import pandas as pd import matplotlib.pyplot as plt

SEOUL_CLASS_1_preprocessing

- 서울시 상권 데이터를 업종 별로 파악할 수 있는 데이터
- 업종별 isrisky 비율을 알 수 있는 'CREDIT_CLASS_1.xlsx' 파일에서 isrisky 비율 값만 가져오기
- 업종별 데이터를 제공하는 추정 매출액과 점포 데이터 값을 컬럼으로 추가하기
- 나머지 컬럼들은 업종별 데이터를 제공하지 않으므로 기존 상권별 데이터 'SEOUL_CLUSTER.csv'에서 데이터 가져오기

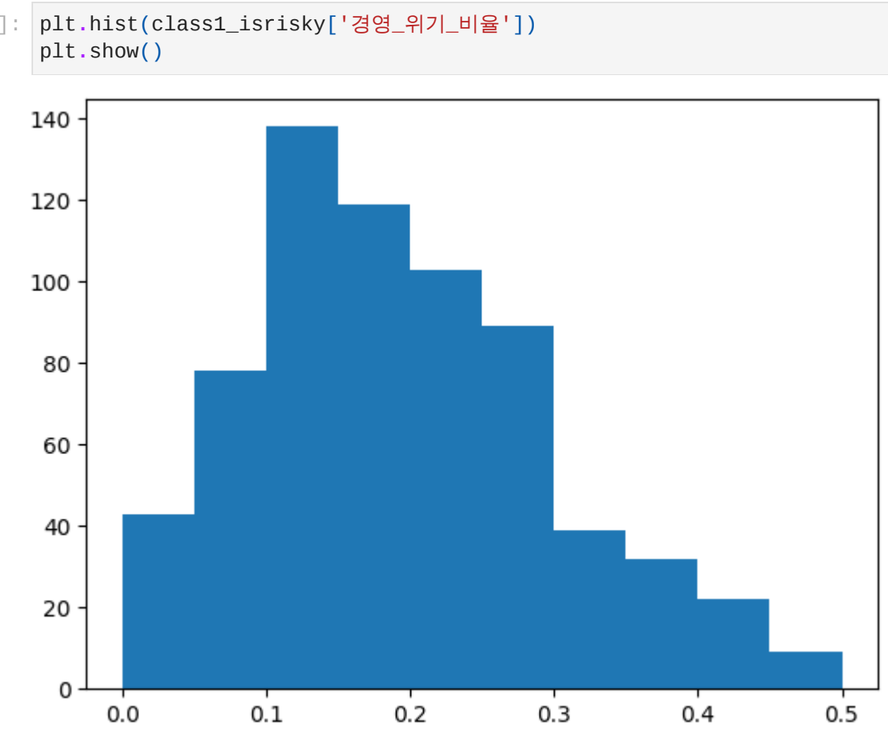
In [46]:
```
class1 = pd.read_excel('../data/CREDIT_CLASS_1.xlsx')
class1 = class1[class1['year'] == 2022]
```

In [4]:
```
class1.columns
```

Out[4]:
```
Index(['trdar_nm', 'year', 'quarter', 'class_1_name', 'average(age)',
       'average(duration)', 'average(is_franchise)',
       'average(business_square_size)', 'average(is_risky)',
       'average(monthly_rental_fee)', 'average(regular_employees_count)',
       'average(rental_deposit)', 'average(sum_customer_cnt)',
       'average(sum_new_customer_cnt)', 'average(sum_purchase_card)',
       'average(sum_purchase_cash)', 'average(sum_purchase_invoice)',
       'average(sum_sales_card)', 'average(sum_sales_delivery)',
       'average(sum_sales_invoice)', 'average(sum_weekend_sales_card)',
       'average(sum_weekend_sales_delivery)'],
      dtype='object')
```

In [49]:
```
class1_isrisky = class1[['trdar_nm', 'year', 'quarter', 'class_1_name', 'average(is_risky)']]
class1_isrisky.columns = ['상권_코드_명', '기준_년_코드', '기준_분기_코드', '업종_대분류', '경영_위기_비율']
```

In [51]:
```
plt.hist(class1_isrisky['경영_위기_비율'])
plt.show()
```



In [33]:
```
len(class1_isrisky['상권_코드_명'].unique())
```

Out[33]: 113

In [23]:
```
seoul = pd.read_csv('../data/SEOUL_CLUSTER.csv')
seoul.drop('경영_위기_비율', axis=1, inplace=True)
seoul
```

Out[23]:

	상권_코드_명	상권_코드	기준_년_코드	기준_분기_코드	유사_업종_점포_수	개업_율	폐업_율	주중_매출_비율	주말_매출_비율	남성_매출_비율	...	연령대_1020_매출_비율	연령대_3040_매출_비율	연령대_5060_매출_비율	집객시설_수	교통_인프라	총 상주인구 수	총_직장_인구_수	총_생활인구_수	월_평균_소득_금액	클러스터
0	DMC(디지털미디어시티)	2120098	2022	1	766	0.031353	0.016502	0.882753	0.117247	0.532487	...	0.166844	0.583242	0.249914	72	24	1285	8846.0	318202	3567709.0	3
1	DMC(디지털미디어시티)	2120098	2022	2	768	0.042763	0.032895	0.874108	0.125892	0.529133	...	0.175676	0.593012	0.231312	72	24	1285	8846.0	358782	3567709.0	3
2	DMC(디지털미디어시티)	2120098	2022	3	774	0.035714	0.027597	0.881428	0.118572	0.526975	...	0.184997	0.587134	0.227869	72	24	1285	8846.0	352688	3567709.0	3
3	DMC(디지털미디어시티)	2120098	2022	4	778	0.041935	0.037097	0.868980	0.131020	0.527500	...	0.183274	0.584706	0.232020	72	24	1285	8846.0	355095	3567709.0	3
4	가락시장역	2120234	2022	1	1203	0.016393	0.022769	0.794609	0.205391	0.546158	...	0.100028	0.389655	0.510317	62	3	2694	10268.0	1430185	3494644.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
447	홍대입구역(홍대)	2120103	2022	4	2979	0.033623	0.030369	0.617551	0.382449	0.431874	...	0.501809	0.371155	0.127036	91	22	4789	9580.0	3897068	2934520.0	0
448	화곡역	2120120	2022	1	636	0.024074	0.016667	0.761005	0.238995	0.485096	...	0.197469	0.451777	0.350755	88	16	2748	927.0	1102013	2624496.0	1
449	화곡역	2120120	2022	2	634	0.036969	0.038817	0.744602	0.255398	0.489530	...	0.208147	0.463911	0.327942	88	16	2748	927.0	1168431	2624496.0	1
450	화곡역	2120120	2022	3	627	0.016822	0.029907	0.765109	0.234891	0.494202	...	0.200589	0.445831	0.353580	88	16	2748	927.0	1163547	2624496.0	1
451	화곡역	2120120	2022	4	631	0.024164	0.020446	0.741782	0.258218	0.504047	...	0.191336	0.450515	0.358150	88	16	2748	927.0	1208993	2624496.0	1

