

## ITS 국가 교통 정보 센터

- <https://www.its.go.kr/opendata/opendataList?service=event>
- 공사 사고 정보

```
In [14]: import re
import os
import pandas as pd
from zipfile import ZipFile
from IPython.display import clear_output
```

```
In [15]: # zip폴더의 파일을 추출할 경로,
extract_path = './2018_csv_file/'

if not os.path.isdir(extract_path) :
    os.mkdir(extract_path)

# 0000년도 zip 폴더에서 extract_path에 압축된 폴더 추출,
for root, dirs, files in os.walk(f'2020/') :
    for file_name in files :
        ZipFile(f'2020/{file_name}').extractall(extract_path)

# 압축된 폴더로부터 csv 파일의 명칭을 리스트에 담기.
file_name = []
for r, d, fi in os.walk(f'{extract_path}') :
    for a in fi :
        file_name.append(a)

# 날짜 명칭만 가져오기.
variable_name = [f.split('.')[0].split('_')[-1] for f in file_name]

# for f_name, v_name in zip(file_name, variable_name) :
#     globals()[f'df_{v_name}'] = pd.read_csv(f'test/{f_name}', error_bad_lines=False, warn_bad_lines = False , sep = ',')
```

```
In [25]: df_all = pd.DataFrame()
for file in file_name :
    df = pd.read_csv(f'{extract_path}/{file}', error_bad_lines=False,
                    warn_bad_lines = False, sep = ',', encoding='cp949', header=None)
    df_all = pd.concat([df_all, df])

# 결합하였으니 인덱스 재설정.
df_all.index = list(range(df_all.shape[0]))
```

```
In [26]: df2 = df_all.iloc[:, [0, 4, 5, 6, 7, 8, 9, 10]]
df2.columns = ['date', 'num', 'event', 'address', 'lon', 'lat', 'nan1', 'nan2']
df2
```

```
Out[26]:
```

	date	num	event	address	lon	lat	nan1	nan2
0	2018-01-02 00:17:00	1.0	사고	다산로 청구역 악수역 1차로 주의운전 차량고장 -	127.013397	37.559435	NaN	NaN
1	2018-01-02 02:27:00	1.0	사고	서울외곽순환고속도로 서운JC 계양IC 1차로 차량고장 주의운전 -	126.751386	37.5356	NaN	NaN
2	2018-01-02 04:26:30	1.0	사고	창경궁로 중구청앞사거리 을지로4가 1차로 주의운전 교통사고 -	126.998509	37.566201	NaN	NaN
3	2018-01-02 06:22:01	1.0	사고	중부고속도로 남이천IC 일죽IC 1차로 차량사고 주의운전 -	127.44106	37.136445	NaN	NaN
4	2018-01-02 06:57:00	1.0	사고	석동로59번길 냉천3사거리 천지가앞사거리 1차로 주의운전 교통사고 -	128.703638	35.156766	NaN	NaN
...	...	...	...	...	...	...	...	...
120598	2018-12-31 22:05:30	1	사고	삼일대로 필동2가삼익주택 퇴계로2가 1차로 주의운전 교통사고 -	126.989626	37.561237	NaN	NaN
120599	2018-12-31 22:20:30	1	사고	강변북로 토평강변로연결 강변북로분기1 1차로 주의운전 교통사고 -	127.127702	37.570763	NaN	NaN
120600	2018-12-31 23:14:44	1	사고	[서해안선]영광부근(65K) (2차로) 승용차 관련 사고 처리중 2차로 차단 사고	126.5834	35.3571	NaN	NaN
120601	2018-12-31 23:25:31	1	사고	경부고속도로 판교JC 판교IC 5차로 차량고장 주의운전 -	127.100187	37.399173	NaN	NaN
120602	2018-12-31 23:53:20	1	사고	[청주영덕선]화서(휴)부근(46.6K) (갯길)화물차고장 처리중 고장	127.9061	36.4462	NaN	NaN

