광주광역시

광주 인공지능사관학교

슬기로운
소비데이터

데이터 시각화 및 Discovery

2기 수료생 홍인영

목차

PROCESS

분석 목표

[데이터 분석 과정의 Discovery]
일부 코드 등 설명의 Hard Skills, 분석과정의 Soft Skills 등을
활용해 사회적 문제/이슈 를 발견하고 이를 해결하는 방안제시

개요 / 전처리

데이터 와 진행방향에 대한 간단한 설명과
전처리 과정을 설명

EDA / 시각화

데이터를 시각화하고 탐색하며
이해 및 분석 방향을 설정

분석 / 문제

데이터를 분석하고
발견한 문제점 제시

결론 /해결방안

아쉬운 점과 분석과정을 겪으며
생성한 idea 및 solution

광주광역시

개요/전처리

광주광역시

광주 인공지능사관학교

DATA_INFO

RELATIONSHIP DIAGRAM

DATA 전처리

DATA PREPROCESSING

데이터 전처리란?

특정 분석에 적합하게 데이터를 가공하는 작업

CNT_ 접두사 제거

전

'CNT_DAY_월'

'CNT_온라인쇼핑'

후

'DAY_월'

'온라인쇼핑'

'CNT_' 형태로 붙어 있는 접두사를 제거

데이터의 기존 Column 에서 'DAY_ ' 와 같이 접두사가 붙어있는 경우 어려움이 존재

```
def drop_prefix(self, prefix):
    self.columns = self.columns.str.lstrip(prefix)
    return self

pd.core.frame.DataFrame.drop_prefix = drop_prefix
```

광주광역시

DATA 전처리

DATA PREPROCESSING

정규화 란?

특이점 또는 불연속점을 지우는 과정

추정소득 Column 변경

전

A : 3천미만
B : 3 ~ 5천만원 미만
C : 5 ~ 7천만원 미만
D : 7천만 ~ 1억 미만
E : 1억 이상

문자형은 구분의 의미로 사용함
숫자형(1~5)으로 변환 후에도 구분가능, 향후 정규화를 진행한 수치형 Column들과 상관관계 파악에 용이

후

A	1
B	2
C →	3
D	4
E	5

```
clean_원본['추정소득']=clean_원본['추정소득'].apply(lambda x:x.replace("A","1"))
clean_원본['추정소득']=clean_원본['추정소득'].apply(lambda x:x.replace("B","2"))
clean_원본['추정소득']=clean_원본['추정소득'].apply(lambda x:x.replace("C","3"))
clean_원본['추정소득']=clean_원본['추정소득'].apply(lambda x:x.replace("D","4"))
clean_원본['추정소득']=clean_원본['추정소득'].apply(lambda x:x.replace("E","5"))

print("전체데이터 중 예시5개")
clean_원본.head()
```

수치형DATA의 **상관관계**

HEAT MAP

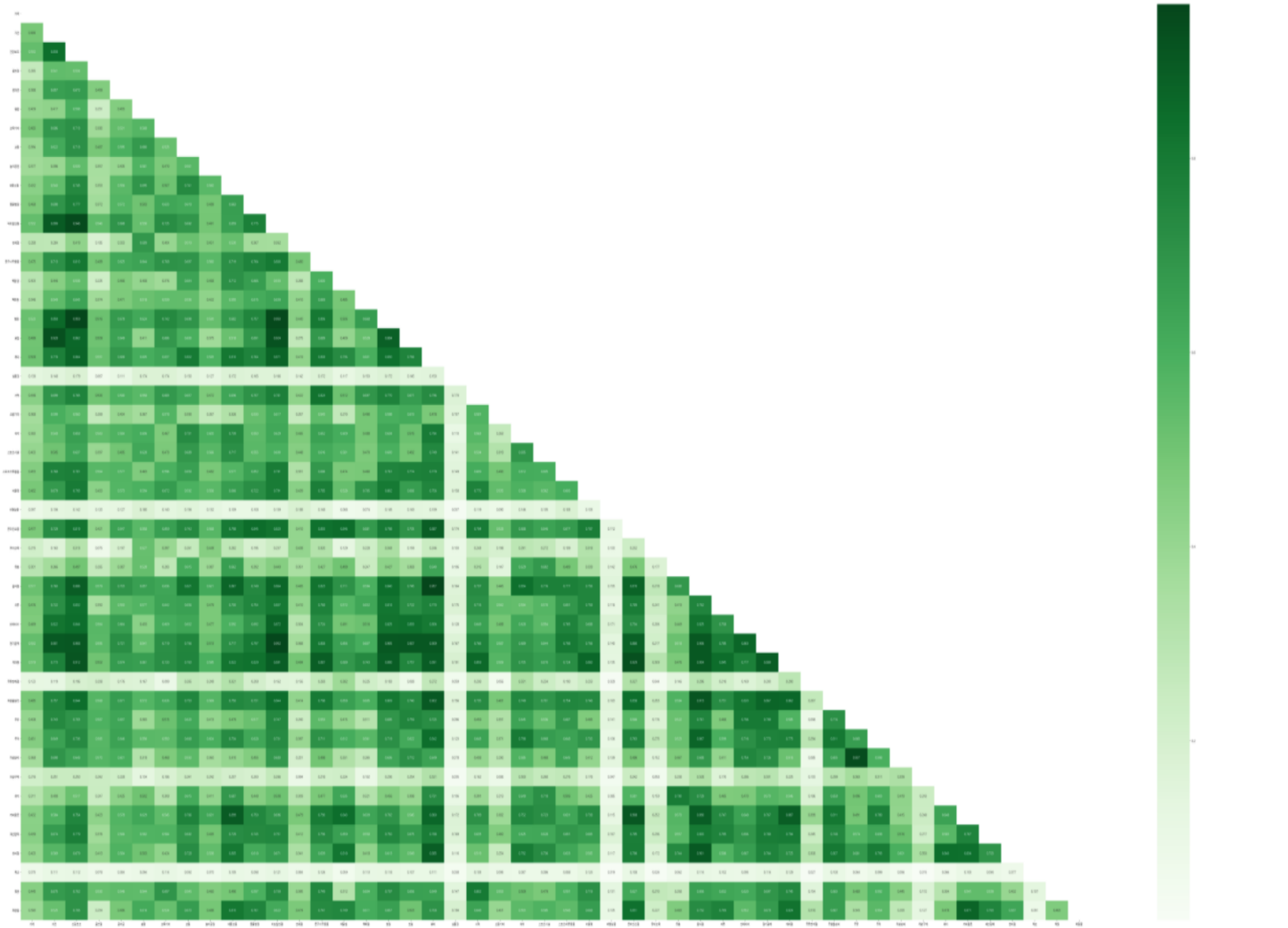
수치형 변수(소비종류)들 간의 상관관계를 알아보고자 히트맵 생성

```
def make_heatmap(inputdata):

    mask = np.zeros_like(inputdata.corr())
    mask[np.triu_indices_from(mask)] = True
    plt.rcParams["figure.figsize"] = (100,50) # 그림 크기 조정
    plt.rc('font', size=10)          # 기본 폰트 크기
    plt.rc('axes', labelsize=10)     # x,y축 label 폰트 크기
    plt.rc('xtick', labelsize=10)    # x축 눈금 폰트 크기
    plt.rc('ytick', labelsize=10)    # y축 눈금 폰트 크기
    plt.rc('legend', fontsize=10)    # 범례 폰트 크기
    plt.rc('figure', titlesize=10) # figure title 폰트 크기

    return sb.heatmap(data = inputdata.corr(), mask=mask, annot=True,
                      fmt = '.3f', linewidths=0, cmap='Greens')
```