452 rows × 21 columns

In [24]:
```
SEOUL_CLASS_1 = pd.merge(class1_isrisky, seoul, on=['상권_코드_명', '기준_년_코드', '기준_분기_코드'], how='left')
```

In [8]:
```
SEOUL_CLASS_1
```

Out[8]:

	상권_코드_명	기준_년_코드	기준_분기_코드	업종_대분류	경영_위기_비율	상권_코드	유사_업종_점포_수	개업_율	폐업_율	주중_매출_비율	...	연령대_1020_매출_비율	연령대_3040_매출_비율	연령대_5060_매출_비율	집객시설_수	교통_인프라	총 상주인구 수	총_직장_인구_수	총_생활인구_수	월_평균_소득_금액	클러스터
0	DMC(디지털미디어시티)	2022	1	외식업	0.150943	2120098	766	0.031353	0.016502	0.882753	...	0.166844	0.583242	0.249914	72	24	1285	8846.0	318202	3567709.0	3
1	DMC(디지털미디어시티)	2022	2	외식업	0.207547	2120098	768	0.042763	0.032895	0.874108	...	0.175676	0.593012	0.231312	72	24	1285	8846.0	358782	3567709.0	3
2	DMC(디지털미디어시티)	2022	3	외식업	0.142857	2120098	774	0.035714	0.027597	0.881428	...	0.184997	0.587134	0.227869	72	24	1285	8846.0	352688	3567709.0	3
3	DMC(디지털미디어시티)	2022	4	외식업	0.291667	2120098	778	0.041935	0.037097	0.868980	...	0.183274	0.584706	0.232020	72	24	1285	8846.0	355095	3567709.0	3
4	가락시장역	2022	1	외식업	0.235294	2120234	1203	0.016393	0.022769	0.794609	...	0.100028	0.389655	0.510317	62	3	2694	10268.0	1430185	3494644.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
667	홍대입구역(홍대)	2022	4	유통업	0.387755	2120103	2979	0.033623	0.030369	0.617551	...	0.501809	0.371155	0.127036	91	22	4789	9580.0	3897068	2934520.0	0
668	화곡역	2022	1	서비스업	0.166667	2120120	636	0.024074	0.016667	0.761005	...	0.197469	0.451777	0.350755	88	16	2748	927.0	1102013	2624496.0	1
669	화곡역	2022	2	서비스업	0.473684	2120120	634	0.036969	0.038817	0.744602	...	0.208147	0.463911	0.327942	88	16	2748	927.0	1168431	2624496.0	1
670	화곡역	2022	3	서비스업	0.380952	2120120	627	0.016822	0.029907	0.765109	...	0.200589	0.445831	0.353580	88	16	2748	927.0	1163547	2624496.0	1
671	화곡역	2022	4	서비스업	0.380952	2120120	631	0.024164	0.020446	0.741782	...	0.191336	0.450515	0.358150	88	16	2748	927.0	1208993	2624496.0	1

672 rows × 23 columns

In [9]:
```
SEOUL_CLASS_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 23 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   상권_코드_명         672 non-null    object
 1   기준_년_코드         672 non-null    int64
 2   기준_분기_코드        672 non-null    int64
 3   업종_대분류          672 non-null    object
 4   경영_위기_비율        672 non-null    float64
 5   상권_코드           672 non-null    int64
 6   유사_업종_점포_수      672 non-null    int64
 7   개업_율            672 non-null    float64
 8   폐업_율            672 non-null    float64
 9   주중_매출_비율        672 non-null    float64
 10  주말_매출_비율        672 non-null    float64
 11  남성_매출_비율        672 non-null    float64
 12  여성_매출_비율        672 non-null    float64
 13  연령대_1020_매출_비율  672 non-null    float64
 14  연령대_3040_매출_비율  672 non-null    float64
 15  연령대_5060_매출_비율  672 non-null    float64
 16  집객시설_수          672 non-null    int64
 17  교통_인프라          672 non-null    int64
 18  총 상주인구 수        672 non-null    int64
 19  총_직장_인구_수       672 non-null    float64
 20  총_생활인구_수        672 non-null    int64
 21  월_평균_소득_금액      672 non-null    float64
 22  클러스터            672 non-null    int64
dtypes: float64(12), int64(9), object(2)
memory usage: 120.9+ KB
```

In [66]:
```
SEOUL_CLASS_1.columns
```

Out[66]:
```
Index(['상권_코드_명', '기준_년_코드', '기준_분기_코드', '업종_대분류', '경영_위기_비율', '상권_코드',
       '유사_업종_점포_수', '개업_율', '폐업_율', '주중_매출_비율', '주말_매출_비율', '남성_매출_비율',
       '여성_매출_비율', '연령대_1020_매출_비율', '연령대_3040_매출_비율', '연령대_5060_매출_비율',
       '집객시설_수', '교통_인프라', '총 상주인구 수', '총_직장_인구_수', '총_생활인구_수', '월_평균_소득_금액'],
      dtype='object')
```

In [10]:
```
SEOUL_CLASS_1.drop(['유사_업종_점포_수', '개업_율', '폐업_율', '주중_매출_비율', '주말_매출_비율', '남성_매출_비율',
       '여성_매출_비율', '연령대_1020_매출_비율', '연령대_3040_매출_비율', '연령대_5060_매출_비율',], axis=1, inplace=True)
```

추청 매출 데이터

In [3]:
```
income = pd.read_csv('./SEOUL_raw/서울시 상권분석서비스(상권-추정매출).csv', encoding='cp949')
income
```

Out[3]:

	기준_년_코드	기준_분기_코드	상권_구분_코드	상권_구분_코드_명	상권_코드	상권_코드_명	서비스_업종_코드	서비스_업종_코드_명	분기당_매출_금액	분기당_매출_건수	...	시간대_건수~24_매출_건수	남성_매출_건수	여성_매출_건수	연령대_10_매출_건수	연령대_20_매출_건수	연령대_30_매출_건수	연령대_40_매출_건수	연령대_50_매출_건수	연령대_60_이상_매출_건수	점포_수
0	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300043	전자상거래업	4770649	1016	...	0	34	982	77	779	87	24	44	5	6
1	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300032	가전제품	1364810577	4605	...	132	2208	1920	90	1044	1278	894	666	156	8
2	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300031	가구	8512320761	3912	...	0	1421	2491	0	106	212	781	785	2028	6
3	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300028	화초	658389405	9080	...	36	4594	3400	54	2085	2412	1687	571	1183	5
4	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300026	완구	383996294	3093	...	0	1687	1349	228	857	687	1063	200	0	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
136815	2022	1	A	골목상권	2110001	이북5도청사	CS200001	일반교습학원	46710839	136	...	0	96	40	0	0	0	120	16	0	3
136816	2022	1	A	골목상권	2110001	이북5도청사	CS100010	커피-음료	20931521	2018	...	0	955	1009	0	382	342	380	581	279	*
136817	2022	1	A	골목상권	2110001	이북5도청사	CS100009	호프-간이주점	5119924	159	...	18	115	18	0	0	18	18	54	44	*
136818	2022	1	A	골목상권	2110001	이북5도청사	CS100008	분식전문점	92420017	10197	...	0	5872	3689	26	401	1191	2273	3093	2576	3
136819	2022	1	A	골목상권	2110001	이북5도청사	CS100001	한식음식점	681810073	26389	...	98	15940	7749	110	926	2092	4267	7202	9093	11

136820 rows × 80 columns

In [13]:
```
unique_values = income['서비스_업종_코드_명'].unique()
print(unique_values)
```

```
['전자상거래업' '가전제품' '가구' '화초' '완구' '운동/경기용품' '화장품' '문구' '의약품' '시계및귀금속' '안경'
 '가방' '신발' '일반의류' '반찬가게' '핸드폰' '컴퓨터및주변장치판매' '편의점' '슈퍼마켓' '피부관리실' '네일숍'
 '미용실' '스포츠클럽' '치과의원' '일반의원' '커피-음료' '분식전문점' '패스트푸드점' '제과점' '양식음식점'
 '일식음식점' '중식음식점' '한식음식점' '인테리어' '의료기기' '노래방' '고시원' '여관' '세탁소' '자동차수리'
 '골프연습장' '당구장' '한의원' '스포츠 강습' '호프-간이주점' '치킨전문점' '조명용품' '철물점' '애완동물' '섬유제품'
 '서적' '청과상' '수산물판매' '육류판매' '미곡판매' '가전제품수리' 'PC방' '예술학원' '외국어학원' '일반교습학원'
 '자동차미용' '부동산중개업' '자전거 및 기타운송장비']
```

In [4]:
```
service = ['피부관리실', '네일숍', '미용실', '스포츠클럽', '치과의원', '일반의원', '노래방', '고시원', '여관', '세탁소', '자동차수리',
          '골프연습장', '당구장', '한의원', '스포츠 강습', '가전제품수리', 'PC방', '예술학원', '외국어학원', '일반교습학원', '자동차미용', '부동산중개업']
supply = ['전자상거래업', '가전제품', '가구', '화초', '완구', '운동/경기용품', '화장품', '문구', '의약품', '시계및귀금속', '안경',
         '가방', '신발', '일반의류', '핸드폰', '컴퓨터및주변장치판매', '편의점', '슈퍼마켓', '의료기기', '조명용품', '철물점',
         '섬유제품', '서적', '청과상', '수산물판매', '육류판매', '미곡판매', '자전거 및 기타운송장비', '인테리어', '애완동물']
food = ['커피-음료', '분식전문점', '패스트푸드점', '제과점', '양식음식점', '일식음식점', '중식음식점', '한식음식점', '호프-간이주점', '치킨전문점', '반찬가게']
```

In [5]:
```
income['업종_대분류'] = income['서비스_업종_코드_명'].apply(lambda x: '서비스업' if x in service else ('유통업' if x in supply else '외식업'))
```

In [6]:
```
income['연령대_1020_매출_비율'] = income['연령대_10_매출_비율'] + income['연령대_20_매출_비율']
income['연령대_3040_매출_비율'] = income['연령대_30_매출_비율'] + income['연령대_40_매출_비율']
income['연령대_5060_매출_비율'] = income['연령대_50_매출_비율'] + income['연령대_60_이상_매출_비율']
```

