	숙박일자	상태
0	141	덕유산 자연휴양림	야영데크(127)	2021-09-26	체크아웃
1	187	희리산 자연휴양림	607호 수선화	2021-10-04	체크아웃
2	191	오서산 자연휴양림	402호 복숭아꽃	2021-09-24	체크아웃
3	224	운악산 자연휴양림	[A동]꾀꼬리	2021-09-30	체크아웃
4	106	청태산 자연휴양림	분비나무	2021-09-29	체크아웃

In [19]:
```
df.tail(n=5) #아래에서 5개 출력
```

Out[19]:

기관아이디	기관이름	상품이름	숙박일자	상태

In [28]:
```
df.T # 열과 행을 바꾼 형태의 데이터 프레임
```

Out[28]:

	0	1	2	3	4	5	6	7	8	9	...	142654	142655	14265
기관아이디	141	187	191	224	106	106	112	189	224	116	...	113	101	30
기관이름	덕유산 자연휴양림	희리산 자연휴양림	오서산 자연휴양림	운악산 자연휴양림	청태산 자연휴양림	청태산 자연휴양림	미천골 자연휴양림	변산 자연휴양림	운악산 자연휴양림	화천숲 속 야영장	...	가리왕산 자연휴양림	유명산 자연휴양림	상당산성 자연휴양
상품이름	야영데크(127)	607호 수선화	402호 복숭아꽃	[A동] 꾀꼬리	분비나무	분비나무	희망	고사포항	[B동] 다람쥐	오토캠핑장 42(쇄석)	...	파랑새	[연]오동나무	1_203 백리
숙박일자	2021-09-26	2021-10-04	2021-09-24	2021-09-30	2021-09-29	2021-09-30	2021-09-26	2021-09-26	2021-09-25	2021-09-24	...	2022-03-14	2022-03-14	2022-03-1
상태	체크아웃	체크아웃	체크아웃	체크아웃	체크아웃	체크아웃	체크아웃	체크아웃	체크아웃	체크아웃	...	체크아웃	체크아웃	체크

5 rows × 142664 columns

In [31]:
```
df.sort_index(axis=1, ascending=False) # .sort_index() 메서드: 행과 열 이름을 정
# axis=0이면 인덱스 기준으로 정렬(기본값)
# axis=1이면 컬럼을 기준으로 정렬
# 정렬의 방향에 대한 파라미터는 ascending 이용
# ascending=True: 오름차순 정렬(기본값)
# ascendng=False: 내림차순 정렬
```

Out[31]:

	숙박일자	상품이름	상태	기관이름	기관아이디
0	2021-09-26	야영데크(127)	체크아웃	덕유산 자연휴양림	141
1	2021-10-04	607호 수선화	체크아웃	희리산 자연휴양림	187
2	2021-09-24	402호 복숭아꽃	체크아웃	오서산 자연휴양림	191
3	2021-09-30	[A동]꾀꼬리	체크아웃	운악산 자연휴양림	224
4	2021-09-29	분비나무	체크아웃	청태산 자연휴양림	106
...	...	...	...	...	...
142659	2022-03-14	206호 창출	체크아웃	용현 자연휴양림	220
142660	2022-03-14	하단2관 단풍취	체크아웃	신불산 자연휴양림	105
142661	2022-03-14	하단2관 고사리	체크아웃	신불산 자연휴양림	105
142662	2022-03-14	1휴_207호금낭화	체크아웃	상당산성 자연휴양림	300
142663	2022-03-14	오동나무(2층)	체크아웃	방장산 자연휴양림	181

142664 rows × 5 columns

데이터 선택하기 (Selection)

데이터 프레임 자체가 갖고 있는 `[]` 슬라이싱 기능 이용

특정 '컬럼'의 값들만 가져오고 싶다면?

=> df['컬럼명'] 이는 `df.컬럼명`과 동일, 리턴되는 값은 Series 의 자료구조를 가지고 있음

In [37]:
```
df['기관이름']
```

Out[37]:
```
0            덕유산 자연휴양림
1            희리산 자연휴양림
2            오서산 자연휴양림
3            운악산 자연휴양림
4            청태산 자연휴양림
              ...
142659         용현 자연휴양림
142660       신불산 자연휴양림
142661       신불산 자연휴양림
142662      상당산성 자연휴양림
142663       방장산 자연휴양림
Name: 기관이름, Length: 142664, dtype: object
```

In [38]:
```
type(df['기관이름'])
```

Out[38]: pandas.core.series.Series

컬럼명을 이용해 데이터 선택하기(컬럼 선택)

특정 '행 범위'를 가져오고 싶다면 다음과 같이 리스트를 슬라이싱 할 때와 같이 []를 이용할 수 있습니다.

df[0:3] 라고 하면 0, 1, 2번째 행을 가져옵니다(데이터프레임의 첫번째 행을 0번째 행이라고 가정). [0:3] 이라고 입력했지만 3번째 행을 가져오지 않음에 유의합니다.

또 다른 방법으로 df['20130102':'20130104'] 인덱스명을 직접 넣어서 해당하는 '행 범위'를 가져올 수도 있습니다. 이 때에는 숫자를 이용하여 슬라이싱 할 때와 달리 처음과 끝의 행이 모두 포함된 결과를 가져옵니다.

In [41]:
```
df[:3] # 맨처음 3개의 행 가져오기
```

Out[41]:

	기관아이디	기관이름	상품이름	숙박일자	상태
0	141	덕유산 자연휴양림	야영데크(127)	2021-09-26	체크아웃
1	187	희리산 자연휴양림	607호 수선화	2021-10-04	체크아웃
2	191	오서산 자연휴양림	402호 복숭아꽃	2021-09-24	체크아웃

In [45]:
```
df[119:127] # 인덱스명에 해당되는 값들을 가져오기
```

Out[45]:

	기관아이디	기관이름	상품이름	숙박일자	상태
119	106	청태산 자연휴양림	야영데크(127)	2021-09-25	체크아웃
120	106	청태산 자연휴양림	야영데크(117)	2021-09-24	체크아웃
121	106	청태산 자연휴양림	야영데크(117)	2021-09-25	체크아웃
122	106	청태산 자연휴양림	분비나무	2021-09-24	체크아웃
123	106	청태산 자연휴양림	야영데크(128)	2021-09-25	체크아웃
124	106	청태산 자연휴양림	소나무	2021-09-24	체크아웃
125	106	청태산 자연휴양림	야영데크(103)	2021-09-24	체크아웃
126	106	청태산 자연휴양림	은행나무	2021-09-24	체크아웃

이름을 이용해 선택하기 : .loc

라벨의 이름을 이용해 선택할 수 있는 `.loc` 를 이용할 수도 있다.

In [51]:
```
df.loc[:,['기관이름','상품이름']]
```

Out[51]:

	기관이름	상품이름
0	덕유산 자연휴양림	야영데크(127)
1	희리산 자연휴양림	607호 수선화
2	오서산 자연휴양림	402호 복숭아꽃
3	운악산 자연휴양림	[A동]꾀꼬리
4	청태산 자연휴양림	분비나무
...	...	...
142659	용현 자연휴양림	206호 창출
142660	신불산 자연휴양림	하단2관 단풍취
142661	신불산 자연휴양림	하단2관 고사리
142662	상당산성 자연휴양림	1휴_207호금낭화
142663	방장산 자연휴양림	오동나무(2층)

142664 rows × 2 columns

위치를 이용해 선택하기: .iloc

위치를 나타내는 인덱스 번호를 이용해 데이터를 선택할 수 있다.

여기서 인덱스 번호는 python에서 사용하는 인덱스와 같은 개념으로 이해하자!

인덱스 번호는 0부터 시작하므로, 첫번째 데이터는 인덱스 번호가 0이고, 두번째 데이터는 인덱스 번호가 1이라는 뜻이다.

In [54]:
```
df.iloc[3] # 인덱스 번호 3 => 네번째 행 선택
```

Out[56]:

	기관아이디	기관이름
3	224	운악산 자연휴양림
4	106	청태산 자연휴양림

결측치

여러가지 이유로 데이터를 다 측정하지 못하는 경우 종종 발생 측정되지 못해 비어있는 데이터를 '결측치'라고 한다. pandas에서 결측치를 `np.nan` 으로 나타냄. pandas에서는 결측치를 기본적으로 연산에서 제외시킴.

In [ ]: