데이터 4 전처리

- ‘mct_cat_cd’와 'mct_cat_nm’ 은 1:1 관계 -> 상대적 구분이 어려운 'mct_cat_nm’ 열 제거
- ‘mct_cat_nm’, ‘use_cnt’, ‘use_amt’를 활용 -> 각 업종별 월별 ‘결제금액’, ‘결제건수’열 생성
- 거주지 중심 발생 음식물쓰레기 데이터에 포함되지 않을 식당(‘한식’, ‘양식’, ‘아시아음식’, ‘부페’, ‘주점 및 주류판매’) 데이터 통합

  ex) 2018-01, 제주시, 애월읍, 식당_결제건수, 식당_결제금액, 패스트푸드_결제건수, 배달앱_결제금액

In [1]:
```
#라이브러리 로드
import numpy as np
import pandas as pd

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

In [2]:
```
# 업종별 데이터 새로운 열으로 생성하는 함수
def addColumns(df,df_n,s):    #df_n : 새로 생성할 데이터프레임, df: 사용할 데이터프레임
    condition = df.mct_cat_nm == s
    temp = df[condition]
    df_n[s+'_결제건수'] = temp.use_cnt
    df_n[s+'_결제금액'] = temp.use_amt
    return df_n

def showData(df,df_n,s):    #df_n : 새로 생성할 데이터프레임, df: 사용할 데이터프레임
    condition = df.mct_cat_nm == s
    temp = df[condition]
    return temp.use_amt
```

In [3]:
```
# 데이터 로드
df = pd.read_csv("04_음식관련 카드소비_CARD_SPENDING.csv",encoding = 'cp949',parse_dates=
df.info()
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 544805 entries, 0 to 544804
Data columns (total 8 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   base_date   544805 non-null  datetime64[ns]
 1   city        544805 non-null  object
 2   emd_cd      544805 non-null  object
 3   emd_nm      544805 non-null  object
 4   mct_cat_cd  544805 non-null  object
 5   mct_cat_nm  544805 non-null  object
 6   use_cnt     544805 non-null  int64
 7   use_amt     544805 non-null  int64
dtypes: datetime64[ns](1), int64(2), object(5)
memory usage: 33.3+ MB
```

Out[3]:

	base_date	city	emd_cd	emd_nm	mct_cat_cd	mct_cat_nm	use_cnt	use_amt
0	2018-01-06	제주시	50110650	연동	C00100	한식	3643	127777300
1	2018-01-09	제주시	50110650	연동	C00500	패스트푸드	432	6711675

	base_date	city	emd_cd	emd_nm	mct_cat_cd	mct_cat_nm	use_cnt	use_amt
2	2018-01-15	제주시	50110650	연동	C01200	농축수산물	236	16089579
3	2018-01-15	제주시	50110650	연동	A00200	마트/슈퍼마켓	4031	121979867
4	2018-01-20	제주시	50110650	연동	C01000	식품	633	30410674

In [4]:
```
# 추자면, 우도면 제거
df = df[df.emd_nm != '추자면']
df = df[df.emd_nm != '우도면']
df.loc[df['emd_nm']=='알수없음','city'] = '알수없음'
df
```

Out[4]:

	base_date	city	emd_cd	emd_nm	mct_cat_cd	mct_cat_nm	use_cnt	use_amt
0	2018-01-06	제주시	50110650	연동	C00100	한식	3643	127777300
1	2018-01-09	제주시	50110650	연동	C00500	패스트푸드	432	6711675
2	2018-01-15	제주시	50110650	연동	C01200	농축수산물	236	16089579
3	2018-01-15	제주시	50110650	연동	A00200	마트/슈퍼마켓	4031	121979867
4	2018-01-20	제주시	50110650	연동	C01000	식품	633	30410674
...	...	...	...	...	...	...	...	...
544800	2020-10-16	알수없음	XXXXXXXX	알수없음	C01400	배달	5	127250
544801	2019-08-31	알수없음	XXXXXXXX	알수없음	C01400	배달	5	39974
544802	2020-10-03	알수없음	XXXXXXXX	알수없음	C01400	배달	10	329832
544803	2020-10-23	알수없음	XXXXXXXX	알수없음	C01400	배달	5	68715
544804	2020-12-21	알수없음	XXXXXXXX	알수없음	C01400	배달	10	187452

530618 rows × 8 columns

In [5]:
```
# # column 이름 변경, 행정동 코드 제거
# df = df.rename(columns={'base_date':'y_m'})
# df = df.rename(columns={'emd_nm':'location'})
# del df['emd_cd'],df['mct_cat_cd']
# df.head()
```

In [6]:
```
# base_date 0000-00 형태로 변환
df = df.set_index('base_date')
df['month']=df.index.month
df['year']=df.index.year
```

In [7]:
```
# 월별 날짜 열 생성
df['month'] = df['month'].apply(lambda x: "{:0>2d}".format(x))

df['year']=df['year'].astype('str')
df['month']=df['month'].astype('str')

df = df.reset_index()
df['base_date'] = df['year']+'-'+df['month']
del df['month'],df['year']
```

```
df.head()
```

Out[7]:

	base_date	city	emd_cd	emd_nm	mct_cat_cd	mct_cat_nm	use_cnt	use_amt
0	2018-01	제주시	50110650	연동	C00100	한식	3643	127777300
1	2018-01	제주시	50110650	연동	C00500	패스트푸드	432	6711675
2	2018-01	제주시	50110650	연동	C01200	농축수산물	236	16089579
3	2018-01	제주시	50110650	연동	A00200	마트/슈퍼마켓	4031	121979867
4	2018-01	제주시	50110650	연동	C01000	식품	633	30410674

In [8]:
```
# column 이름 변경, 행정동 코드 제거
df = df.rename(columns={'base_date':'y_m'})
df = df.rename(columns={'emd_nm':'location'})
del df['emd_cd'],df['mct_cat_cd']
df.head()
```

Out[8]:

	y_m	city	location	mct_cat_nm	use_cnt	use_amt
0	2018-01	제주시	연동	한식	3643	127777300
1	2018-01	제주시	연동	패스트푸드	432	6711675
2	2018-01	제주시	연동	농축수산물	236	16089579
3	2018-01	제주시	연동	마트/슈퍼마켓	4031	121979867
4	2018-01	제주시	연동	식품	633	30410674

In [9]:
```
# 지역별, 월별 그룹화

df_g = df.groupby(['y_m','city','location','mct_cat_nm'])['use_cnt','use_amt'].sum()
df_g = df_g.reset_index()
df_g.head()
```

Out[9]:

	y_m	city	location	mct_cat_nm	use_cnt	use_amt
0	2018-01	서귀포시	남원읍	간식	3073	36003854
1	2018-01	서귀포시	남원읍	농축수산물	1753	132565191
2	2018-01	서귀포시	남원읍	마트/슈퍼마켓	13786	700738510
3	2018-01	서귀포시	남원읍	배달	265	5163749
4	2018-01	서귀포시	남원읍	부페	15	836173

In [10]:
```
# 업종별 결제건수, 결제금액 열 생성
cat_nm = ['한식', '아시아음식', '패스트푸드', '간식', '농축수산물', '마트/슈퍼마켓', '식품

