▣ 연습문제

- 대전충청권출발 지역간 월누계교통량
  - http://data.ex.co.kr/portal/fdwn/view?type=TCS&num=C4&requestfrom=dataset
  - 파일: TCS_C4_04_02_818628.csv
  - encoding="cp949"

In [1]:
```
### Packages
import pandas as pd
from matplotlib import pyplot as plt
```

In [2]:
```
### 그래프에 굴림 글꼴 사용
plt.rcParams['font.family'] = 'Gulim'    # 'AppleGothic' in mac
```

In [3]:
```
### 파일 선택
data_file = 'TCS_C4_04_02_818628.csv'
```

1. 파일 자료의 관측값(행)의 수와 변수(열)의 수를 구하시오.

In [4]:
```
### 파일 읽기
df = pd.read_csv(data_file, encoding="cp949")
df.head()
df.shape
```

Out[4]: (949468, 16)

1. 모든 변수에 대한 기술통계량을 구하시오. - df.describe()

In [5]:
```
df.describe(include='all')
```

Out[5]:

	집계일자	출발영업소코드	도착영업소코드	출발영업소명	도착영업소명	출발권역코드	도착권역코드	출발권역명	도착권역명	도착지방
count	9.494680e+05	949468.000000	949468.000000	949468	949468	949468	949468	949468	817532	949468.0
unique	NaN	NaN	NaN	76	403	6	9	2	8	
top	NaN	NaN	NaN	북상주	양산			대전충남본부	부산경남본부	
freq	NaN	NaN	NaN	12493	2356	662129	539524	499720	157852	
mean	2.018082e+07	392.000000	445.615385	NaN	NaN	NaN	NaN	NaN	NaN	1.1
std	8.944277e+00	219.653371	532.131748	NaN	NaN	NaN	NaN	NaN	NaN	24.0
min	2.018080e+07	107.000000	101.000000	NaN	NaN	NaN	NaN	NaN	NaN	0.0
25%	2.018081e+07	179.500000	202.000000	NaN	NaN	NaN	NaN	NaN	NaN	0.0
50%	2.018082e+07	521.500000	503.000000	NaN	NaN	NaN	NaN	NaN	NaN	0.0
75%	2.018082e+07	572.250000	605.000000	NaN	NaN	NaN	NaN	NaN	NaN	0.0
max	2.018083e+07	796.000000	9999.000000	NaN	NaN	NaN	NaN	NaN	NaN	4477.0

1. **도착지방향총교통량**이 0보다 큰 자료를 df_0에 저장하고, 관측값의 수와 변수의 수를 구하시오. - df.loc[]

In [6]:
```
### 도착지방향총교통량이 0이상인 자료 추출
df_0 = df.loc[df['도착지방향총교통량']>0]
df_0.shape
```

Out[6]: (40539, 16)

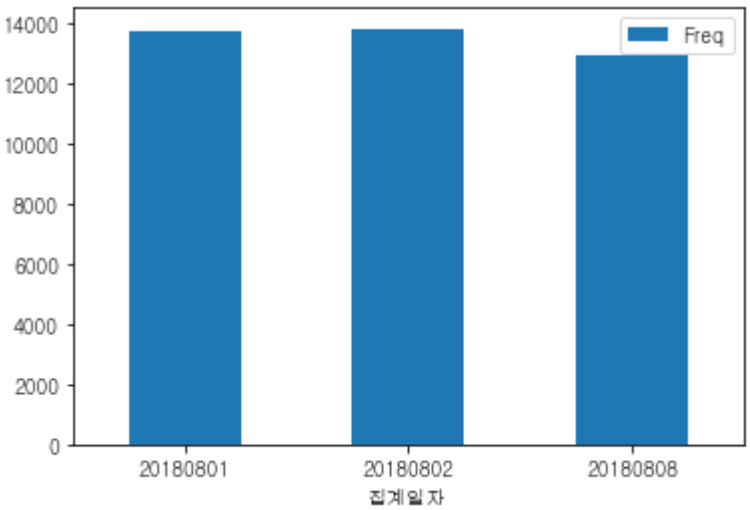
1. 3번에서 저장한 df_0 자료에 대하여, **집계일자**에 대한 막대그래프를 그리시오.

In [7]:
```
tb = pd.crosstab(df_0['집계일자'], 'Freq', colnames=[''])
tb
```

Out[7]:

집계일자	Freq
20180801	13756
20180802	13820
20180808	12963

In [8]:
```
### 막대그래프
ax = tb.plot.bar(rot=0)
```



1. df_0 자료에서 8월 2일 자료만 추출하여 df_8_2에 저장하고, 관측값의 수와 변수의 수를 구하시오. - df.loc[]

In [9]:
```
df_8_2 = df_0.loc[df_0['집계일자']==20180802]
df_8_2.shape
```

Out[9]: (13820, 16)

1. df_8_2 자료에서 **출발영업소명**이 '안성'인 자료를 추출하여 df_as에 저장하고, 관측값의 수와 변수의 수를 구하시오.- df.loc[]

In [10]:
```
df_as = df_8_2.loc[df_8_2['출발영업소명']=='안성']
df_as.shape
```

Out[10]: (252, 16)

1. df_as 자료에서 **도착지방향총교통량**에 대한 기술통계량을 구하시오. - df.describe()

In [11]:

```
df_as.describe()
```

Out[11]:

	집계일자	출발영업소코드	도착영업소코드	도착지방향1종교통량	도착지방향2종교통량	도착지방향3종교통량	도착지방향4종교통량	도착지방향5종교통량	도착지방향6종교통량
count	252.0	252.0	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000
mean	20180802.0	107.0	362.666667	59.289683	1.690476	3.107143	1.099206	0.765873	4.384921
std	0.0	0.0	212.822617	263.032695	7.249634	23.859825	3.952164	3.219207	19.564345
min	20180802.0	107.0	101.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	20180802.0	107.0	177.750000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	20180802.0	107.0	264.000000	4.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	20180802.0	107.0	567.000000	17.000000	0.000000	1.000000	1.000000	0.000000	1.000000
max	20180802.0	107.0	766.000000	3257.000000	80.000000	365.000000	41.000000	34.000000	224.000000

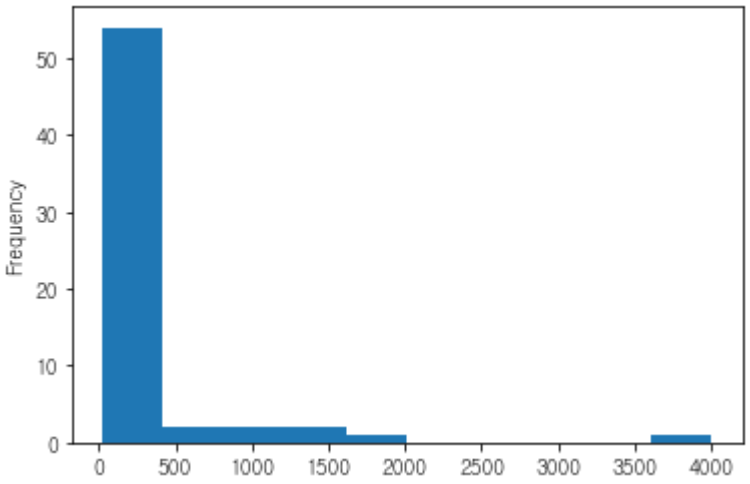
1. df_as 자료에서 **도착지방향총교통량**이 20보다 큰 자료를 추출하여 df_as_20에 저장하고, 관측값의 수와 변수의 수를 구하시오.

In [12]:
```
df_as_20 = df_as.loc[df_as['도착지방향총교통량']>20]
df_as_20.shape
```

Out[12]: (62, 16)

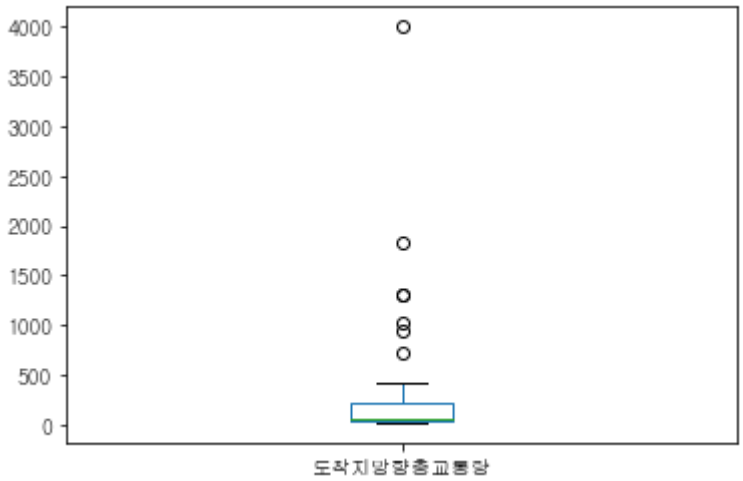
1. df_as_20 자료에서 **도착지방향총교통량**에 대한 히스토그램을 그리시오.

In [13]:
```
### 히스토그램
ax = df_as_20['도착지방향총교통량'].plot.hist()
```



1. df_as_20 자료의 **도착지방향총교통량**에 대한 상자그림을 그리시오.

In [14]:
```
### 상자그림
ax = df_as_20['도착지방향총교통량'].plot.box()
```



1. df_as_20 자료의 **도착지방향1종교통량**과 **도착지방향2종교통량**에 대한 산점도를 그리시오.

In [15]:
```
### 산점도 - 변수 2개
ax = df_as_20.plot.scatter(x='도착지방향1종교통량', y='도착지방향2종교통량', c='darkblue')
```