이 때, 기존의 데이터 값이 서로 매우 상이하기 때문에 정규화를 진행



Min-Max 정규화

MIN-MAX NORMALIZATION

$$\frac{x - \min(x)}{\max(x) - \min(x)}$$

특성들의 범위가 0 에서 1 사이 (0 ~ 1)가 되도록 비례적으로 맞춤

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
scaler.fit(사용처)
scaled_사용처 = scaler.transform(사용처)

MinMax_사용처 = pd.DataFrame(data=scaled_사용처,columns = 사용처.columns)
```

이 때, 정규화는 정규분포가 아니라 단순히 특성들의 범위를 맞추는 것을 뜻함

RobustScaler 도 사용했지만, 별 차이 없었다

**RobustScaler?
평균과 분산 대신, 중간값과 사분위 값을 조정하여 아주 동 떨어진 데이터를 제거

	가례	가전	건강보조	골프장	공과금	관람	교육기타	교통	놀이공원	대중교통	...	주차	차량
0	0.000000	0.000000	0.016432	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.008289	...	0.000000	0.000
1	0.000000	0.000000	0.015962	0.000000	0.000000	0.016207	0.000000	0.024036	0.000000	0.008042	...	0.000000	0.000
2	0.000000	0.000000	0.015023	0.000000	0.000000	0.000000	0.000000	0.005590	0.000000	0.014012	...	0.000000	0.000
3	0.000000	0.030702	0.013146	0.000000	0.000000	0.000000	0.000000	0.015092	0.000000	0.002072	...	0.000000	0.000
4	0.000000	0.000000	0.013146	0.000000	0.000000	0.000000	0.000000	0.006708	0.000000	0.003750	...	0.000000	0.000
...	...	...	...	...	...	...	...	...	...	...	...	...	
254389	0.000000	0.151316	0.125352	0.000000	0.045205	0.090762	0.128889	0.094466	0.043478	0.098081	...	0.164659	0.025