df_n = df_g.copy()
for i in range(len(cat_nm)):
    addColumns(df_g,df_n,cat_nm[i])
df_n
```

Out[10]:

	y_m	city	location	mct_cat_nm	use_cnt	use_amt	한식_결제건수	한식_결제금액	아시아음식_결제건수	아시아음결제금
0	2018-01	서귀포시	남원읍	간식	3073	36003854	NaN	NaN	NaN	N
1	2018-01	서귀포시	남원읍	농축수산물	1753	132565191	NaN	NaN	NaN	N
2	2018-01	서귀포시	남원읍	마트/슈퍼마켓	13786	700738510	NaN	NaN	NaN	N
3	2018-01	서귀포시	남원읍	배달	265	5163749	NaN	NaN	NaN	N
4	2018-01	서귀포시	남원읍	부페	15	836173	NaN	NaN	NaN	N
...	...	...	...	...	...	...	...	...	...	
19146	2021-06	제주시	화북동	아시아음식	4475	141865970	NaN	NaN	4475.0	1418659
19147	2021-06	제주시	화북동	양식	1572	48038685	NaN	NaN	NaN	N
19148	2021-06	제주시	화북동	주점및주류판매	685	23014262	NaN	NaN	NaN	N
19149	2021-06	제주시	화북동	패스트푸드	7944	125227233	NaN	NaN	NaN	N
19150	2021-06	제주시	화북동	한식	50517	1411014487	50517.0	1.411014e+09	NaN	N

19151 rows × 28 columns

In [11]:
```
# 데이터 프레임 정리
df_new = df_n.drop(columns=['mct_cat_nm','use_cnt','use_amt'], axis=1)
df_new = df_new.groupby(['y_m','city','location']).sum().reset_index()
df_new.head()
```

Out[11]:

	y_m	city	location	한식_결제건수	한식_결제금액	아시아음식_결제건수	아시아음식_결제금액	패스트푸드_결제건수	패스트푸드_결제금액	간식_결제건수	...	4
0	2018-01	서귀포시	남원읍	15412.0	5.459299e+08	1633.0	87237934.0	3337.0	57030976.0	3073.0	...	91
1	2018-01	서귀포시	대륜동	19905.0	6.700034e+08	2284.0	92484907.0	4889.0	81574888.0	5266.0	...	3(
2	2018-01	서귀포시	대정읍	22768.0	7.621949e+08	2412.0	96196033.0	5221.0	98015164.0	8310.0	...	5(
3	2018-01	서귀포시	대천동	18175.0	5.688884e+08	2199.0	92292288.0	5136.0	80300156.0	5351.0	...	28
4	2018-01	서귀포시	동홍동	33125.0	1.231469e+09	3315.0	152938153.0	6908.0	126657868.0	7959.0	...	54

5 rows × 25 columns

In [12]:
```
# 식당 열 생성, 해당 열 제거
df_new['식당_결제건수'] = df_new['한식_결제건수']+df_new['양식_결제건수']+df_new['아시아음
df_new['식당_결제금액'] = df_new['한식_결제금액']+df_new['양식_결제금액']+df_new['아시아음
df_final = df_new.drop(columns=['한식_결제건수','한식_결제금액','양식_결제건수','양식_결제
df_final
```

Out[12]:

	y_m	city	location	패스트푸드_결제건수	패스트푸드_결제금액	간식_결제건수	간식_결제금액	농축수산물_결제건수	농축수산물_결제금액	마트/슈퍼마켓_결제건수	
0	2018-01	서귀포시	남원읍	3337.0	57030976.0	3073.0	36003854.0	1753.0	132565191.0	13786.0	7
1	2018-01	서귀포시	대륜동	4889.0	81574888.0	5266.0	62481472.0	2026.0	133506435.0	25909.0	1
2	2018-01	서귀포시	대정읍	5221.0	98015164.0	8310.0	88847534.0	1959.0	146060716.0	30433.0	1

	y_m	city	location	패스트푸드_결제건수	패스트푸드_결제금액	간식_결제건수	간식_결제금액	농축수산물_결제건수	농축수산물_결제금액	마트/슈퍼마켓_결제건수	
3	2018-01	서귀포시	대천동	5136.0	80300156.0	5351.0	56156408.0	1722.0	116891485.0	23893.0	9
4	2018-01	서귀포시	동홍동	6908.0	126657868.0	7959.0	89540973.0	4031.0	263004249.0	55399.0	2
...	...	...	...	...	...	...	...	...	...	...	
1759	2021-06	제주시	일도2동	10469.0	135801500.0	16303.0	186182590.0	8846.0	519191996.0	60791.0	1
1760	2021-06	제주시	조천읍	6678.0	113730242.0	8324.0	112097929.0	3773.0	307434356.0	25722.0	1
1761	2021-06	제주시	한경면	1406.0	26876799.0	1507.0	23045501.0	1040.0	96593409.0	8888.0	3
1762	2021-06	제주시	한림읍	4315.0	81433739.0	5256.0	74235134.0	2075.0	221504210.0	20129.0	7
1763	2021-06	제주시	화북동	7944.0	125227233.0	10916.0	129374585.0	8098.0	664687009.0	48143.0	1

1764 rows × 17 columns

In [13]:
```
df_final.to_csv("4번데이터_전처리.csv",encoding = "cp949", index = False)
```