In [7]:
```
selected_col = ['기준_년_코드', '기준_분기_코드', '상권_코드', '상권_코드_명', '업종_대분류', '연령대_1020_매출_비율', '연령대_3040_매출_비율', '연령대_5060_매출_비율', '주중_매출_비율',
               '주말_매출_비율', '남성_매출_비율', '여성_매출_비율']
selected_income = income[selected_col]
selected_income
```

Out[7]:

	기준_년_코드	기준_분기_코드	상권_코드	상권_코드_명	업종_대분류	연령대_1020_매출_비율	연령대_3040_매출_비율	연령대_5060_매출_비율	주중_매출_비율	주말_매출_비율	남성_매출_비율	여성_매출_비율
0	2022	4	1001496	강남 마이스 관광특구	유통업	83	12	4	55	45	3	97
1	2022	4	1001496	강남 마이스 관광특구	유통업	22	57	22	69	31	59	41
2	2022	4	1001496	강남 마이스 관광특구	유통업	3	26	72	64	36	42	59
3	2022	4	1001496	강남 마이스 관광특구	유통업	17	55	28	79	21	58	42
4	2022	4	1001496	강남 마이스 관광특구	유통업	34	63	3	45	56	63	37
...	...	...	...	...	...	...	...	...	...	...	...	...
136815	2022	1	2110001	이북5도청사	서비스업	0	86	15	100	0	73	27
136816	2022	1	2110001	이북5도청사	외식업	19	36	46	67	33	47	53
136817	2022	1	2110001	이북5도청사	외식업	0	28	73	40	60	81	19
136818	2022	1	2110001	이북5도청사	외식업	3	36	62	49	52	56	44
136819	2022	1	2110001	이북5도청사	외식업	3	26	71	57	43	67	34

136820 rows × 12 columns

In [10]:
```
avg_income = selected_income.groupby(['기준_년_코드', '기준_분기_코드', '상권_코드', '상권_코드_명', '업종_대분류']).mean().reset_index()
avg_columns = ['연령대_1020_매출_비율', '연령대_3040_매출_비율', '연령대_5060_매출_비율', '주중_매출_비율', '주말_매출_비율', '남성_매출_비율', '여성_매출_비율']
avg_income[avg_columns] = avg_income[avg_columns].apply(lambda x: round(x,2))
avg_income
```

Out[10]:

	기준_년_코드	기준_분기_코드	상권_코드	상권_코드_명	업종_대분류	연령대_1020_매출_비율	연령대_3040_매출_비율	연령대_5060_매출_비율	주중_매출_비율	주말_매출_비율	남성_매출_비율	여성_매출_비율
0	2022	1	1001491	이태원 관광특구	서비스업	15.20	65.80	19.27	77.13	22.93	54.00	46.13
1	2022	1	1001491	이태원 관광특구	외식업	40.45	45.64	14.18	60.55	39.45	52.91	47.18
2	2022	1	1001491	이태원 관광특구	유통업	18.58	53.16	28.47	75.11	24.89	49.21	50.84
3	2022	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	서비스업	26.00	51.65	22.59	81.94	18.12	38.47	61.59
4	2022	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	외식업	21.91	52.00	26.36	82.36	17.82	54.73	45.55
...	...	...	...	...	...	...	...	...	...	...	...	...
18979	2022	4	2130325	명일전통시장	외식업	5.00	38.00	56.75	65.75	34.50	44.50	55.75
18980	2022	4	2130325	명일전통시장	유통업	3.12	22.62	74.88	76.62	23.62	34.88	65.38
18981	2022	4	2130326	고덕 골목형상점가	서비스업	19.12	47.50	33.88	72.25	27.75	54.00	46.00
18982	2022	4	2130326	고덕 골목형상점가	외식업	11.50	58.00	30.67	65.00	35.00	55.00	45.17
18983	2022	4	2130326	고덕 골목형상점가	유통업	8.91	36.09	55.27	68.09	31.91	41.36	58.64

18984 rows × 12 columns

In [12]:
```
avg_income.sort_values(by='상권_코드_명')
```

Out[12]:

	기준_년_코드	기준_분기_코드	상권_코드	상권_코드_명	업종_대분류	연령대_1020_매출_비율	연령대_3040_매출_비율	연령대_5060_매출_비율	주중_매출_비율	주말_매출_비율	남성_매출_비율	여성_매출_비율
15240	2022	4	2110346	4.19민주묘지역 2번	서비스업	6.50	45.17	48.33	65.00	35.00	40.50	59.67
999	2022	1	2110346	4.19민주묘지역 2번	유통업	11.43	25.57	63.14	78.57	21.43	45.14	54.86
998	2022	1	2110346	4.19민주묘지역 2번	외식업	18.83	36.17	45.50	64.33	35.67	54.67	45.33
997	2022	1	2110346	4.19민주묘지역 2번	서비스업	2.67	58.33	39.17	67.83	32.17	64.00	36.00
5751	2022	2	2110346	4.19민주묘지역 2번	유통업	7.14	30.86	62.14	76.86	23.43	50.71	49.43
...	...	...	...	...	...	...	...	...	...	...	...	...
9353	2022	2	2130270	흑석시장	유통업	3.67	22.67	73.33	72.67	27.67	36.00	64.67
14097	2022	3	2130270	흑석시장	유통업	3.00	22.33	74.67	79.33	20.67	36.67	63.33
14096	2022	3	2130270	흑석시장	외식업	7.33	34.67	57.67	79.33	21.00	48.00	52.33
4603	2022	1	2130270	흑석시장	유통업	4.00	21.67	74.67	72.67	27.33	40.33	60.33
4602	2022	1	2130270	흑석시장	외식업	10.33	33.00	57.00	76.33	23.67	53.00	47.33