120603 rows × 8 columns

In [56]:

```
# 각 컬럼명 데이터를 원하는 방향으로 타입 변환.
df2['num'] = pd.to_numeric(df2['num'])
df2['event'] = [str(i).strip() for i in df2['event']]
df2['lon'] = ['한글' if str(i) == 'nan' else i for i in df2['lon']]
df2['lat'] = [0 if str(i) == 'nan' else i for i in df2['lat']]
df2['nan1'] = [0 if str(i) == 'nan' else i for i in df2['nan1']]
df2['nan2'] = [0 if str(i) == 'nan' else i for i in df2['nan2']]
```

<ipython-input-56-8bb0f7c339c7>:1: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
df2['num'] = pd.to_numeric(df2['num'])
```

<ipython-input-56-8bb0f7c339c7>:2: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
df2['event'] = [str(i).strip() for i in df2['event']]
```

<ipython-input-56-8bb0f7c339c7>:3: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
df2['lon'] = ['한글' if str(i) == 'nan' else i for i in df2['lon']]
```

<ipython-input-56-8bb0f7c339c7>:4: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
df2['lat'] = [0 if str(i) == 'nan' else i for i in df2['lat']]
```

<ipython-input-56-8bb0f7c339c7>:5: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
df2['nan1'] = [0 if str(i) == 'nan' else i for i in df2['nan1']]
```

<ipython-input-56-8bb0f7c339c7>:6: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
df2['nan2'] = [0 if str(i) == 'nan' else i for i in df2['nan2']]
```

In [57]:

```
df2.dropna(inplace=True)
```

<ipython-input-57-f66736151044>:1: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
df2.dropna(inplace=True)
```

In [58]:

```
df2.index = list(range(df2.shape[0]))
```

In [59]:

```
# 밀린 데이터로 인해, 특정 컬럼내에 한글이 존재하면 이에 해당하는 인덱스 반환.
index_list = []
for j, i in enumerate(df2['lon']) :
    a = re.sub('[^0-9.]+', '#', str(i))
    if '#' in a :
        index_list.append(j)
clear_output(wait = True)
print(f'{j+1}/{df2.shape[0]}번째')
```

IOPub message rate exceeded.

The notebook server will temporarily stop sending output  
to the client in order to avoid crashing it.  
To change this limit, set the config variable

```
`--NotebookApp.iopub_msg_rate_limit`.
```

```
Current values:  
NotebookApp.iopub_msg_rate_limit=1000.0 (msgs/sec)  
NotebookApp.rate_limit_window=3.0 (secs)
```

```
In [60]: df2.drop(index_list, inplace=True)
```

```
C:\Users\user\anaconda3\lib\site-packages\pandas\core\frame.py:4308: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy  
return super().drop()
```

```
In [61]: df2.drop(['nan1', 'nan2'], inplace=True, axis=1)  
df2.isna().sum()
```

```
Out[61]: date      0  
num      0  
event     0  
address    0  
lon       0  
lat       0  
dtype: int64
```

```
In [62]: df2
```

```
Out[62]:
```

	date	num	event	address	lon	lat
0	2018-01-02 00:17:00	1.0	사고	다산로 청구역 약수역 1차로 주의운전 차량고장 -	127.013397	37.559435
1	2018-01-02 02:27:00	1.0	사고	서울외곽순환고속도로 서운JC 계양IC 1차로 차량고장 주의운전 -	126.751386	37.5356
2	2018-01-02 04:26:30	1.0	사고	창경궁로 중구청앞사거리 을지로4가 1차로 주의운전 교통사고 -	126.998509	37.566201
3	2018-01-02 06:22:01	1.0	사고	중부고속도로 남이천IC 일죽IC 1차로 차량사고 주의운전 -	127.44106	37.136445
4	2018-01-02 06:57:00	1.0	사고	석동로59번길 냉천3사거리 천지가앞사거리 1차로 주의운전 교통사고 -	128.703638	35.156766
...	...	...	...	...	...	...
120221	2018-12-31 22:05:30	1.0	사고	삼일대로 필동2가삼익주택 퇴계로2가 1차로 주의운전 교통사고 -	126.989626	37.561237
120222	2018-12-31 22:20:30	1.0	사고	강변북로 토평강변로연결 강변북로분기1 1차로 주의운전 교통사고 -	127.127702	37.570763
120223	2018-12-31 23:14:44	1.0	사고	[서해안선]영광부근(65K) (2차로) 승용차 관련 사고 처리중 2차로 차단 사고	126.5834	35.3571
120224	2018-12-31 23:25:31	1.0	사고	경부고속도로 판교JC 판교IC 5차로 차량고장 주의운전 -	127.100187	37.399173
120225	2018-12-31 23:53:20	1.0	사고	[청주영덕선]화서(휴)부근(46.6K) (갓길)화물차고장 처리중 고장	127.9061	36.4462

119284 rows × 6 columns

## 저장 및 확인

```
In [63]: # 저장.  
df2.to_csv('2018.csv', encoding='cp949', index=False, header=True)
```

```
In [64]: # 확인.  
pd.read_csv('2018.csv', encoding='cp949')
```

```
Out[64]:
```

	date	num	event	address	lon	lat
0	2018-01-02 00:17:00	1.0	사고	다산로 청구역 약수역 1차로 주의운전 차량고장 -	127.013397	37.5594348
1	2018-01-02 02:27:00	1.0	사고	서울외곽순환고속도로 서운JC 계양IC 1차로 차량고장 주의운전 -	126.751386	37.5355995023003
2	2018-01-02 04:26:30	1.0	사고	창경궁로 중구청앞사거리 을지로4가 1차로 주의운전 교통사고 -	126.998509	37.5662011
3	2018-01-02 06:22:01	1.0	사고	중부고속도로 남이천IC 일죽IC 1차로 차량사고 주의운전 -	127.441060	37.1364446890986
4	2018-01-02 06:57:00	1.0	사고	석동로59번길 냉천3사거리 천지가앞사거리 1차로 주의운전 교통사고 -	128.703638	35.156765684386
...	...	...	...	...	...	...

119279	2018-12-31 22:05:30	1.0	사고	삼일대로 필동2가삼익주택 퇴계로2가 1차로 주의운전 교통사고 -	126.989626	37.5612375
119280	2018-12-31 22:20:30	1.0	사고	강변북로 토평강변로연결 강변북로분기1 1차로 주의운전 교통사고 -	127.127702	37.5707632
119281	2018-12-31 23:14:44	1.0	사고	[서해안선]영광부근(65K) (2차로) 승용차 관련 사고 처리중 2차로 차단 사고	126.583400	35.3571
119282	2018-12-31 23:25:31	1.0	사고	경부고속도로 판교JC 판교IC 5차로 차량고장 주의운전 -	127.100187	37.3991730651403
119283	2018-12-31 23:53:20	1.0	사고	[청주영덕선]화서(휴)부근(46.6K) (갓길)화물차고장 처리중 고장	127.906100	36.4462

119284 rows × 6 columns

## 2018 ~ 2019년 사고 데이터

### 2018년 데이터

In [70]:

```
df = pd.read_csv('2018.csv', encoding='cp949')
df2 = df[df['event'] == '사고']
df2.index = list(range(df2.shape[0]))
df2.to_csv('2018_사고.csv', encoding='cp949', index=False, header=True)
print(f'{df2.shape[0]}개')
```

50941개

### 2019년 데이터

In [71]:

```
df = pd.read_csv('2019.csv', encoding='cp949')
df2 = df[df['event'] == '사고']
df2.index = list(range(df2.shape[0]))
df2.to_csv('2019_사고.csv', encoding='cp949', index=False, header=True)
print(f'{df2.shape[0]}개')
```

63488개

### 2020년 데이터

In [72]:

```
df = pd.read_csv('2020.csv', encoding='cp949')
df2 = df[df['event'] == '사고']
df2.index = list(range(df2.shape[0]))
df2.to_csv('2020_사고.csv', encoding='cp949', index=False, header=True)
print(f'{df2.shape[0]}개')
```

88734개