254390	0.086331	0.412281	0.269484	0.107468	0.058904	0.066451	0.111111	0.099497	0.079051	0.057181	...	0.287149	0.082
254391	0.107914	0.412281	0.335681	0.087432	0.087671	0.058347	0.204444	0.129122	0.086957	0.062065	...	0.180723	0.096
254392	0.129496	0.300439	0.390141	0.038251	0.113014	0.170178	0.257778	0.252096	0.221344	0.277665	...	0.315261	0.077
254393	0.187050	0.282895	0.290610	0.063752	0.121233	0.173420	0.000000	0.440470	0.162055	0.337510	...	0.510040	0.143

DATA의 한계

LIMITATIONS OF DATA|ANALYSIS

소비기간의 공백
Gap in Consumption Period

2020.06~2020.11 소비기간의 공백

- 수집기간이 전체연속이 아닌 일부 공백기간이 존재
- 시계열 분석을 시도하지못해 계절성 등의 Discovery 불가

표본의 다양성
Specimen Diversity

전국적으로 다양한 고객

- 총 254개의 개별적인 지역에서 다양한 종류의 고객들이 수집됨
- 서울을 비롯한 수도권 지역이 비교적 많았지만, 다양한 지역의 약 25 만명의 사람들을 대상으로 하여 전국 단위의 분석시도

다양한 변수
Various Variables

48종의 소비 변수

- 다양한 수치형 변수로 인해 상관분석에 어려움이 있음

광주광역시

EDA / 시각화

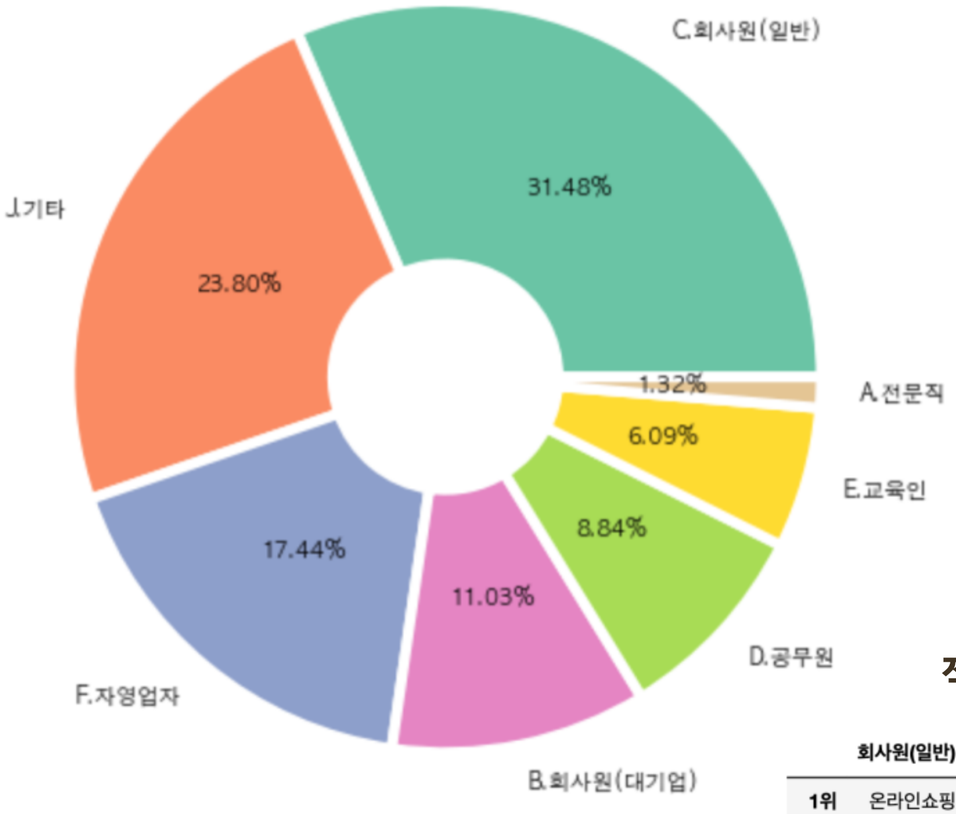
광주광역시

광주 인공지능사관학교

직업별 분석

ANALYSIS BY OCCUPATION

DATA 상의 직업별 비율



1위 : 회사원(일반)
2위 : 기타
3위 : 자영업자
4위 : 회사원(대기업)
5위 : 공무원
6위 : 교육인
7위 : 전문직

직업군이 소비한 품목들을 평균화

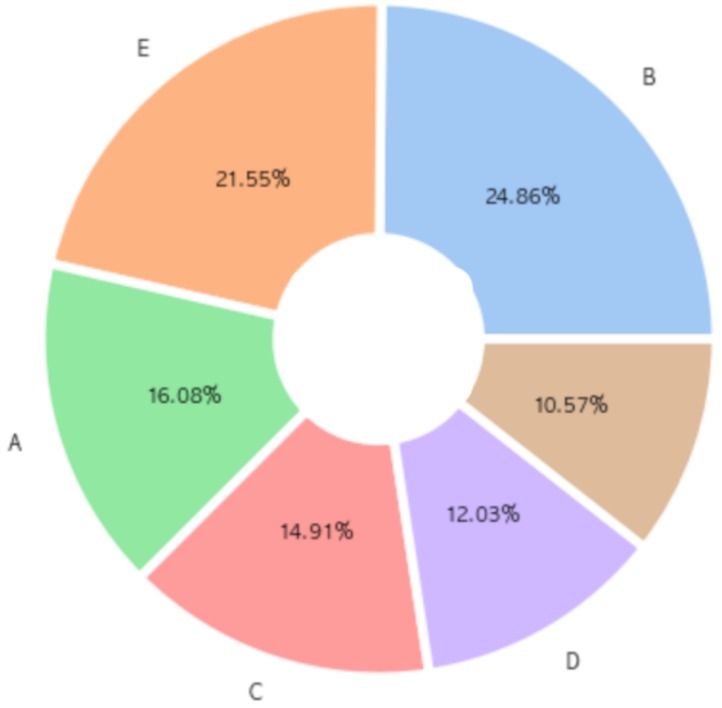
직업별로 카드사용 사용처가 많은 순위

	회사원(일반)	기타	자영업자	회사원(대기업)	공무원	교육인	전문직
1위	온라인쇼핑	마트할인점	마트할인점	온라인쇼핑	온라인쇼핑	온라인쇼핑	온라인쇼핑
2위	마트할인점	온라인쇼핑	온라인쇼핑	음식점	마트할인점	마트할인점	음식점
3위	음식점	음식점	음식점	대중교통	음식점	음식점	마트할인점
4위	편의점	편의점	편의점	마트할인점	편의점	대중교통	대중교통
5위	대중교통	대중교통	주유	편의점	대중교통	편의점	편의점
6위	커피음료	정기결제	대중교통	커피음료	정기결제	커피음료	커피음료
7위	배달앱	병원	정기결제	배달앱	커피음료	정기결제	백화점
8위	주유	건강보조	커피음료	정기결제	주유	배달앱	제과점
9위	정기결제	주유	병원	주유	병원	병원	정기결제
10위	병원	커피음료	배달앱	병원	건강보조	주유	주유

연령성별 분석

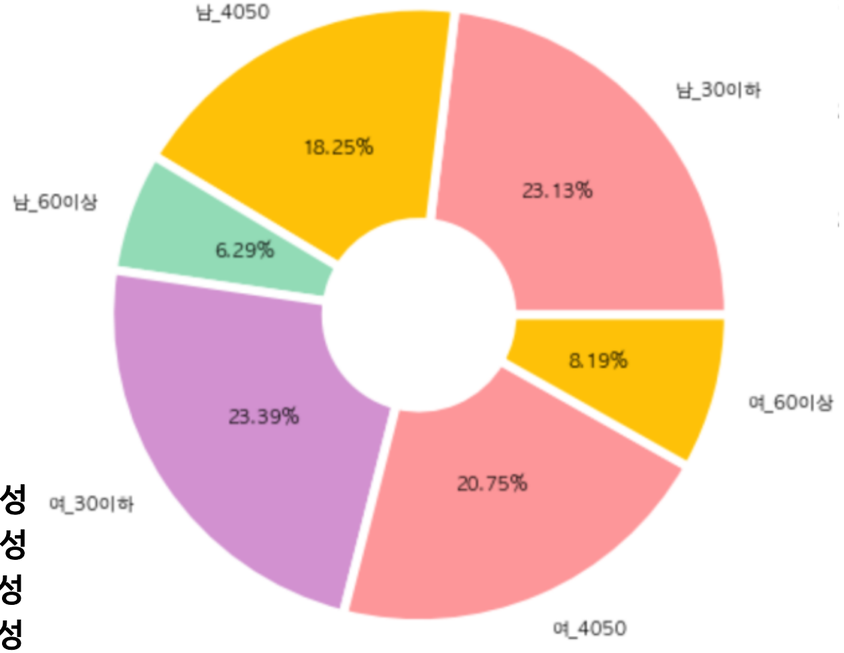
ANALYSIS BY AGE|GENDER

DATA 상의 연령성별 비율



1위 : 40~50대 남성
2위 : 40~50대 여성
3위 : 30대 이하 남성
4위 : 60대 이상 남성
5위 : 30대 이하 여성
6위 : 60대 이상 여성

연령성별_구매건 비율

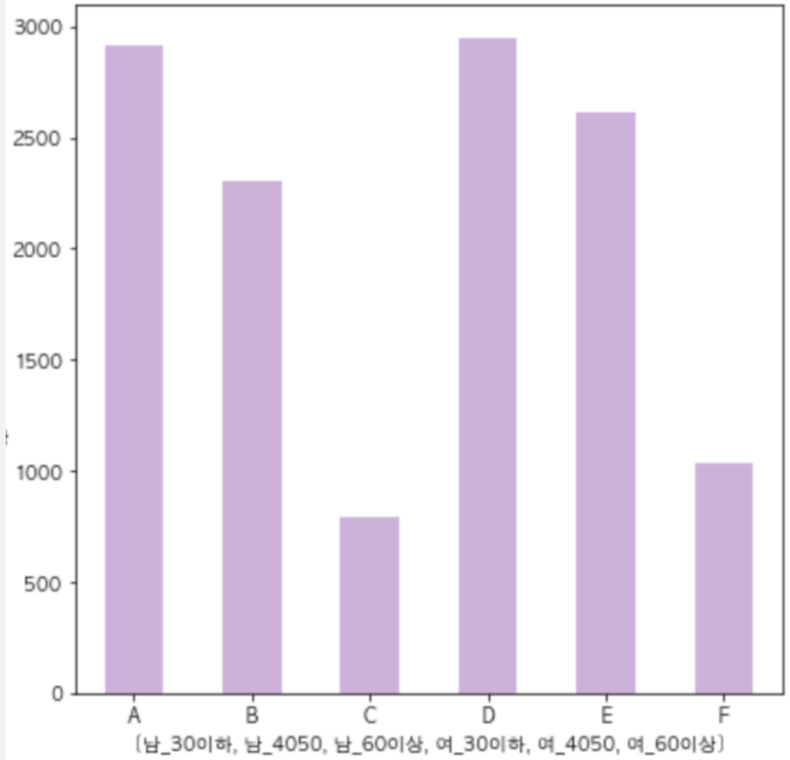


1위 : 30대 이하 여성
2위 : 30대 이하 남성
3위 : 40~50대 여성
4위 : 40~50대 남성
5위 : 60대 이상 여성
6위 : 60대 이상 남성

연령성별 분석

ANALYSIS BY AGE|GENDER

(청년 > 장년 | 여성 > 남성)



연령성별로 카드사용 사용처가 많은 순위

	남_4050	여_4050	남_30이하	남_60이상	여_30이하	여_60이상
1위	마트할인점	온라인쇼핑	온라인쇼핑	마트할인점	온라인쇼핑	마트할인점
2위	음식점	마트할인점	편의점	음식점	대중교통	온라인쇼핑
3위	온라인쇼핑	음식점	음식점	주유	음식점	음식점
4위	편의점	대중교통	대중교통	대중교통	편의점	병원
5위	대중교통	편의점	마트할인점	편의점	마트할인점	대중교통
6위	주유	정기결제	배달앱	온라인쇼핑	배달앱	건강보조
7위	정기결제	커피음료	커피음료	병원	커피음료	편의점
8위	커피음료	병원	주유	건강보조	정기결제	정기결제
9위	병원	건강보조	정기결제	정기결제	제과점	보험
10위	건강보조	보험	취미	보험	병원	주유

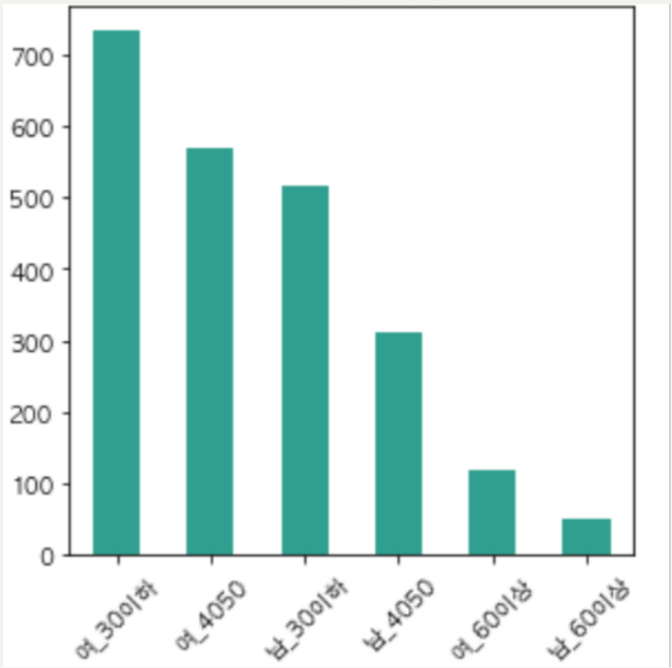
온라인쇼핑 | 마트할인점

INFORMATION

온라인은 청년층 | 오프라인은 장년층

아래 그래프를 통해 온라인쇼핑, 마트할인점에 대해서 청-장년 층의 순위가 상반되는 모양

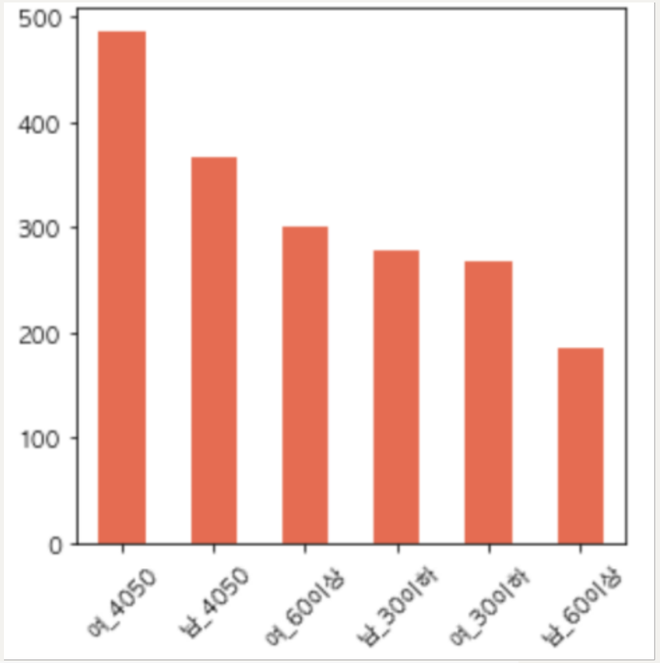
청년층에서 온라인쇼핑을 더 많이 찾음




```
plt.subplot(221)
온라인쇼핑이용층 = 연령성별사용처.transpose().sort_values(by="온라인쇼핑",
                                                    ascending=False).head(10)['온라인쇼핑']

온라인쇼핑이용층.plot.bar(rot=45,color='#2a9d8f')
```

중.장년 층 여성이 많이 찾음




```
plt.subplot(222)
마트할인점이용층 = 연령성별사용처.transpose().sort_values(by="마트할인점",
                                                    ascending=False).head(10)['마트할인점']

마트할인점이용층.plot.bar(rot=45,color='#e76f51')
```

분석 / 문제

광주광역시

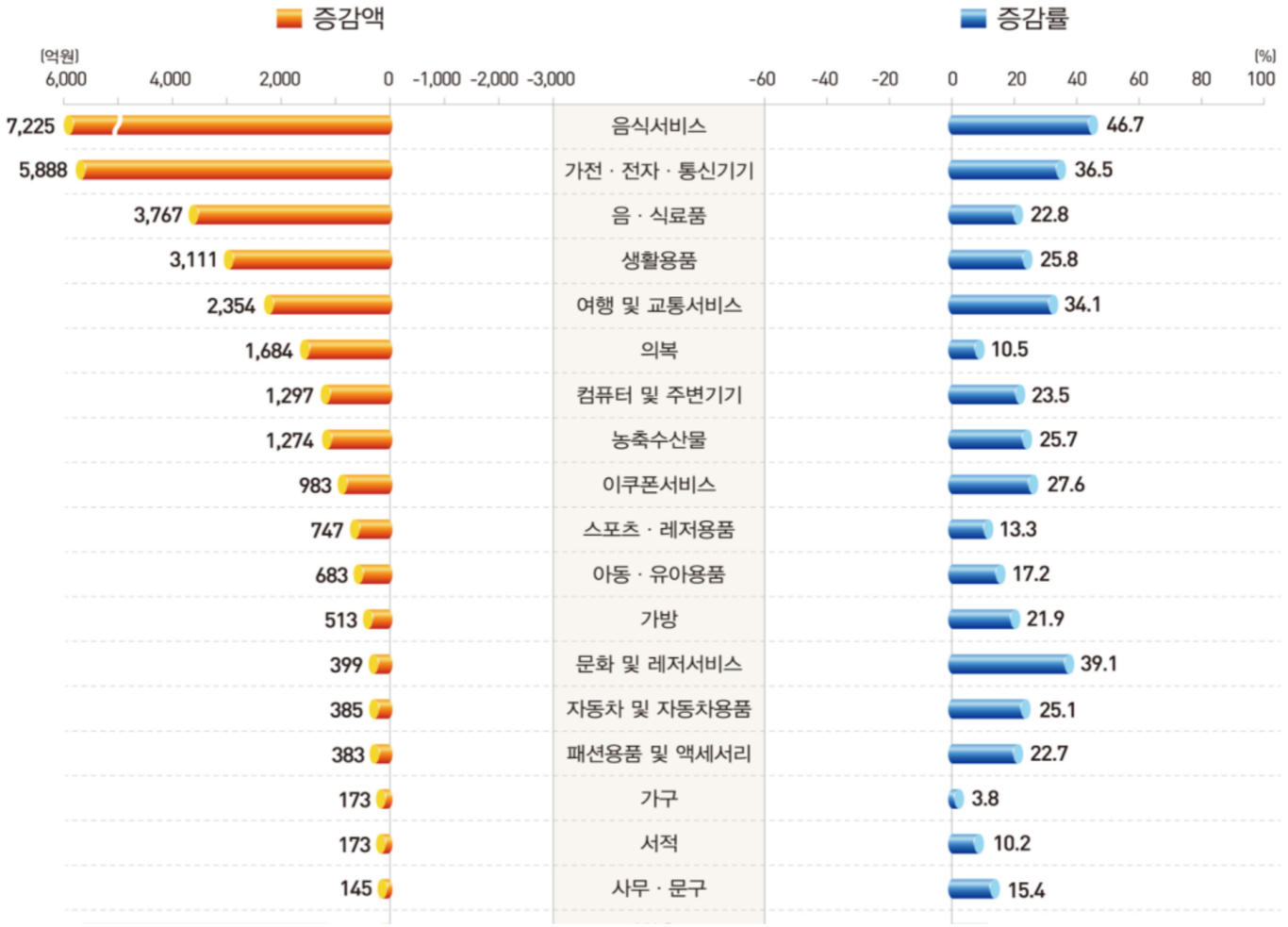
온라인쇼핑 **동향**

ONLINE SHOPPING TRENDS

* 2021.12.3 최신 조사에 따르면
약 16조 원의 온라인쇼핑 거래총액에서
1위인 음식서비스는 2조 2,688억원,
3위인 음식료품은 2조 259억원 이라는 높은 비중을 차지함

* 전년동월대비 증감은
음식서비스가 7,225억원 증가로 약 46.7%증가
음식료품이 3,767억원 증가로 약 22.8%증가
하여 소비의 많은 비중이 먹거리로 향하고 있음

상품군별 온라인쇼핑 거래액 증감액/증감률



통계청(온라인쇼핑동향조사).온라인쇼핑 동향.(2021.12.3)
이민경,김성규

문제 탐색

EXPLORE THE PROBLEM

* 2021.12.3 최신 조사에 따르면
약 16조 원의 온라인쇼핑 거래총액에서
1위인 음식서비스는 2조 2,688억원,
3위인 음식료품은 2조 259억원 이라는 높은 비중을 차지함

* 전년동월대비 증감은
음식서비스가 7,225억원 증가로 약 46.7%증가
음식료품이 3,767억원 증가로 약 22.8%증가
하여 소비의 많은 비중이 먹거리로 향하고 있음

환경부.음식물핸드북.(2013. 4)
가현기획

음식물쓰레기란?

식품의 생산, 유통, 가공, 조리과정에서 발생하는 농·수·축산물 쓰레기와 먹고 남긴 음식찌꺼기

음식물쓰레기는 푸짐한 상차림과 국물 음식을 즐기는 우리나라 음식문화와 인구증가, 생활수준 향상, 식생활의 고급화 등으로 인해 매년 3%가량 증가

국내에서 발생하는 음식물쓰레기는 하루 1만 4천여 톤으로, 전체 쓰레기 발생량의 28.7%를 차지

HOW?

WHERE?

주 발생원은 가정/소형음식점!
전체의 약 70%

문제 제시

RAISE AN ISSUE

전국 | 광주 1인 가구는 해마다 증가하는 추세

1인 가구 에서 '온라인쇼핑'과 '배달앱' 이용이 증가

코로나로 인해 이용률이 증가에 가속화

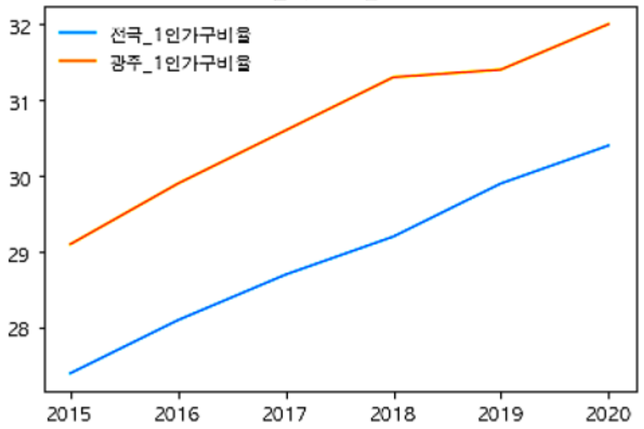
적정량의 음식물 섭취가 힘듬

음식물 쓰레기량 지속적으로 증가

* 참고문헌에 따르면 음식물 소량구매가 음식물 쓰레기 감량에 도움을 줌

다항로짓 추정결과에 의하면 일반 수퍼마켓에서 식품을 구입하는 소비자들 은 소량으로 자주 구입하고 유통기한을 확인하는 방식으로 음식물 쓰레기 감소에 노력하는 것으로 분석

전국|광주 1인 가구 비율(%)



충남대학교 농업과학연구소.가정내 음식물 쓰레기 감량을 위한 소비자 특성별 행위와 요인분석.(2012. 6)