18984 rows × 12 columns

점포 데이터

In [13]:
```
store = pd.read_csv('./SEOUL_raw/서울시 상권분석서비스(상권-점포).csv', encoding='cp949')
store
```

Out[13]:

	기준_년_코드	기준_분기_코드	상권_구분_코드	상권_구분_코드_명	상권_코드	상권_코드_명	서비스_업종_코드	서비스_업종_코드_명	점포_수	유사_업종_점포_수	개업_율	개업_점포_수	폐업_률	폐업_점포_수	프랜차이즈_점포_수
0	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300043	전자상거래업	6	6	0	0	0	0	0
1	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300042	주유소	1	1	0	0	0	0	0
2	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300041	예술품	10	10	0	0	0	0	0
3	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300038	자동차부품	10	10	0	0	0	0	0
4	2022	4	U	관광특구	1001496	강남 마이스 관광특구	CS300036	조명용품	14	14	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
898063	2020	1	A	골목상권	2110001	이북5도청사	CS200001	일반교습학원	3	3	0	0	0	0	0
898064	2020	1	A	골목상권	2110001	이북5도청사	CS100010	커피-음료	2	2	0	0	0	0	0
898065	2020	1	A	골목상권	2110001	이북5도청사	CS100009	호프-간이주점	2	3	0	0	0	0	1
898066	2020	1	A	골목상권	2110001	이북5도청사	CS100008	분식전문점	3	3	0	0	0	0	0
898067	2020	1	A	골목상권	2110001	이북5도청사	CS100001	한식음식점	10	11	0	0	0	0	1

898068 rows × 15 columns

In [51]:
```
unique_values2 = store['서비스_업종_코드_명'].unique()
print(unique_values2)
```

```
['전자상거래업' '주유소' '예술품' '자동차부품' '조명용품' '인테리어' '악기' '철물점' '가전제품' '가구' '애완동물'
 '화초' '섬유제품' '완구' '운동/경기용품' '미용재료' '화장품' '문구' '서적' '의료기기' '의약품' '시계및귀금속'
 '안경' '가방' '신발' '유아의류' '일반의류' '반찬가게' '청과상' '육류판매' '미곡판매' '주류도매' '핸드폰'
 '컴퓨터및주변장치판매' '편의점' '슈퍼마켓' '여행사' '건축물청소' '통번역서비스' '사진관' 'DVD방' '부동산중개업'
 '세탁소' '피부관리실' '네일숍' '미용실' '자동차미용' '스포츠클럽' 'PC방' '세무사사무소' '기타법무서비스'
 '법무사사무소' '변호사사무소' '치과의원' '일반의원' '외국어학원' '일반교습학원' '커피-음료' '분식전문점' '패스트푸드점'
 '제과점' '양식음식점' '일식음식점' '중식음식점' '한식음식점' '재생용품 판매점' '중고차판매' '자전거 및 기타운송장비'
 '수산물판매' '의류임대' '비디오/서적임대' '녹음실' '노래방' '고시원' '여관' '가전제품수리' '자동차수리'
 '통신기기수리' '복권방' '기타오락장' '볼링장' '골프연습장' '당구장' '회계사사무소' '변리사사무소' '동물병원' '한의원'
 '스포츠 강습' '컴퓨터학원' '예술학원' '호프-간이주점' '치킨전문점' '한복점' '가정용품임대' '독서실' '게스트하우스'
 '모터사이클수리' '전자게임장' '모터사이클및부품' '중고가구']
```

In [14]:
```
service = ['여행사', '건축물청소', '통번역서비스', '사진관', 'DVD방', '부동산중개업', '세탁소', '피부관리실', '네일숍', '미용실', '자동차미용',
          '스포츠클럽', 'PC방', '세무사사무소', '기타법무서비스', '법무사사무소', '변호사사무소', '치과의원', '일반의원', '외국어학원', '일반교습학원',
          '의류임대', '비디오/서적임대', '녹음실', '노래방', '고시원', '여관', '가전제품수리', '자동차수리', '통신기기수리', '복권방', '기타오락장',
          '볼링장', '골프연습장', '당구장', '회계사사무소', '변리사사무소', '동물병원', '한의원', '스포츠 강습', '컴퓨터학원', '예술학원','한복점',
          '가정용품임대', '독서실', '게스트하우스', '모터사이클수리', '전자게임장']
supply = ['전자상거래업', '주유소', '예술품', '자동차부품', '조명용품', '인테리어', '악기', '철물점', '가전제품', '가구', '애완동물', '화초',
         '섬유제품', '완구', '운동/경기용품', '미용재료', '화장품', '문구', '서적', '의료기기', '의약품', '시계및귀금속', '안경', '가방',
         '신발', '유아의류', '일반의류', '반찬가게', '청과상', '육류판매', '미곡판매', '주류도매', '핸드폰', '컴퓨터및주변장치판매', '편의점', '슈퍼마켓',
         '재생용품 판매점', '중고차판매', '자전거 및 기타운송장비', '수산물판매', '모터사이클및부품', '중고가구']
food = ['커피-음료', '분식전문점', '패스트푸드점', '제과점', '양식음식점', '일식음식점', '중식음식점', '한식음식점', '호프-간이주점', '치킨전문점']
```

In [15]:
```
store['업종_대분류'] = store['서비스_업종_코드_명'].apply(lambda x: '서비스업' if x in service else ('유통업' if x in supply else '외식업'))
```

In [54]:
```
store.columns
```

Out[54]:
```
Index(['기준_년_코드', '기준_분기_코드', '상권_구분_코드', '상권_구분_코드_명', '상권_코드', '상권_코드_명',
       '서비스_업종_코드', '서비스_업종_코드_명', '점포_수', '유사_업종_점포_수', '개업_율', '개업_점포_수',
       '폐업_률', '폐업_점포_수', '프랜차이즈_점포_수', '업종_대분류'],
      dtype='object')
```

