데이터 1 전처리

- 'emd_cd' 와 'emd_nm' 은 1:1 관계 -> 상대적 구분이 어려운 'emd_cd' 열 제거.
- 행정동별 배출거점 지역코드의 개수 카운트 -> 배출량 주요 요인 분석을 위한 파생변수 'area_cnt(배출거점지역 개수)' 열 생성, 기존 'em_area_cd'는 제거
- ['y_m', 'city', 'emd_nm'] 열들을 기준으로 그룹화

In [1]:
```
# 필요 라이브러리 로드
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

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

In [2]:
```
# 파일 불러오기
df = pd.read_csv('01_음식물쓰레기_FOOD_WASTE_210811_update.csv',encoding='cp949')
df.head()
```

Out[2]:

	base_date	city	emd_cd	emd_nm	em_area_cd	em_cnt	em_g	pay_amt
0	2018-01-01	제주시	50110670	외도동	W6X062	2	15500	464
1	2018-01-01	제주시	50110630	아라동	W6XA97	25	59800	1787
2	2018-01-01	제주시	50110630	아라동	W6XA96	8	25350	758
3	2018-01-01	제주시	50110630	아라동	W6XA95	12	30000	898
4	2018-01-01	제주시	50110630	아라동	W6XA94	14	44050	1317

In [3]:
```
# 행정동명 하나당 emd_cd는 한개이다.
# 'emd_cd' 와 'emd_nm' 은 1:1 관계
for i in df.emd_nm.unique():
    display(i,df[df.emd_nm == i].emd_cd.unique())
```

```
'외도동'
array(['50110670', 50110670], dtype=object)
'아라동'
array(['50110630', 50110630], dtype=object)
'노형동'
array(['50110660', 50110660], dtype=object)
'도두동'
array(['50110690', 50110690], dtype=object)
'오라동'
array(['50110640', 50110640], dtype=object)
'삼양동'
array(['50110610', 50110610], dtype=object)
'이도2동'
array(['50110540', 50110540], dtype=object)
'일도2동'
array(['50110520', 50110520], dtype=object)
'건입동'
array(['50110590', 50110590], dtype=object)
'화북동'
array(['50110600', 50110600], dtype=object)
```

```
'봉개동'
array(['50110620', 50110620], dtype=object)
'삼도1동'
array(['50110550', 50110550], dtype=object)
'이도1동'
array(['50110530', 50110530], dtype=object)
'용담2동'
array(['50110580', 50110580], dtype=object)
'연동'
array(['50110650', 50110650], dtype=object)
'이호동'
array(['50110680', 50110680], dtype=object)
'애월읍'
array(['50110253', 50110253], dtype=object)
'동홍동'
array(['50130570', 50130570], dtype=object)
'대정읍'
array(['50130250', 50130250], dtype=object)
'영천동'
array(['50130560', 50130560], dtype=object)
'서홍동'
array(['50130580', 50130580], dtype=object)
'효돈동'
array(['50130550', 50130550], dtype=object)
'중문동'
array(['50130610', 50130610], dtype=object)
'대륜동'
array(['50130590', 50130590], dtype=object)
'남원읍'
array(['50130253', 50130253], dtype=object)
'대천동'
array(['50130600', 50130600], dtype=object)
'안덕면'
array(['50130310', 50130310], dtype=object)
'천지동'
array(['50130540', 50130540], dtype=object)
'예래동'
array(['50130620', 50130620], dtype=object)
'표선면'
array(['50130320', 50130320], dtype=object)
'알수없음'
array(['알수없음'], dtype=object)
'성산읍'
array(['50130259', 50130259], dtype=object)
'정방동'
array(['50130520', 50130520], dtype=object)
'송산동'
array(['50130510', 50130510], dtype=object)
'중앙동'
array(['50130530', 50130530], dtype=object)
'삼도2동'
array(['50110560', 50110560], dtype=object)
'일도1동'
array(['50110510', 50110510], dtype=object)
'용담1동'
array(['50110570', 50110570], dtype=object)
'조천읍'
array(['50110259', 50110259], dtype=object)
'구좌읍'
array(['50110256', 50110256], dtype=object)
'한림읍'
array(['50110250', 50110250], dtype=object)
'한경면'
array(['50110310', 50110310], dtype=object)
```

In [4]:
```
## column 이름 변경, 행정동 코드 제거
df = df.drop('emd_cd', axis = 1)
df = df.rename(columns={'emd_nm':'location'})
df.head()
```

Out[4]:

	base_date	city	location	em_area_cd	em_cnt	em_g	pay_amt
0	2018-01-01	제주시	외도동	W6X062	2	15500	464
1	2018-01-01	제주시	아라동	W6XA97	25	59800	1787
2	2018-01-01	제주시	아라동	W6XA96	8	25350	758
3	2018-01-01	제주시	아라동	W6XA95	12	30000	898
4	2018-01-01	제주시	아라동	W6XA94	14	44050	1317

In [5]:
```
# 'y_m'(년_월) 열 새로 생성
df["y_m"] = df.apply(lambda x: x.base_date[:7], axis = 1)
df = df.drop('base_date', axis = 1)
df.head()
```

Out[5]:

	city	location	em_area_cd	em_cnt	em_g	pay_amt	y_m
0	제주시	외도동	W6X062	2	15500	464	2018-01
1	제주시	아라동	W6XA97	25	59800	1787	2018-01
2	제주시	아라동	W6XA96	8	25350	758	2018-01
3	제주시	아라동	W6XA95	12	30000	898	2018-01
4	제주시	아라동	W6XA94	14	44050	1317	2018-01

In [6]:
```
# 행정동별 쓰레기 배출거점지역 개수 카운트 열 "area_cnt" 생성
for i in df.location.unique():
    df.loc[df.location == i,"area_cnt"]= df[df.location == i].em_area_cd.nunique()

# em_area_cd 제거
df = df.drop('em_area_cd', axis = 1)
df.head()
```

Out[6]:

	city	location	em_cnt	em_g	pay_amt	y_m	area_cnt
0	제주시	외도동	2	15500	464	2018-01	66.0
1	제주시	아라동	25	59800	1787	2018-01	120.0
2	제주시	아라동	8	25350	758	2018-01	120.0
3	제주시	아라동	12	30000	898	2018-01	120.0
4	제주시	아라동	14	44050	1317	2018-01	120.0

In [7]:
```
df_g = df.groupby(['y_m','city','location','area_cnt']).sum().reset_index()
df_g
```

Out[7]:

	y_m	city	location	area_cnt	em_cnt	em_g	pay_amt
0	2018-01	서귀포시	남원읍	52.0	9570	42437700	1270773

	y_m	city	location	area_cnt	em_cnt	em_g	pay_amt
1	2018-01	서귀포시	대륜동	38.0	21666	57612600	1676850
2	2018-01	서귀포시	대정읍	89.0	10185	38885550	1164122
3	2018-01	서귀포시	대천동	37.0	20280	53858550	1593709
4	2018-01	서귀포시	동홍동	49.0	45936	118701000	3501286
...	...	...	...	...	...	...	...
1661	2021-06	제주시	일도2동	87.0	84360	147438200	4402149
1662	2021-06	제주시	조천읍	141.0	27732	63927750	1911187
1663	2021-06	제주시	한경면	71.0	8031	27060150	809898
1664	2021-06	제주시	한림읍	112.0	25653	82746990	2476292
1665	2021-06	제주시	화북동	84.0	66088	110750050	3306029

1666 rows × 7 columns

In [8]:
```
df_g.to_csv("1번데이터_전처리.csv",encoding = "cp949", index = False)
```