데이터 수집 미니 프로젝트

아이스크림

안다윤
이광호 강사님

빅데이터 분석 기반 AI 알고리즘 개발 과정
2023 06 29 금
메가스터디IT아카데미

아이스크림 : 수집할 정보

수집할 웹사이트 :

CU 와 emart mall 의 특징비교

CU

1. 한 화면에 전체 판매목록이 있지 않고 '더보기' 버튼이 있음
2. 대부분 판매명에 아이스크림 브랜드와 이름이 중간에 ')'로 구분되어 있음 -> 이름과 브랜드명 추출 용이

emart

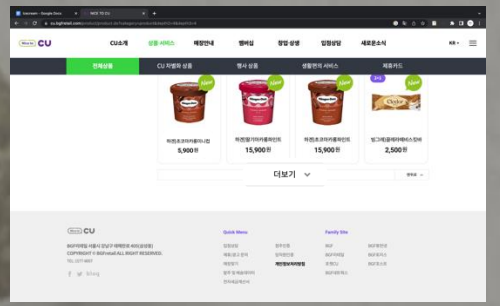
1. 한 화면에 전체 판매목록이 있지 않고 22개의 페이지가 있음
2. 판매명에 아이스크림 이름, 용량, 개수, 브랜드 등이 무작위로 나열되어 있으므로 이름만 추출하기 어려움 -> 판매명 전체를 추출함
3. 브랜드명 따로 표시되어 있는 경우가 있고 판매명에 포함되어 있는 경우가 있음

와 emart mall 의 # 특징비교

CU

1. 한 화면에 전체 판매목록이 있지 않고 '더보기' 버튼이 있음
2. 대부분 판매명에 아이스크림 브랜드와 이름이 중간에 ')'로 구분되어 있음 -> 이름과 브랜드명 추출 용이

https://cu.bgfretail.com/product/product.do?category=product&depth2=4&depth3=4

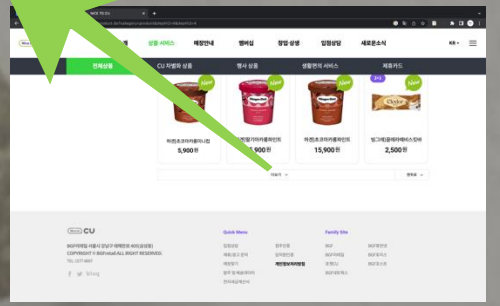


와 emart mall 의 # 특징비교

CU

1. 한 화면에 전체 판매목록이 있지 않고 '더보기' 버튼이 있음
2. 대부분 판매명에 아이스크림 브랜드와 이름이 중간에 ')'로 구분되어 있음 -> 이름과 브랜드명 추출 용이

더보기



https://cu.bgfretail.com/product/product.do?category=product&depth2=4&depth3=4

와

의

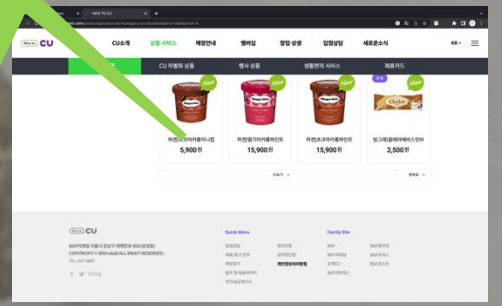
특징비교

CU

1. 한 화면에 전체 판매목록이 있지 않고 '더보기' 버튼이 있음
2. 대부분 판매명에 아이스크림 브랜드와 이름이 중간에 ')'로 구분되어 있음 -> 이름과 브랜드명 추출 용이