In [16]:
```
selected_col = ['기준_년_코드', '기준_분기_코드', '상권_코드', '상권_코드_명', '업종_대분류', '점포_수', '유사_업종_점포_수', '개업_점포_수', '폐업_점포_수']
selected_store = store[selected_col]
selected_store
```

Out[16]:

	기준_년_코드	기준_분기_코드	상권_코드	상권_코드_명	업종_대분류	점포_수	유사_업종_점포_수	개업_점포_수	폐업_점포_수
0	2022	4	1001496	강남 마이스 관광특구	유통업	6	6	0	0
1	2022	4	1001496	강남 마이스 관광특구	유통업	1	1	0	0
2	2022	4	1001496	강남 마이스 관광특구	유통업	10	10	0	0
3	2022	4	1001496	강남 마이스 관광특구	유통업	10	10	0	0
4	2022	4	1001496	강남 마이스 관광특구	유통업	14	14	0	0
...	...	...	...	...	...	...	...	...	...
898063	2020	1	2110001	이북5도청사	서비스업	3	3	0	0
898064	2020	1	2110001	이북5도청사	외식업	2	2	0	0
898065	2020	1	2110001	이북5도청사	외식업	2	3	0	0
898066	2020	1	2110001	이북5도청사	외식업	3	3	0	0
898067	2020	1	2110001	이북5도청사	외식업	10	11	0	0

898068 rows × 9 columns

In [17]:
```
sum_store = selected_store.groupby(['기준_년_코드', '기준_분기_코드', '상권_코드', '상권_코드_명', '업종_대분류']).sum().reset_index()
sum_store
```

Out[17]:

	기준_년_코드	기준_분기_코드	상권_코드	상권_코드_명	업종_대분류	점포_수	유사_업종_점포_수	개업_점포_수	폐업_점포_수
0	2020	1	1001491	이태원 관광특구	서비스업	197	202	4	9
1	2020	1	1001491	이태원 관광특구	외식업	627	678	20	35
2	2020	1	1001491	이태원 관광특구	유통업	874	905	13	30
3	2020	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	서비스업	1604	1641	19	26
4	2020	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	외식업	1380	1826	44	45
...	...	...	...	...	...	...	...	...	...
59460	2022	4	2130325	명일전통시장	외식업	16	16	1	2
59461	2022	4	2130325	명일전통시장	유통업	58	60	1	1
59462	2022	4	2130326	고덕 골목형상점가	서비스업	69	75	4	1
59463	2022	4	2130326	고덕 골목형상점가	외식업	23	38	6	2
59464	2022	4	2130326	고덕 골목형상점가	유통업	42	44	1	0

59465 rows × 9 columns

In [18]:
```
sum_store['개업률'] = round(sum_store['개업_점포_수'] / sum_store['점포_수'] * 100, 2)
sum_store['폐업률'] = round(sum_store['폐업_점포_수'] / sum_store['점포_수'] * 100, 2)
sum_store
```

Out[18]:

	기준_년_코드	기준_분기_코드	상권_코드	상권_코드_명	업종_대분류	점포_수	유사_업종_점포_수	개업_점포_수	폐업_점포_수	개업률	폐업률
0	2020	1	1001491	이태원 관광특구	서비스업	197	202	4	9	2.03	4.57
1	2020	1	1001491	이태원 관광특구	외식업	627	678	20	35	3.19	5.58
2	2020	1	1001491	이태원 관광특구	유통업	874	905	13	30	1.49	3.43
3	2020	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	서비스업	1604	1641	19	26	1.18	1.62
4	2020	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	외식업	1380	1826	44	45	3.19	3.26
...	...	...	...	...	...	...	...	...	...	...	...
59460	2022	4	2130325	명일전통시장	외식업	16	16	1	2	6.25	12.50
59461	2022	4	2130325	명일전통시장	유통업	58	60	1	1	1.72	1.72
59462	2022	4	2130326	고덕 골목형상점가	서비스업	69	75	4	1	5.80	1.45
59463	2022	4	2130326	고덕 골목형상점가	외식업	23	38	6	2	26.09	8.70
59464	2022	4	2130326	고덕 골목형상점가	유통업	42	44	1	0	2.38	0.00

59465 rows × 11 columns

In [19]:
```
sum_store.drop(['점포_수', '개업_점포_수', '폐업_점포_수'], axis=1, inplace=True)
```

In [24]:
```
sum_store
```

Out[24]:

	기준_년_코드	기준_분기_코드	상권_코드	상권_코드_명	업종_대분류	유사_업종_점포_수	개업률	폐업률
0	2020	1	1001491	이태원 관광특구	서비스업	202	2.03	4.57
1	2020	1	1001491	이태원 관광특구	외식업	678	3.19	5.58
2	2020	1	1001491	이태원 관광특구	유통업	905	1.49	3.43
3	2020	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	서비스업	1641	1.18	1.62
4	2020	1	1001492	명동 남대문 북창동 다동 무교동 관광특구	외식업	1826	3.19	3.26
...	...	...	...	...	...	...	...	...
59460	2022	4	2130325	명일전통시장	외식업	16	6.25	12.50
59461	2022	4	2130325	명일전통시장	유통업	60	1.72	1.72
59462	2022	4	2130326	고덕 골목형상점가	서비스업	75	5.80	1.45
59463	2022	4	2130326	고덕 골목형상점가	외식업	38	26.09	8.70
59464	2022	4	2130326	고덕 골목형상점가	유통업	44	2.38	0.00