한재환, 황윤재

광주광역시

1인 가구 | 다 가구

ANALYSIS BY ONE-PERSON HOUSEHOLDS

1인 가구 | 다 가구 얼마나?

아래 카테고리 별로 각각 1인 가구의 비율을 확인
1인 가구의 경우

직업별 1인 가구
By Occupation

각 직업군에 1인가구가 얼마나 있는지?

연령성별 1인 가구
By Age and gender

각 연령|성별로 1인가구가 얼마나 있는지?

소득별 1인 가구
By Income

각 소득층별로 1인가구가 얼마나 있는지?

1인 가구 | 다 가구

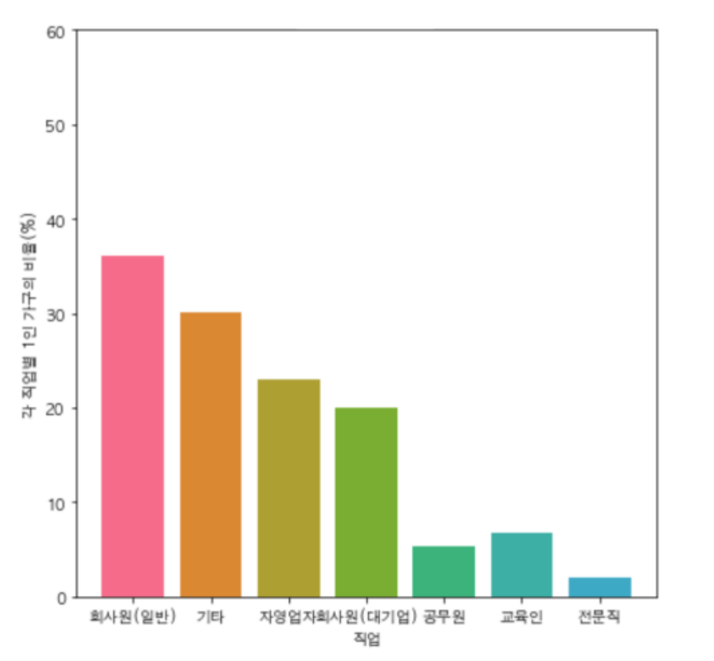
INFORMATION

소득별 = 명확 | 직업별 = 일반직다수 | 연령성별 = 대체로 비슷

소득별로 구분된 1인 가구의 각 비율이 눈에 띄게 계층적임을 확인
소득에 대해 직업은 필요조건을 만족하기 때문에 연관된 양상
연령성별은 대부분 서로 큰 차이 없음

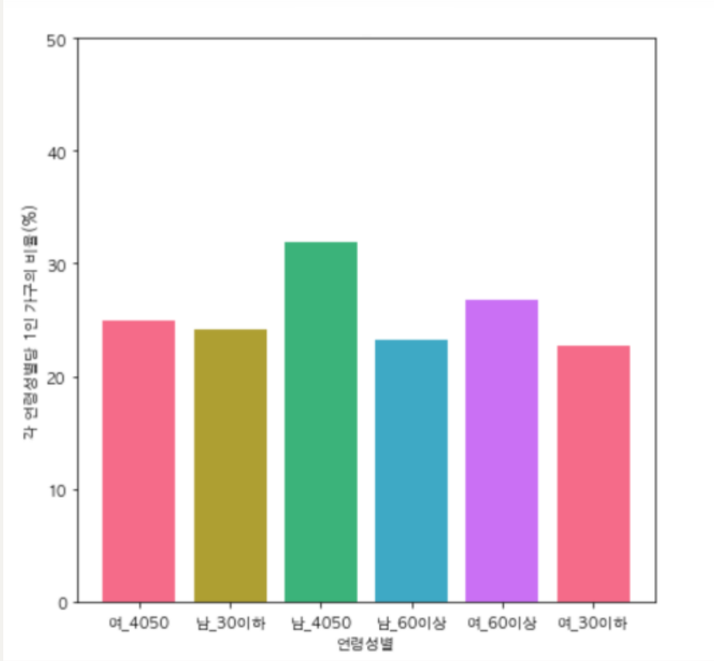
직업별 1인 가구

By Occupation



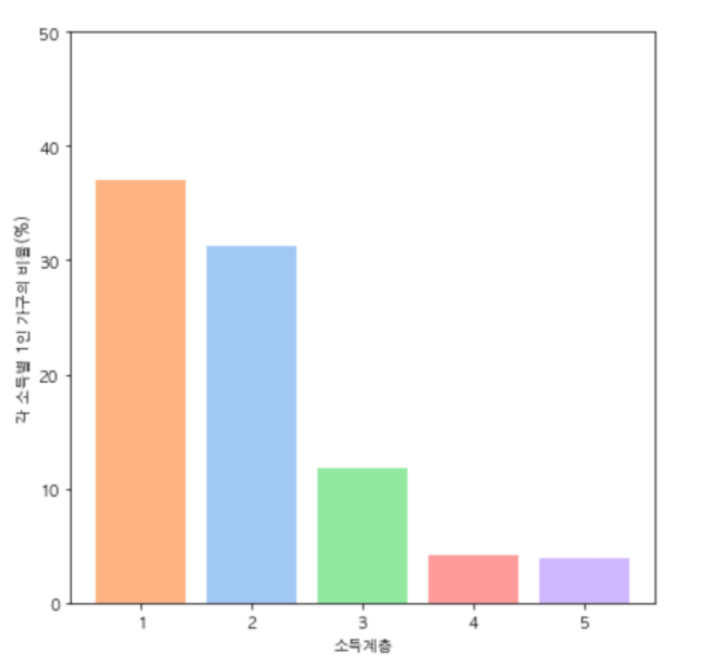
연령성별 1인 가구

By Age and gender



소득별 1인 가구

By Income



1인 가구
요일별

ANALYSIS BY ONE-PERSON HOUSEHOLDS

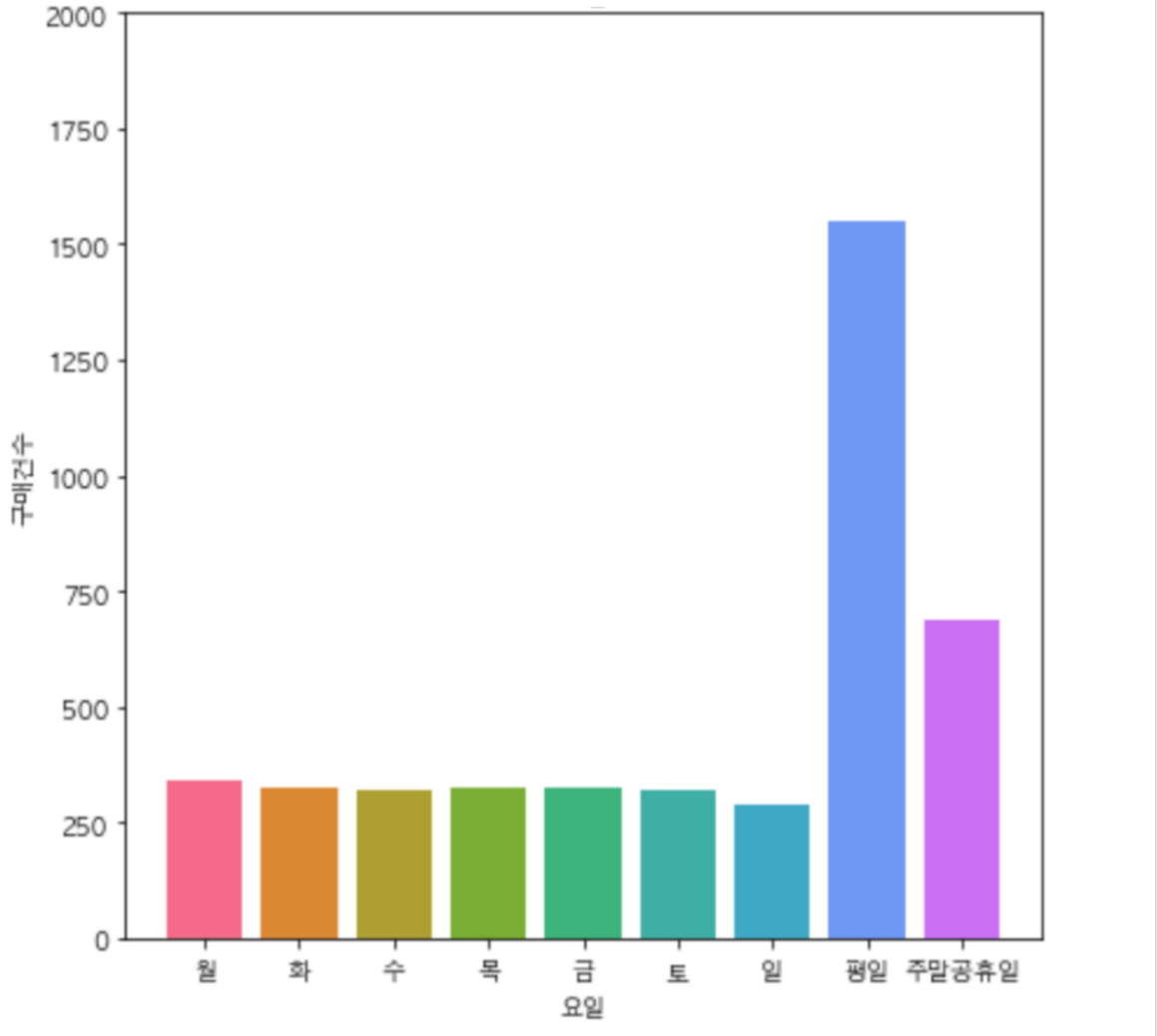
* 모든 요일에서 비슷한 구매정도를 보였고, 구매에 대하여 요일의 영향은 크게 없으나
평일에 근소한 차이로 구매건이 높았습니다

* 요일은 소비와 별 영향이 없다고 판단

```
일인가구요일별 = 일인가구.iloc[:,8:17]
일인가구요일별_평균구매건수 = 일인가구요일별.aggregate(np.mean)
Days = ['월','화','수','목','금','토','일','평일','주말공휴일']

colors = sb.color_palette('husl',10)
f = plt.figure(figsize=(7,7))
plt.bar(Days,일인가구요일별_평균구매건수,color=colors)
plt.title('1인 가구 요일별 _ 평균 구매건수')
plt.xlabel('요일')
plt.ylabel('구매건수')
plt.ylim(0,2000)
plt.show()
```

1인 가구의 요일별_ 구매건수 추이



결론 / 해결방안

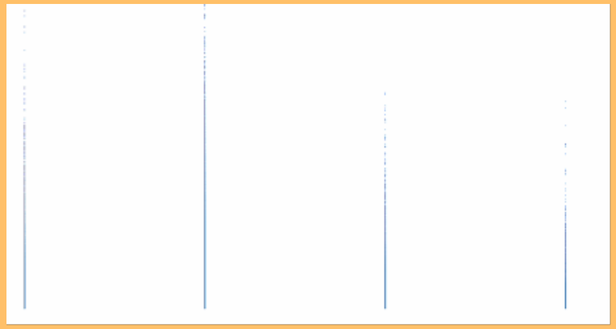
아쉬운 점
보완점

COMPLEMENT

추정소득과 온라인쇼핑간의 상관관계를 파악하기위해

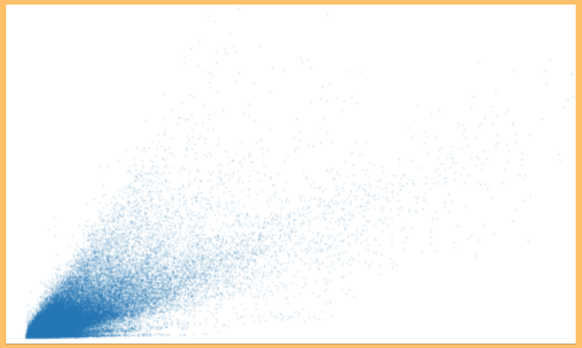
MinmaxScaler/RobustScaler
정규화를 진행했으나 상관관계를 보이지 않아 새로운 관점이나 발견을 못한 것이 아쉬움

추정소득_온라인쇼핑 산점도



추정소득과 온라인 쇼핑의 상관관계를 파악하기 위해 산점도를 생성했으나 연관 없다고 판단

마트할인점_온라인쇼핑 산점도



마트할인점_온라인쇼핑의 상관관계를 파악하기 위해 산점도를 생성, 약한 양의 상관관계를 보이지만 값이 증가하며 분산되어 활용안함

결론 및 해결방안

CONCLUSION AND SOLUTION

지역 음식 기부 플랫폼 구축

가정에서 나온 잔여 식품을 생활이 어려운 계층에 기부하고
기부자에게 소정의 지역크레딧을 지급하여 지역경제 활성화

푸드뱅크
Food Bank

기존에 시행되고 있는
식품 무상제공 서비스를 적극 홍보 및 활용
(www.foodbank1377.org)

배달 공유
Delivery Sharing System

1인 가구특성 상 잔여식품이 많이 나오는
문제를 해결하기 위해
동일 음식을 이웃과 함께 주문배달하는 시스템 계획

음식나눔 지역 바자회
Food Sharing Bazaar

코로나19 종식 후
음식 기부자와 수혜자를 위한
지역 바자회제도 정기적 시행

광주광역시

감사합니다

참고 Data 및 논문

1.통계청(온라인쇼핑동향조사).온라인쇼핑 동향.(2021.12.3)/이민경,김성규
2.통계청.시도별_1인가구data
3.환경부.음식물핸드북.(2013. 4)/가현기획
4.충남대학교 농업과학연구소.가정내 음식물 쓰레기 감량을 위한 소비자 특성별 행위와 요인분석.(2012. 6)/한재환, 황윤재

프로젝트 진행/발표

2기 수료생 홍인영