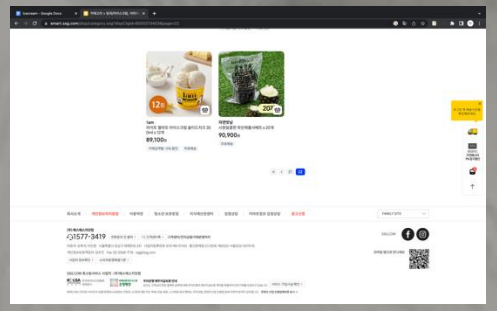
하겐)초코마카롱미니컵

5,900원



https://cu.bgfretail.com/product/product.do?category=product&depth2=4&depth3=4

CU 와 emart mall 의 특징비교

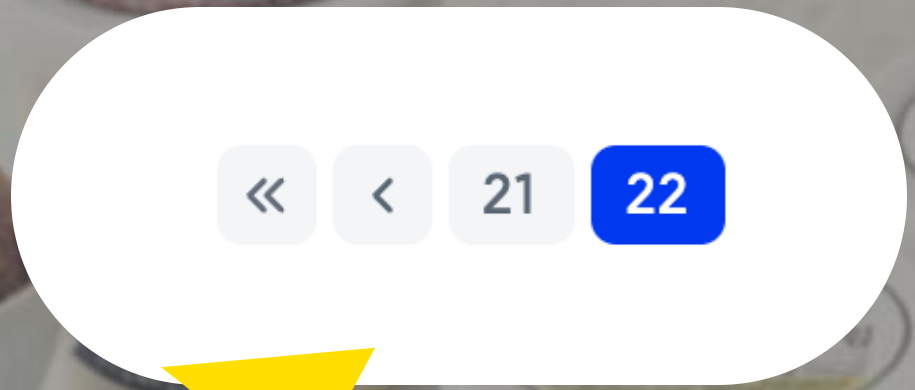


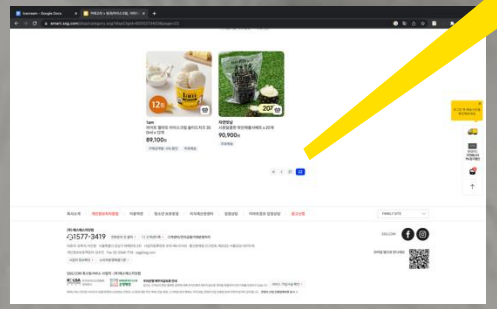
https://emart.ssg.com/disp/category.ssg?dispCtgId=6000213403

emart

1. 한 화면에 전체 판매목록이 있지 않고 22개의 페이지가 있음
2. 판매명에 아이스크림 이름, 용량, 개수, 브랜드 등이 무작위로 나열되어 있으므로 이름만 추출하기 어려움 -> 판매명 전체를 추출함
3. 브랜드명 따로 표시되어 있는 경우가 있고 판매명에 포함되어 있는 경우가 있음

CU 와 emart mall 의 특징비교





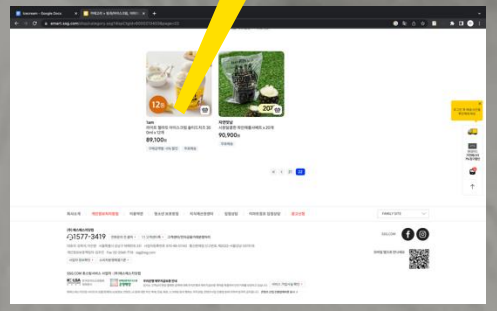
https://emart.ssg.com/disp/category.ssg?dispCtgId=6000213403

emart

1. 한 화면에 전체 판매목록이 있지 않고 22개의 페이지가 있음
2. 판매명에 아이스크림 이름, 용량, 개수, 브랜드 등이 무작위로 나열되어 있으므로 이름만 추출하기 어려움 -> 판매명 전체를 추출함
3. 브랜드명 따로 표시되어 있는 경우가 있고 판매명에 포함되어 있는 경우가 있음

CU 와 emart mall 의 특징비교

1am
라이트 젤라또 아이스크림 솔티드치즈 35
0ml x 12개
,100원

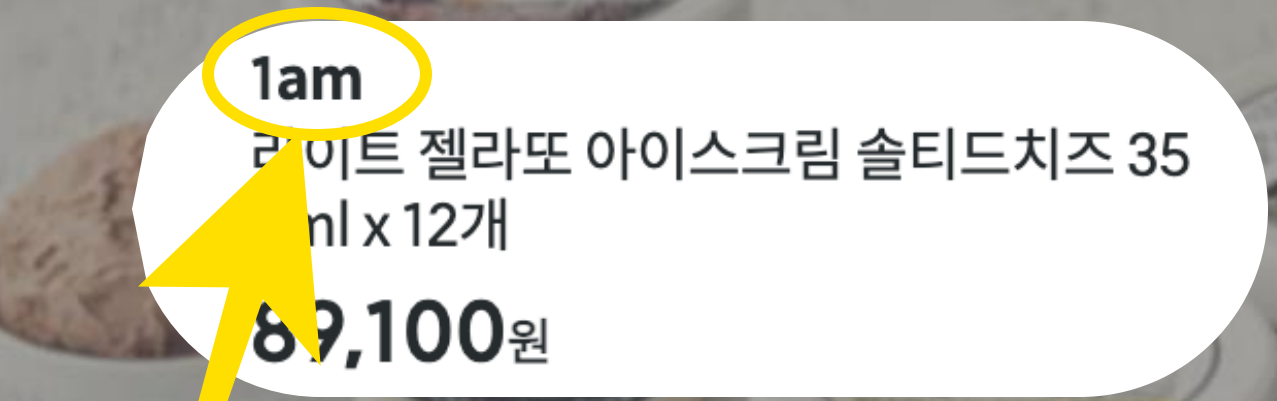


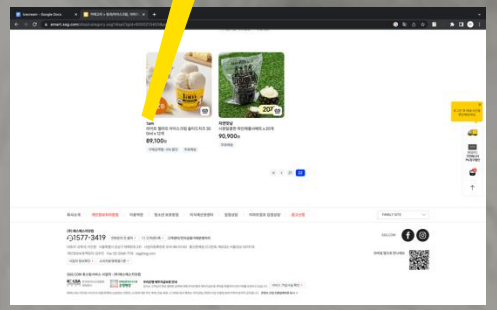
https://emart.ssg.com/disp/category.ssg?dispCtgId=6000213403

emart

1. 한 화면에 전체 판매목록이 있지 않고 22개의 페이지가 있음
2. 판매명에 아이스크림 이름, 용량, 개수, 브랜드 등이 무작위로 나열되어 있으므로 이름만 추출하기 어려움 -> 판매명 전체를 추출함
3. 브랜드명 따로 표시되어 있는 경우가 있고 판매명에 포함되어 있는 경우가 있음

CU 와 emart mall 의 특징비교





https://emart.ssg.com/disp/category.ssg?dispCtgId=6000213403

emart

1. 한 화면에 전체 판매목록이 있지 않고 **22개의 페이지**가 있음
2. 판매명에 아이스크림 **이름, 용량, 개수, 브랜드 등이 무작위로 나열**되어 있으므로 이름만 추출하기 어려움 -> 판매명 전체를 추출함
3. 브랜드명 **따로 표시되어 있는 경우**가 있고 **판매명에 포함되어 있는 경우**가 있음

Selenium: 적용기술

Open Source 자동화 도구 중 하나인 Selenium을 사용하여

Chrome 브라우저를 자동으로 제어하는 기술을 통해

판매하는 아이스크림에 대한 정보가 있는

CU와 emart 웹 사이트에 접속하여

각 웹사이트에서 원하고자 하는 정보를 수집하였다

수집 프로그램 구현

for

Step 1:

필요한 Package 참조

```
import chromedriver_autoinstaller
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.common.by import By
import time

from bs4 import BeautifulSoup
from pandas import DataFrame
```

Step 2:

Chrome Browser 가동

```
chromedriver_autoinstaller.install()
driver = webdriver.Chrome()

# 크롬브라우저가 준비될 때 까지 최대 5초씩 대기
driver.implicitly_wait(5)
```

Step 3:

CU 아이스크림 페이지로 이동

```
driver.get("https://cu.bgfretail.com/product/product.do?category=product&depth2=4&depth3=4")
# 1초간 대기
time.sleep(1)
```

Step 4:

모든 아이스크림을 조회하기 위해 '더보기' 버튼 누르고 5초간 대기 (총 6번)

```
a = 0
for a in range(6):
    if a < 6:
        more_button = WebDriverWait(driver, 3).until(
            lambda x: x.find_element(By.CSS_SELECTOR, ".prodListBtn-w")
        )
        more_button.click()
        # 5초간 대기
        time.sleep(5)
        a += 1
```

Step 5:

내용 수집

Step 5-1) 모든 정보 수집할 리스트 생성 및 아이스크림의 정보가 표시되는 html & css 추출

```
icecream_data = []

soup = BeautifulSoup(driver.page_source)
icecreamList = soup.select(".prod_wrap")
```

Step 5:

내용 수집

Step 5-2) 추출된 아이스크림 목록 수 만큼 반복하며 이름, 브랜드명, 가격 가져오기

```
for i in icecreamList:
    # 이름
    name = i.select(".name")
    if name:
        nameValue = name[0].text.strip()
        if ')' in nameValue:
            indexBeforeName = nameValue.index(')')
            nameValue = nameValue[indexBeforeName + 1:]
        else:
            nameValue = nameValue
    else:
        nameValue = None

    # 브랜드명
    brand = i.select(".name")
    if brand:
        brandValue = brand[0].text.strip()
        if ')' in brandValue:
            indexBeforeName = brandValue.index(')')
            brandValue = brandValue[:indexBeforeName]
        else:
            brandValue = None
    else:
        brandValue = None

    # 가격
    price = i.select(".price")
    if price:
        priceValue = price[0].text.strip().replace('원', '').replace(',', '')
    else:
        priceValue = None
```

```
    # 결과 병합
    resultDic = {'이름': nameValue,
                 '브랜드': brandValue,
                 '가격 (원)': priceValue}
    icecream_data.append(resultDic)
```

Step 6:

Excel 파일로 수집한 내용 저장

```
df = DataFrame(icecream_data)
df.to_excel("icecream_list_cu.xlsx")
```

수집 결과

icecream_list_cu.xlsx

	A	B	C	D
1		이름	브랜드	가격(원)
2	0	그루비초콜릿바	IEK	3900
3	1	그루비카라멜아몬드바	IEK	3900
4	2	팥모나카	라라스윗	3300
5	3	폴바셋카라멜파인트	엠즈	13900
6	4	모구모구리치요거바	제스	2200
7	5	요맘때딸기홈	빙그레	9000
8	6	청도홍시빙수	라벨리	3500
9	7	메로나망고	빙그레	1500
10	8	폴라포매실	해태	1800
11	9	레모나아이스	해태	2200
12	10	서울앵무새모나카	삼우	2500
13	11	초콜릿빵샌드	라라스윗	3300
14	12	생우유빵샌드	라라스윗	3300
15	13	제로밀크소프트콘	롯데	2500
16	14	제로밀크모나카	롯데	2500
17	15	1000콘초코	삼우	1000
18	16	1000콘바닐라	삼우	1000
19	17	나바나나	서주	400
20	18	우유마루홈	해태	7000
21	19	레어케이크	서주	2500
22	20	따옴바포도	빙그레	2000
23	21	따옴바파인애플	빙그레	2000
24	22	빅치즈마루홈	해태	9000
25	23	약콩크런치초코바	밥스누	2500
26	24	빅구슬밀크카라멜	동학	1800
27	25	이정도는약과지바	라벨리	1800
28	26	쌍쌍바메로나	해태	1500
29	27	소금바닐라샌드	서주	2500
30	28	레모나바	해태	1800
31	29	비비빅바밤바	빙그레	1500
32	30	밀탑팥엔밀크바	삼우	1500

...

	A	B	C	D
233	231	딸기파인트	하겐	15900
234	232	월드콘	롯데	2200
235	233	스크류바	롯데	1200
236	234	더위사냥액티브	빙그레	1800
237	235	더블비얀코	롯데	2200
238	236	설레임밀크쉐이크	롯데	2200
239	237	쿠앤크바	빙그레	1500
240	238	빅구슬복숭아키위	동학	1800
241	239	구구크러스터홈	롯데	6000
242	240	빵빠레초코	롯데	2200
243	241	HEYROO우유팥빙수컵		2800
244	242	찰떡아이스	롯데	2200
245	243	돼지바	롯데푸드	1200
246	244	옥동자	롯데	1200
247	245	와샤베트	롯데	2200
248	246	HEYROO파르페플러리		2700
249	247	HEYROO파르페딸기		2700
250	248	쌍쌍바	해태	1500
251	249	비비빅	빙그레	1500
252	250	수박바	롯데	1200
253	251	구구콘	롯데	2200
254	252	죠스바	롯데	1200
255	253	빠삐코	롯데푸드	1500
256	254	메로나	빙그레	1500
257	255	부라보콘바닐라	해태	2200
258	256	본가찰옥수수	롯데	2200
259	257	초콜릿파인트	나뚜루	14900
260	258	딸기파인트	나뚜루	14900
261	259	녹차파인트	나뚜루	14900
262	260	바닐라아몬드바	나뚜루	4800
263	261	녹차미니컵	나뚜루	4800
264				

Next.

수집 프로그램 구현

for emart mall

Step 1:

필요한 Package 참조

```
import chromedriver_autoinstaller
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
import time

from bs4 import BeautifulSoup
from pandas import DataFrame
```

emart mall

Step 2:

Chrome Browser 가동

```
chromedriver_autoinstaller.install()
driver = webdriver.Chrome()

# 크롬브라우저가 준비될 때 까지 최대 5초씩 대기
driver.implicitly_wait(5)
```

emart mall

수집 프로그램 구현

Step 3:

emart 페이지 별 아이스크림에 접근하기 위한 변수값 설정 (총 22 페이지)

```
params = list(range(1, 23))
url_emart = "https://emart.ssg.com/disp/category.ssg?dispCtgId=6000213403&page={0}"
```

emart mall

Step 4:

내용 수집

Step 4-1) 모든 정보 수집할 리스트 및 판매명에 포함되어 있는 브랜드명이 담긴 리스트 생성

```
icecream_data = []
brandName = ['빙그레', '해태', 'No Brand', '노브랜드', '롯데', '피코크', 'SSG푸드마켓', '매일유업',
             '나뚜루', '나뚜르', '하겐다즈', '설빙', '바른씨', '아이러브아이스크림', '제이큐',
             '글로벌푸드', '우리밀', '유니레버', '골든벨', '동서', '서울우유', '허쉬', '화과방',
             '매그넘', '상상앤드', '고디바','우리가스토리', '자연맛남', '앤드류허쉬', '1am',
             '비비수산', '인정식탁', '널담', '몽테이블', '부르스터스', '네추럴킹덤', '오설록',
             '델몬트', '아이스올리', '라벨리', '라라스윗', '할로', '끌레도르', '벤엔제리스', '상하목장',
             '삼립', '다이센', '네니아', '킷캣', '미닛메이드', '엠즈씨드', '사몬타나', '헤일로탑',
             '바라톨리노', '마이노멀', '테이트', '하이디', '맥키스','매키스']
```

emart mall

Step 4:

내용 수집

Step 4-2) 페이지 별 아이스크림의 정보가 표시되는 html & css 추출

```
# 페이지별 반복
for p in params:
    url = url_emart.format(p)
    driver.get(url)
    time.sleep(1)

    soup = BeautifulSoup(driver.page_source)
    icecreamList = soup.select(".mnemitem_grid_item")
```

emart mall

Step 4:

내용 수집

Step 4-3) 앞의 for문 안에서 추출된 아이스크림 목록 수 만큼 반복하여 판매명, 브랜드명, 가격 가져오기

```
# 추출된 아이스크림 목록 수 만큼 반복
for i in icecreamList:
    # 판매명
    name = i.select(".mnemitem_goods_tit")
    if name:
        nameValue = name[0].text.strip()
    else:
        nameValue = None

    # 브랜드명
    brand = i.select(".mnemitem_goods_brand")
    if brand:
        brandValue = brand[0].text.strip()
    else:
        for b in brandName:
            if b in nameValue:
                brandValue = b
                break
            else:
                brandValue = None

    # 가격
    price = i.select(".new_price > .ssg_price")
    if price:
        priceValue = price[0].text.strip().replace(',', '')
    else:
        priceValue = None
```

```
        # 결과 병합
        resultDic = {'판매명': nameValue,
                     '브랜드': brandValue,
                     '가격(원)': priceValue}
        icecream_data.append(resultDic)
```

emart mall

Step 5:

Excel 파일로 수집한 내용 저장

```
df = DataFrame(icecream_data)
df.to_excel("icecream_list_emart.xlsx")
```

수집 결과

emart mall

icecream_list_emart.xlsx

(편의상 2 페이지만 수집한 결과)

	A	B	C	D
1		판매명	브랜드	가격(원)
2	0	[롯데]일품 팥빙수	롯데	1600
3	1	[빙그레] 메로나 75ml*8	빙그레	4800
4	2	롯데 월드콘XQ멀티160ml*5	롯데	6000
5	3	[빙그레] 비비빅 70ml*8	빙그레	4800
6	4	빙그레 더위사냥 140ml*5	빙그레	4000
7	5	[나뚜루] 녹차 아이스크림 파인트 474ml	나뚜루	7450
8	6	(Q)빙그레 싸만코150ml*5	빙그레	6000
9	7	해태 폴라포포도 120ml*6	해태	4368
10	8	롯데 빠삐코130ml*6	롯데칠성	4800
11	9	롯데 구구콘 아이스크림 160ml*5	롯데칠성	6000
12	10	(Q)빙그레 빵또아 180ml*5	빙그레	6000
13	11	초코칩아이스660ml	노브랜드	4980
14	12	(Q)롯데 스크류바 75ml*6	롯데	3600
15	13	빙그레 쿠앤크750ml	빙그레	5000
16	14	향긋한바닐라향아이스5L	노브랜드	13480
17	15	쿠키칩아이스크림660ml	노브랜드	4980
18	16	(Q)빙그레 투게더 900ml	빙그레	5600
19	17	프로틴 아이스크림 초코 474ml	피코크	7980
20	18	[나뚜루] 바닐라 아이스크림 파인트 474ml	나뚜루	7450
21	19	밀크아이스크림660ml	노브랜드	4980
22	20	[롯데] 티코 다크 초코 510ml	롯데	6000
23	21	[나뚜루] 초코 아이스크림 파인트 474ml	나뚜루	7450
24	22	[롯데] 찰옥수수 700ml(140ml*5개입)	롯데	6000
25	23	롯데 셀렉션 아이스크림 500ml	롯데	6000
26	24	[빙그레] 더 엑설런트 오리지널 800ml	빙그레	6400
27	25	나뚜루 피페리타 민트초코 파인트	나뚜루	7450
28	26	롯데 설레임밀크160ml*5	롯데	7500
29	27	나뚜루 눈꽃 위드 레드빈 파인트 474ml	나뚜루	7450
30	28	롯데 구구 100ml*5입	롯데칠성	6000
31	29	[끌레도르]레드카펫 치즈케익 파인트474ml	끌레도르	6250
32	30	욜로우 녹차 474ml	피코크	5980
33	31	스트로베리치즈케익 아이스크림 파인트 474ml	나뚜루	7450
34	32	롯데 구구크러스터660ml	롯데칠성	6000
35	33	해태 호두마루 660ml	해태	4500
36	34	롯데 셀렉션 싱글 초코 500ml	롯데	6000

...

	A	B	C	D
124	122	[라라스윗] 말차 474ml	SSG푸드마켓	7900
125	123	[라라스윗] 딸기 474ml	SSG푸드마켓	7900
126	124	빅구슬멀티팩37ml*8		7840
127	125	라라스윗 팥 모나카 아이스크림 140ml * 4	라라스윗	9400
128	126	나뚜루 비건 그린티&초코너츠	나뚜루	7450
129	127	[서울우유] 초콜릿우유 아이스크림 474ml	서울우유	9900
130	128	라라스윗 초콜릿 빵샌드 180ml * 4개입	라라스윗	9400
131	129	달달한 디저트/아이스크림 모음전!		4980
132	130	분위기UP시켜주는 디저트/아이스크림		7450
133	131	[하겐다즈] 미니컵멀티팩(초콜렛 퐁당+스트로베리+그린티	하겐다즈	18200
134	132	[빙그레] 끌레도르 제주청정 녹차 90ml	빙그레	2150
135	133	[비요뜨] 요거트 아이스크림 블루베리 474ml		9900
136	134	[라라스윗] 치즈케이크 474ml	SSG푸드마켓	7900
137	135	멀티컵 아이스크림 (100ml*4입)	나뚜루	17280
138	136	[라라스윗] 생우유 474ml	SSG푸드마켓	6320
139	137	[벤앤제리스] 카라멜 수트라 코어 파인트 473ml		14900
140	138	끌레도르 마다가스카르 바닐라아몬드 90ml	끌레도르	2150
141	139	조안나 쿠앤크 900ml		6500
142	140	상하목장 얼려먹는 아이스크림 밀크&초코 85ml X 12입	매일유업	9480
143	141	디저트, 홈파티, 달콤한 테이블		15500
144	142	믿고 구매하는 하겐다즈 모음전	하겐다즈	15500
145	143	부르스터스 아이스크림 (바닐라+오레오) 100ml*4	부르스터스	15600
146	144	아이스크림 파인트 2+1 / 총3개	하겐다즈	31000
147	145	나뚜루 바 컬렉션	나뚜루	17280
148	146	[하겐다즈] 캐러멜 비스킷 앤 크림 아이스크림 473ml	하겐다즈	15500
149	147	빙그레 투게더 바닐라맛 900ml 오리지널 아이스크림	바른씨	4900
150	148	아이스크림 멀티바 바닐라카라멜아몬드(3개입)	하겐다즈	9900
151	149	아이스크림 멀티바 스트로베리앤크림(3개입)	하겐다즈	9900
152	150	콘,파우치 아이스크림 12개 골라담기 월드콘,설레임,구구콘,돼지콘 9종		16500
153	151	따옴바 3종 혼합 30개 (납작복숭아6 + 딸기6 + 패션프루트(	빙그레	26900
154	152	뽕따 소다맛 30개 /쭈쭈바	빙그레	23500
155	153	상하목장 얼려먹는아이스크림 밀크 85ml 24입	매일유업	22900
156	154	비비빅 오리지널(팥) 40개	빙그레	24500
157	155	설레임 밀크쉐이크160ml * 24개		27900
158	156	상하목장 얼려먹는 아이스크림 초코 85ml 24입	매일유업	22900
159	157	해태 탱크보이 배맛 30개 /쭈쭈바/아이스크림/간식	빙그레	24900
160	158	아이스크림 바 40개 모음(쿠앤크 10개+엔초 10개+메로나	빙그레	27000
161	159	상하목장 얼려먹는아이스크림 망고 85ml 24입	매일유업	22900
162				

감사합니다

데이터 수집: CU & emart mall 아이스크림

안다윤
이광호 강사님
빅데이터 분석 기반 AI 알고리즘 개발 과정
2023 06 29 금
메가스터디IT아카데미