59465 rows × 8 columns

Merge

In [26]:
```
merged_SEOUL_CLASS_1 = pd.merge(SEOUL_CLASS_1, avg_income, on=['기준_년_코드', '기준_분기_코드', '상권_코드', '업종_대분류'], how = 'left')
merged_SEOUL_CLASS_1 = pd.merge(merged_SEOUL_CLASS_1, sum_store, on=['기준_년_코드', '기준_분기_코드', '상권_코드', '업종_대분류'], how = 'left')
merged_SEOUL_CLASS_1
```

Out[26]:

	상권_코드_명_x	기준_년_코드	기준_분기_코드	업종_대분류	경영_위기_비율	상권_코드	유사_업종_점포_수_x	개업_율	폐업_율	주중_매출_비율_x	...	연령대_3040_매출_비율_y	연령대_5060_매출_비율_y	주중_매출_비율_y	주말_매출_비율_y	남성_매출_비율_y	여성_매출_비율_y	상권_코드_명	유사_업종_점포_수_y	개업률	폐업률
0	DMC(디지털미디어시티)	2022	1	외식업	0.150943	2120098	766	0.031353	0.016502	0.882753	...	62.10	17.60	87.80	12.40	57.50	42.50	DMC(디지털미디어시티)	421	4.08	2.04
1	DMC(디지털미디어시티)	2022	2	외식업	0.207547	2120098	768	0.042763	0.032895	0.874108	...	60.80	18.70	86.40	13.60	56.40	43.80	DMC(디지털미디어시티)	424	6.08	4.39
2	DMC(디지털미디어시티)	2022	3	외식업	0.142857	2120098	774	0.035714	0.027597	0.881428	...	61.00	18.00	85.80	14.40	56.40	43.70	DMC(디지털미디어시티)	426	5.00	4.67
3	DMC(디지털미디어시티)	2022	4	외식업	0.291667	2120098	778	0.041935	0.037097	0.868980	...	60.40	18.80	83.00	17.00	56.80	43.20	DMC(디지털미디어시티)	427	6.98	6.64
4	가락시장역	2022	1	외식업	0.235294	2120234	1203	0.016393	0.022769	0.794609	...	49.73	35.09	77.82	22.27	61.82	38.18	가락시장역	294	2.74	4.57
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
667	홍대입구역(홍대)	2022	4	유통업	0.387755	2120103	2979	0.033623	0.030369	0.617551	...	30.11	31.68	71.47	28.63	43.63	56.37	홍대입구역(홍대)	874	3.14	1.93
668	화곡역	2022	1	서비스업	0.166667	2120120	636	0.024074	0.016667	0.761005	...	48.25	30.00	81.12	18.94	49.19	50.94	화곡역	254	1.22	3.27
669	화곡역	2022	2	서비스업	0.473684	2120120	634	0.036969	0.038817	0.744602	...	46.00	34.33	81.27	18.93	54.93	45.13	화곡역	258	3.23	1.61
670	화곡역	2022	3	서비스업	0.380952	2120120	627	0.016822	0.029907	0.765109	...	43.60	38.27	81.40	18.67	52.40	47.87	화곡역	256	1.21	2.02
671	화곡역	2022	4	서비스업	0.380952	2120120	631	0.024164	0.020446	0.741782	...	41.80	41.67	80.47	19.73	52.93	47.27	화곡역	256	2.83	2.43

672 rows × 35 columns

최종데이터 imputaton

In [69]:
```
merged_SEOUL_CLASS_1.columns
```

Out[69]:
```
Index(['상권_코드_명', '기준_년_코드', '기준_분기_코드', '업종_대분류', '경영_위기_비율', '상권_코드',
       '집객시설_수', '교통_인프라', '총 상주인구 수', '총_직장_인구_수', '총_생활인구_수', '월_평균_소득_금액',
       '연령대_1020_매출_비율', '연령대_3040_매출_비율', '연령대_5060_매출_비율', '주중_매출_비율',
       '주말_매출_비율', '남성_매출_비율', '여성_매출_비율', '유사_업종_점포_수', '개업률', '폐업률'],
      dtype='object')
```

In [71]:
```
merged_SEOUL_CLASS_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   상권_코드_명         672 non-null    object
 1   기준_년_코드         672 non-null    int64
 2   기준_분기_코드        672 non-null    int64
 3   업종_대분류          672 non-null    object
 4   경영_위기_비율        672 non-null    float64
 5   상권_코드           672 non-null    int64
 6   집객시설_수          672 non-null    int64
 7   교통_인프라          672 non-null    int64
 8   총 상주인구 수        672 non-null    int64
 9   총_직장_인구_수       672 non-null    float64
 10  총_생활인구_수        672 non-null    int64
 11  월_평균_소득_금액      672 non-null    float64
 12  연령대_1020_매출_비율  668 non-null    float64
 13  연령대_3040_매출_비율  668 non-null    float64
 14  연령대_5060_매출_비율  668 non-null    float64
 15  주중_매출_비율        668 non-null    float64
 16  주말_매출_비율        668 non-null    float64
 17  남성_매출_비율        668 non-null    float64
 18  여성_매출_비율        668 non-null    float64
 19  유사_업종_점포_수      672 non-null    int64
 20  개업률             672 non-null    float64
 21  폐업률             672 non-null    float64
dtypes: float64(12), int64(8), object(2)
memory usage: 115.6+ KB
```

In [75]:
```
merged_SEOUL_CLASS_1[merged_SEOUL_CLASS_1.isna().any(axis=1)]
```

Out[75]:

