In [1]:
```python
import numpy as np
import pandas as pd

import requests
import time
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
import json
import googlemaps
import pprint


import matplotlib.pyplot as plt
import seaborn as sns


sns.set()

#-------------------- 차트 관련 속성 (한글처리, 그리드) -----------
plt.rcParams['font.family']= 'Malgun Gothic'
plt.rcParams['axes.unicode_minus'] = False

#-------------------- 주피터 , 출력결과 넓이 늘리기 ---------------
# from IPython.core.display import display, HTML
from IPython.display import display, HTML

display(HTML("<style>.container{width:100% !important;}</style>"))
pd.set_option('display.max_rows', 100)
pd.set_option('display.max_columns', 100)
pd.set_option('max_colwidth', None)

import warnings
warnings.filterwarnings(action='ignore')
```

In [2]:
```python
# #--------------------------------------------- 크롬 옵션 객체 생성
# # options = webdriver.ChromeOptions()
# # options.add_argument("window-size=1000x800") # 화면크기(전체화면)
# # user_agent = "Mozilla/5.0 (Windows NT 4.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/37.0.2049.0 Safari/537.36 "
# # options.add_argument('user-agent=' + user_agent)
# # options.add_argument('headless') # headless 모드 설정
# # options.add_argument("disable-gpu")
# # options.add_argument("disable-infobars")
# # options.add_argument("--disable-extensions")
# # options.add_argument("--mute-audio") #mute
# # options.add_argument('--blink-settings=imagesEnabled=false') #브라우저에서 이미지 로딩을 하지 않습니다.
# # options.add_argument('incognito') #시크릿 모드의 브라우저가 실행됩니다.
# # options.add_argument("--start-maximized")
# # driver = webdriver.Chrome('./chromedriver_102.0.5005.27.exe', options=options)

# #--------------------------------------------- 크롬 드라이버 로드  110.0.5481.177
# # https://chromedriver.chromium.org/downloads
# # https://chromedriver.storage.googleapis.com/index.html?path=110.0.5481.77/
# # ---------------------------------------------
# driver = webdriver.Chrome('../DATA_COLLECTION/chromedriver_110.exe')
# driver.get("https://api.visitjeju.net/api/contents/list?_siteId=jejuavj&locale=kr&device=pc&cate1cd=cate0000000002&tag=&sorting=reviewcnt+desc,+title_kr+asc&region1cd=&region2cd=&pageSize=12&page=1")

# #--------------------------------------------- 스크롤 다운
# # driver.find_element(By.TAG_NAME, "body").send_keys(Keys.END)

# htmlstr = driver.page_source
# htmlstr = htmlstr.replace("\n","").replace("\t","")

# print(htmlstr)
```

Visit Jeju

- API : https://m.visitjeju.net/kr/visitJejuApi
- 여행 : cate1cd=cate0000000002 (./datasets_jeju/02.여행.json) 360 rows
- 쇼핑 : cate1cd=cate0000000003 (./datasets_jeju/03.쇼핑.json) 146 rows
- 숙박 : cate1cd=cate0000000004 (./datasets_jeju/04.숙박.json) 360 rows
- 음식 : cate1cd=cate0000000005 (./datasets_jeju/05.음식.json) 360 rows
- 여행 : https://api.visitjeju.net/api/contents/list?_siteId=jejuavj&locale=kr&device=pc&cate1cd=cate0000000002&tag=&sorting=reviewcnt+desc,+title_kr+asc%C2%AEion1cd=%C2%AEion2cd=&pageSize=360&page=1
- 쇼핑 : https://api.visitjeju.net/api/contents/list?_siteId=jejuavj&locale=kr&device=pc&cate1cd=cate0000000003&tag=&sorting=reviewcnt+desc,+title_kr+asc%C2%AEion1cd=%C2%AEion2cd=&pageSize=360&page=1
- 숙박 : https://api.visitjeju.net/api/contents/list?_siteId=jejuavj&locale=kr&device=pc&cate1cd=cate0000000004&tag=&sorting=reviewcnt+desc,+title_kr+asc%C2%AEion1cd=%C2%AEion2cd=&pageSize=360&page=1
- 음식 : https://api.visitjeju.net/api/contents/list?_siteId=jejuavj&locale=kr&device=pc&cate1cd=cate0000000005&tag=&sorting=reviewcnt+desc,+title_kr+asc%C2%AEion1cd=%C2%AEion2cd=&pageSize=360&page=1

In [3]:
```python
col_list = ["contentsid","alltag","label","title", "address", "tag", "introduction",
               "readcnt", "likecnt", "reviewcnt", "markcnt", "snssharecnt", "schedulecnt", "visitcnt", "evelpt",
               "latitude", "longitude", "phoneno", "sbst","img","thumb"]

file_list = ["02.여행","03.쇼핑","04.숙박","05.음식"]

df_list = []

for fname in file_list :
    with open(f"./datasets_jeju/{fname}.json", "r", encoding="utf-8") as f:
        data_txt = f.read()
    #print(data[:100])
    dic = json.loads(data_txt)
    print ( len(dic["items"]) )

    #-------------------------------------------------------
    data_list = []
    for item in dic["items"]:
        data_list.append([
                        item['contentsid'],     #pk
                        item['alltag'],         #text-----
                        item['contentscd']['label'],
                        item['title'],          #text-----
                        item['address'],        #text-----
                        item['tag'],            #text-----
                        item['introduction'],   #text-----
                        item["readcnt"],        #607790,
                        item["likecnt"],        #322,
                        item["reviewcnt"],      #590,
                        item["markcnt"],        #3979,
                        item["snssharecnt"],    #801,
                        item["schedulecnt"],    #0,
                        item["visitcnt"],       #21,
                        item["evelpt"],         #5,
                        item["latitude"],       #33.462147,
                        item["longitude"],      #126.936424,
                        item["phoneno"],        #"064-783-0959",
                        item["sbst"],           #text-----
                        item["repPhoto"]["photoid"]["imgpath"],
                        item["repPhoto"]["photoid"]["thumbnailpath"],
                     ])
    df = pd.DataFrame(data_list, columns=col_list)
    #df.info()
    df_list.append(df)

for key in dic["items"][0].keys():
    print(f'"{key}",', end=" ")

print("")
df = pd.concat(df_list)
print(df.shape)
df.to_csv(f"./datasets_jeju/data.csv", index=False)
```

```
360
146
360
360
"alltag", "contentsid", "contentscd", "title", "region1cd", "region2cd", "address", "roadaddress", "tag", "introduction", "readcnt", "likecnt", "reviewcnt", "markcnt", "snssharecnt", "schedulecnt", "visitcnt", "evelpt", "latitude", "longitude", "postcode", "phoneno", "sbst", "created", "changed",
"catemappList", "festivalcontents", "repPhoto", "PHOTOSHARECOUNT", "reservelink",
(1226, 21)
```

In [4]:
```python
null_idx = df[df['latitude'].isna()].index.values
null_idx
```

Out[4]:
```
array([ 41, 262], dtype=int64)
```

In [5]:
```python
print(df.shape)
df = df.drop(null_idx, axis=0)
print(df.shape)
```

```
(1226, 21)
(1219, 21)
```

In [6]:
```python
df['evelpt'] = df['evelpt'].fillna(0)
df['evelpt'] = df['evelpt'].astype('int')
df[['alltag','phoneno','sbst']] = df[['alltag','phoneno','sbst']].fillna('')
```

In [7]:
```python
df['tag_orig'] = df['tag']
df['tag'] = df['tag_orig'].apply(lambda x: x.split(",")[0] if len(x) > 0 else cate)
```

In [8]:
```python
# 중복제거
df.drop_duplicates(['contentsid'], keep='first', inplace=True, ignore_index=True)
```

In [9]:
```python
df.to_csv(f"./datasets_jeju/data.csv", index=False)
```

- https://github.com/sqlalchemy/sqlalchemy/issues/4265
- https://cx-oracle.readthedocs.io/en/latest/user_guide/sql_execution.html

In [10]:
```python
df.select_dtypes('object').columns
```

Out[10]:
```
Index(['contentsid', 'alltag', 'label', 'title', 'address', 'tag',
       'introduction', 'phoneno', 'sbst', 'img', 'thumb', 'tag_orig'],
      dtype='object')
```

In [11]:
```python
from sqlalchemy import create_engine
import sqlalchemy as sa
engine = create_engine("oracle+cx_oracle://ai:0000@localhost:1521/XE")
df.to_sql("JEJU_TRAVEL", engine,
          if_exists="replace",  #,fail  , append
          # index=True,
          # index_label = 'contentsid',
          dtype={"latitude": sa.FLOAT(), "longitude": sa.FLOAT(),
                 # 'contentsid':sa.String(4000), 'alltag':sa.String(4000), 'label':sa.String(4000), 'title':sa.String(4000), 'address':sa.String(4000), 'tag':sa.String(4000),
                 # 'phoneno':sa.String(4000), 'sbst':sa.String(4000), 'img':sa.String(4000), 'thumb':sa.String(4000), 'tag_orig':sa.String(4000), #'introduction':sa.String(4000),
                }
         )
```

Out[11]:
```
1213
```

In [12]:
```python
from sqlalchemy import create_engine, text
engine = create_engine("oracle+cx_oracle://ai:0000@localhost:1521/XE")
df = pd.read_sql(text("SELECT * FROM JEJU_TRAVEL"), con = engine.connect())
print(df.shape)
df = df.set_index('index')
print( df.info() )
# df.head(1)
```

```
(1213, 23)
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1213 entries, 0 to 1212
Data columns (total 22 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   contentsid    1213 non-null   object 
 1   alltag        1212 non-null   object 
 2   label         1213 non-null   object 
 3   title         1213 non-null   object 
 4   address       1213 non-null   object 
 5   tag           1213 non-null   object 
 6   introduction  1213 non-null   object 
 7   readcnt       1213 non-null   int64  
 8   likecnt       1213 non-null   int64  
 9   reviewcnt     1213 non-null   int64  
 10  markcnt       1213 non-null   int64  
 11  snssharecnt   1213 non-null   int64  
 12  schedulecnt   1213 non-null   int64  
 13  visitcnt      1213 non-null   int64  
 14  evelpt        1213 non-null   int64  
 15  latitude      1213 non-null   float64
 16  longitude     1213 non-null   float64
 17  phoneno       1176 non-null   object 
 18  sbst          1143 non-null   object 
 19  img           1213 non-null   object 
 20  thumb         1213 non-null   object 
 21  tag_orig      1213 non-null   object 
dtypes: float64(2), int64(8), object(12)
memory usage: 218.0+ KB
None
```

In [13]:
```python
df = pd.read_csv("./datasets_jeju/data.csv")
df['label'].value_counts().index.values
```

Out[13]:
```
array(['음식점', '관광지', '숙박', '쇼핑'], dtype=object)
```

In [14]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1213 entries, 0 to 1212
Data columns (total 22 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   contentsid    1213 non-null   object 
 1   alltag        1212 non-null   object 
 2   label         1213 non-null   object 
 3   title         1213 non-null   object 
 4   address       1213 non-null   object 
 5   tag           1213 non-null   object 
 6   introduction  1213 non-null   object 
 7   readcnt       1213 non-null   int64  
 8   likecnt       1213 non-null   int64  
 9   reviewcnt     1213 non-null   int64  
 10  markcnt       1213 non-null   int64  
 11  snssharecnt   1213 non-null   int64  
 12  schedulecnt   1213 non-null   int64  
 13  visitcnt      1213 non-null   int64  
 14  evelpt        1213 non-null   int64  
 15  latitude      1213 non-null   float64
 16  longitude     1213 non-null   float64
 17  phoneno       1176 non-null   object 
 18  sbst          1143 non-null   object 
 19  img           1213 non-null   object 
 20  thumb         1213 non-null   object 
 21  tag_orig      1213 non-null   object 
dtypes: float64(2), int64(8), object(12)
memory usage: 208.6+ KB
```

In [15]:
```python
df[df['contentsid'].isin( ['CNTS_000000000022562','CNTS_000000000019605','CNTS_000000000018471','CONT_000000000501100'] )].title.values
```

Out[15]:
```
array(['마노르블랑', '키친오즈', '양가형제', '포도호텔'], dtype=object)
```

In [16]:
```python
df.isna().sum()
```

Out[16]:
```
contentsid       0
alltag           1
label            0
title            0
address          0
tag              0
introduction     0
readcnt          0
likecnt          0
reviewcnt        0
markcnt          0
snssharecnt      0
schedulecnt      0
visitcnt         0
evelpt           0
latitude         0
longitude        0
phoneno         37
sbst            70
img              0
thumb            0
tag_orig         0
dtype: int64
```

In [ ]:

In [17]:
```python
df[df["label"]=='음식점'][['title','address','likecnt','reviewcnt','evelpt','thumb']]
```

Out[17]:

	title	address	likecnt	reviewcnt	evelpt	thumb
103	마노르블랑	서귀포시 안덕면 덕수리 2952	13	30	5	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/f8853437-8984-495d-b43d-9897917bf4ce.jpg
110	키친오즈	제주시 한림읍 협재리 958-1	2	28	5	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/b19eda58-2466-48d9-aaa1-7a55bb6bf70c.jpg
142	바다다	제주특별자치도 서귀포시 대포동 2384	9	20	4	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/9b20bc0a-5c29-4336-b3b3-6a67eaf76996.jpg
232	양가형제	제주특별자치도 제주시 한경면 청수리 746-8	3	9	4	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/5c3cd90f-ed82-456c-a84e-e4a8bc87d211.jpg
287	바다를본돼지(판포점)	제주특별자치도 제주시 한경면 판포리 2881-4	1	5	5	https://api.cdn.visitjeju.net/photomng/thumbnailpath/202111/15/01f0fb09-56b2-48ac-85ba-b8bb09487bce.jpg
...	...	...	...	...	...	...
1208	누리해장국	제주특별자치도 제주시 봉개동 1819-3	1	1	5	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/fd9b1579-1763-4f76-bd02-90e7f314da5a.jpg
1209	달마야해물탕	제주특별자치도 제주시 조천읍 북촌리 1363-1	0	1	3	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/b4c4072f-8620-4c9e-a484-7ac6c9b18f2f.jpg
1210	달아래팔	제주특별자치도 제주시 한경면 낙천리 1647-4	0	1	3	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/569b98f4-386e-41dc-85cb-ab9ce71f43fd.jpg
1211	달콤한휴식	제주특별자치도 제주시 애월읍 납읍리 1249-4	0	1	3	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/98a124eb-bf41-469c-9091-f77e9160c869.jpg
1212	대광장식당	제주특별자치도 서귀포시 서귀동 297-1	1	1	4	https://api.cdn.visitjeju.net/photomng/thumbnailpath/201804/30/96c751c4-2aa3-452e-8bf1-860588b562df.jpg

361 rows × 6 columns

상세보기 페이지

- https://api.visitjeju.net/api/node/tourcontents/read.json?id=CONT_000000000500349&_siteId=jejuavj&locale=kr&device=pc&cacheTime=60

주변 관광지/맛집/숙소

- https://api.visitjeju.net/api/bigdata/list?_siteId=jejuavj&locale=kr&device=pc&distance=3&lat=33.462147&lng=126.936424&date=20230302&gender=&years=

댓글 크롤링

- https://www.visitjeju.net/kr/detail/view?contentsid=CONT_000000000500349#
- https://api.visitjeju.net/api/node/tourcontents/read.json?id=CONT_000000000500349&_siteId=jejuavj&locale=kr&device=pc&cacheTime=60

In [ ]:

In [ ]:

In [20]:
```
mm = ['CONT_000000000500349', 'CONT_000000000500477', 'CNTS_000000000001195']
print(",".join(mm))
```

```
CONT_000000000500349,CONT_000000000500477,CNTS_000000000001195
```