	상권_코드_명	기준_년_코드	기준_분기_코드	업종_대분류	경영_위기_비율	상권_코드	집객시설_수	교통_인프라	총 상주인구 수	총_직장_인구_수	...	연령대_1020_매출_비율	연령대_3040_매출_비율	연령대_5060_매출_비율	주중_매출_비율	주말_매출_비율	남성_매출_비율	여성_매출_비율	유사_업종_점포_수	개업률	폐업률
204	마곡역(마곡)	2022	1	서비스업	0.169231	2120118	36	11	1725	763.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	351	11.18	1.18
206	마곡역(마곡)	2022	2	서비스업	0.217391	2120118	36	11	1725	763.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	376	10.16	3.02
208	마곡역(마곡)	2022	3	서비스업	0.202532	2120118	36	11	1725	763.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	390	6.86	3.43
210	마곡역(마곡)	2022	4	서비스업	0.256410	2120118	36	11	1725	763.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	411	7.02	2.01

4 rows × 22 columns

In [26]:
```
# 과거 데이터를 구할 수 없어 결측치가 있는 네개 행만 제거
final_SEOUL_CLASS_1 = merged_SEOUL_CLASS_1.dropna()
```

In [27]:
```
final_SEOUL_CLASS_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 668 entries, 0 to 671
Data columns (total 23 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   상권_코드_명         668 non-null    object
 1   기준_년_코드         668 non-null    int64
 2   기준_분기_코드        668 non-null    int64
 3   업종_대분류          668 non-null    object
 4   경영_위기_비율        668 non-null    float64
 5   상권_코드           668 non-null    int64
 6   집객시설_수          668 non-null    int64
 7   교통_인프라          668 non-null    int64
 8   총 상주인구 수        668 non-null    int64
 9   총_직장_인구_수       668 non-null    float64
 10  총_생활인구_수        668 non-null    int64
 11  월_평균_소득_금액      668 non-null    float64
 12  클러스터            668 non-null    int64
 13  연령대_1020_매출_비율  668 non-null    float64
 14  연령대_3040_매출_비율  668 non-null    float64
 15  연령대_5060_매출_비율  668 non-null    float64
 16  주중_매출_비율        668 non-null    float64
 17  주말_매출_비율        668 non-null    float64
 18  남성_매출_비율        668 non-null    float64
 19  여성_매출_비율        668 non-null    float64
 20  유사_업종_점포_수      668 non-null    int64
 21  개업률             668 non-null    float64
 22  폐업률             668 non-null    float64
dtypes: float64(12), int64(9), object(2)
memory usage: 125.2+ KB
```

In [28]:
```
final_SEOUL_CLASS_1.to_csv('../data/SEOUL_CLASS_1.csv', index=False)
```

In [42]:
```
credit_class_2 = pd.read_csv('../data/CREDIT_CLASS_2.csv')
credit_class_2.columns
```

Out[42]:
```
Index(['trdar_nm', 'year', 'quarter', 'class_2_name', 'average(age)',
       'average(duration)', 'average(is_franchise)',
       'average(business_square_size)', 'average(is_risky)',
       'average(monthly_rental_fee)', 'average(regular_employees_count)',
       'average(rental_deposit)', 'average(sum_customer_cnt)',
       'average(sum_new_customer_cnt)', 'average(sum_purchase_card)',
       'average(sum_purchase_cash)', 'average(sum_purchase_invoice)',
       'average(sum_sales_card)', 'average(sum_sales_delivery)',
       'average(sum_sales_invoice)', 'average(sum_weekend_sales_card)',
       'average(sum_weekend_sales_delivery)'],
      dtype='object')
```

In [52]:
```
credit_class_1 = pd.read_excel('../data/CREDIT_CLASS_1.xlsx')
credit_class_1 = credit_class_1[credit_class_1['year'] == 2022]
credit_class_1
```

Out[52]:

	trdar_nm	year	quarter	class_1_name	average(age)	average(duration)	average(is_franchise)	average(business_square_size)	average(is_risky)	average(monthly_rental_fee)	...	average(sum_customer_cnt)	average(sum_new_customer_cnt)	average(sum_purchase_card)	average(sum_purchase_cash)	average(sum_purchase_invoice)	average(
0	DMC(디지털미디어시티)	2022	1	외식업	41.952830	5.247170	0.169811	40.290215	0.150943	1.859623	...	21.962075	5.313774	4.283320	0.393568	7.738196	
1	DMC(디지털미디어시티)	2022	2	외식업	42.252830	5.215094	0.169811	39.639838	0.207547	1.909811	...	26.239808	6.588654	3.694610	0.533263	7.924871	
2	DMC(디지털미디어시티)	2022	3	외식업	43.504082	5.420408	0.183673	40.630641	0.142857	1.953469	...	24.373878	6.333061	3.634459	0.528276	8.867591	
3	DMC(디지털미디어시티)	2022	4	외식업	44.156250	5.506250	0.187500	40.032112	0.291667	1.921250	...	22.805625	5.896250	3.356604	0.511717	9.789668	
6	가락시장역	2022	1	외식업	44.205882	5.305882	0.529412	95.744118	0.235294	3.812941	...	4.714706	3.259412	3.810062	0.079800	22.236505	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
989	홍대입구역(홍대)	2022	4	유통업	38.300000	7.065306	0.061224	85.057612	0.387755	3.416122	...	1.780227	1.360000	3.823159	0.392102	34.569925	
996	화곡역	2022	1	서비스업	43.916667	4.566667	0.000000	502.754467	0.166667	11.800000	...	8.894444	2.745000	39.839828	1.298314	50.917351	
997	화곡역	2022	2	서비스업	44.452632	4.384211	0.000000	482.594758	0.473684	11.284211	...	9.993684	3.061053	51.727521	1.030745	49.884303	
998	화곡역	2022	3	서비스업	45.371429	4.071429	0.000000	448.035257	0.380952	10.400000	...	9.550000	2.933333	56.513540	1.113395	55.321360	
999	화곡역	2022	4	서비스업	45.371429	4.071429	0.000000	448.035257	0.380952	10.400000	...	8.536190	2.605238	78.228019	0.870120	67.192633	

672 rows × 22 columns

In [ ]: