'Instacart' 재주문을 장려하기 위한

마케팅 인사이트 도출 및 상품 추천시스템 제안

Codestates [AIB] 9기 안나

Contents

주제 선정 이유 및 분석 목적

주제 선정 이유

- **인스타카트 소개**: 미국 온라인 식료품 구매 서비스 인기 앱, 데이터와 AI활용한 지능화 된 개인 서비스를 제공
- **기업 목표**: 많은 비용이 소요되는 신규 고객 확보 외에 **기존 고객의 재주문 장려**하여 **매출을 신장 시키는 것**
- 많은 연구에 따르면 재주문 고객이 기업 매출의 40% 이상 창출[1] 할 수 있고, 고객 유지율을 5% 증가시킬 때 수익이 25~95% 증가[2] 할 수 있습니다.

분석의 목적

재주문을 장려하기 위한 마케팅 인사이트 도출

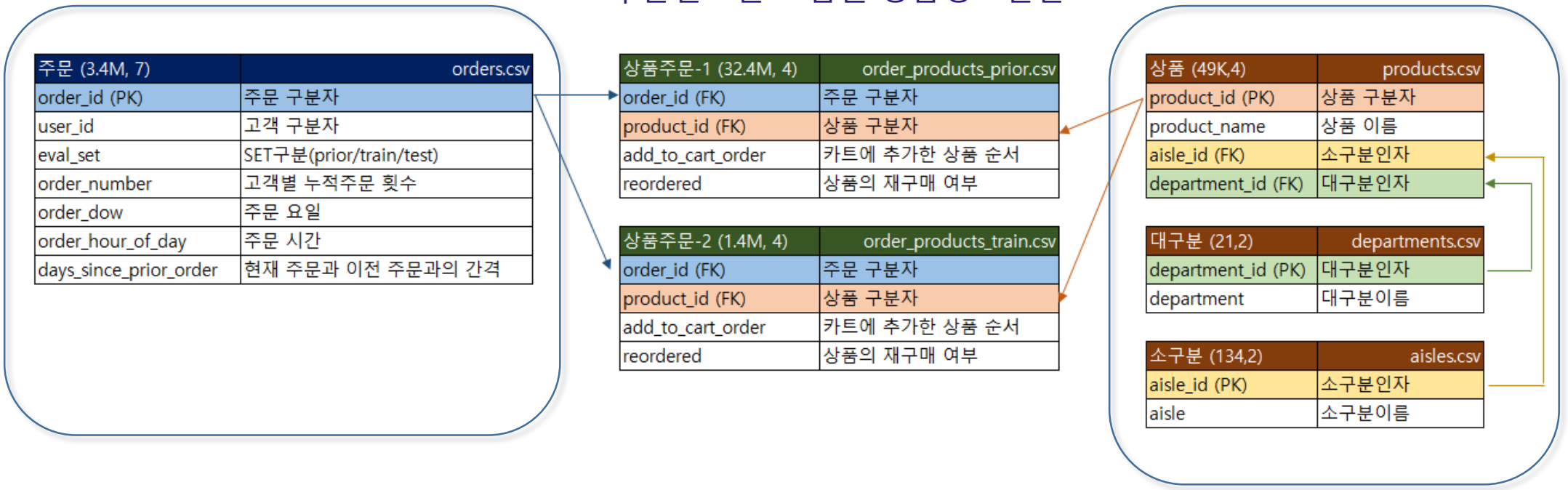
가설 설정: 재주문에는 '패턴'이 있다.

- 주문 시점, 주문 횟수, 주문하는 상품 수, 주문 상품 구성에 특징이 있음을 검증
- 해당 패턴을 파악하고, 재주문을 장려하기 위한 마케팅 인사이트를 도출

1) Adobe Digital Index, 2) Harvard Business Review

데이터 특성

주문번호별 포함된 상품정보연결



- 선정 데이터: **Instacart** Orders 오픈소스 데이터 (**총 20만 명 고객의 340만 건에 대한 주문 정보**)
- 크게 **주문정보, 상품정보**로 구분되며, 하나의 주문시 포함된 상품 목록들이 연결된 데이터 구조 입니다.
- 상품 주문 파일이 prior, train으로 나뉘어 제공되었는데, 이를 합쳐서 분석에 사용하였습니다.

사용 기술 및 구현 내용

- 필요한 데이터를 정의 및 수집하여 이를 분석하는 전체 프로세스를 따라 진행 하였습니다.

데이터

instacart

kaggle

Instacart Orders
오픈소스 데이터
(캐글 업로드 버전)

분석

jupyter

Google Data Studio

데이터 전처리 및 분석
대시보드 구성

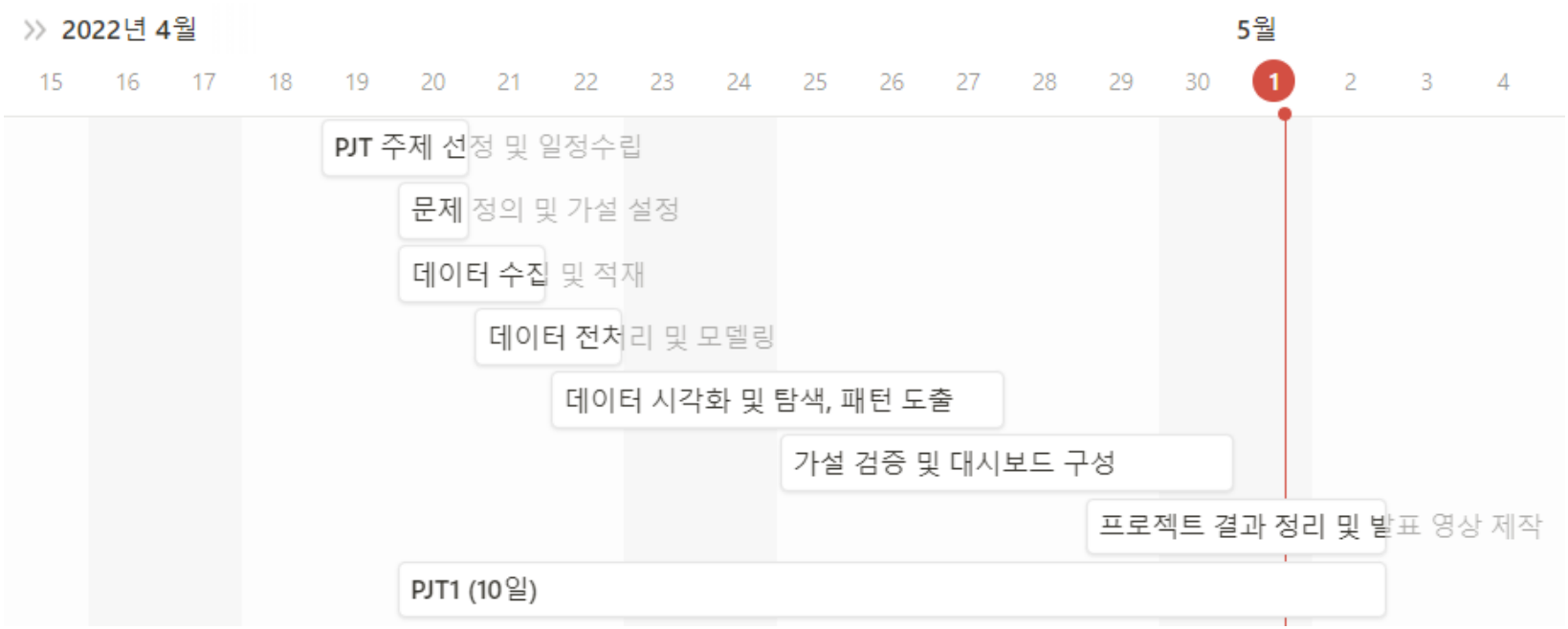
추천 시스템

GENSIM
topic modelling for humans

scikit learn

자연어처리 패키지 (Word2Vec)
KMeans clustering

프로젝트 진행 일정



- 약 10일간 진행 한 프로젝트 이며, 데이터 분석 프로세스에 맞춰 진행하였습니다.
- **데이터**를 **시각화**한 자료를 통해 구매 패턴을 도출하고, **마케팅 인사이트**를 정리하였습니다.
- **연관상품 분석** 및 **상품 추천시스템**을 구성한 내용도 포함됩니다.

1. 인스타카트 사업 현황

- 총 구매자 수: 20만명
- 전체 주문 수: 330만건
- 유저당 평균 주문 수: 16건

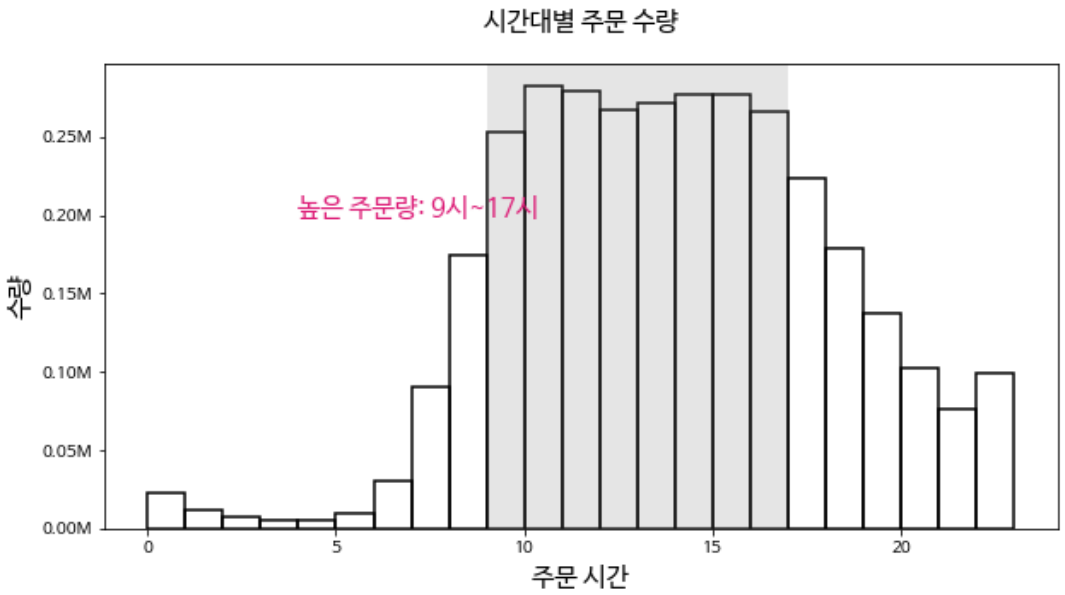
- 총 판매 상품 종류: 4.9만개
- 총 판매 상품 수량: 340만개
- 제품당 평균 주문 수: 680개
- 재주문 상품의 비중: 59%

대시보드: 구글 **데이터스튜디오**로 구성

2. 주문 고객 분석

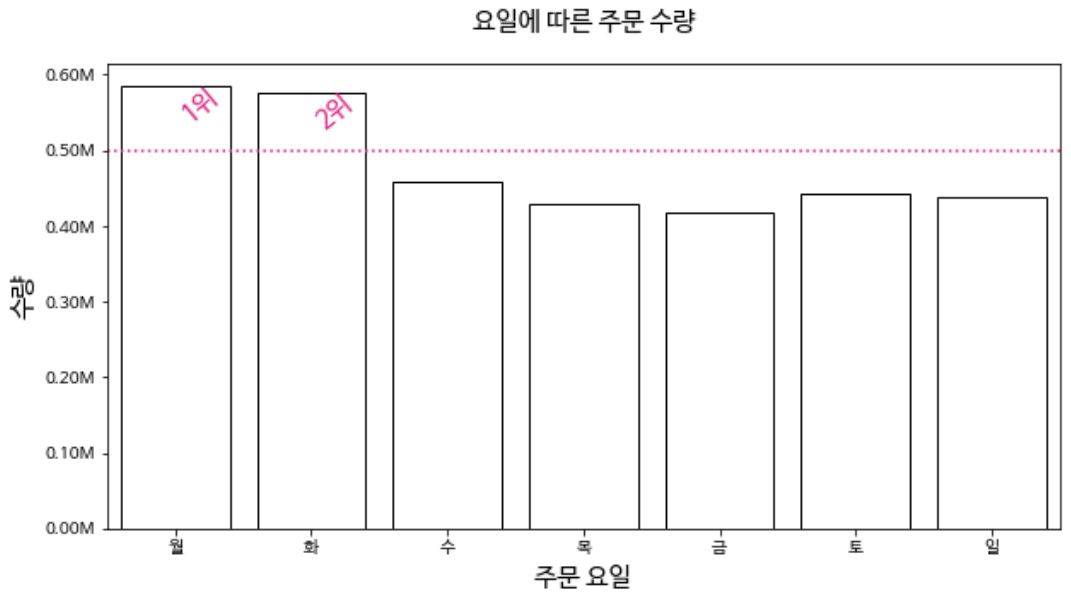
① 가장 주문이 많은 **시간**대는?

- **09:00 - 17:00** 사이 주문량이 일정하게 많습니다.
- (시간당 주문건 20만건 이상)



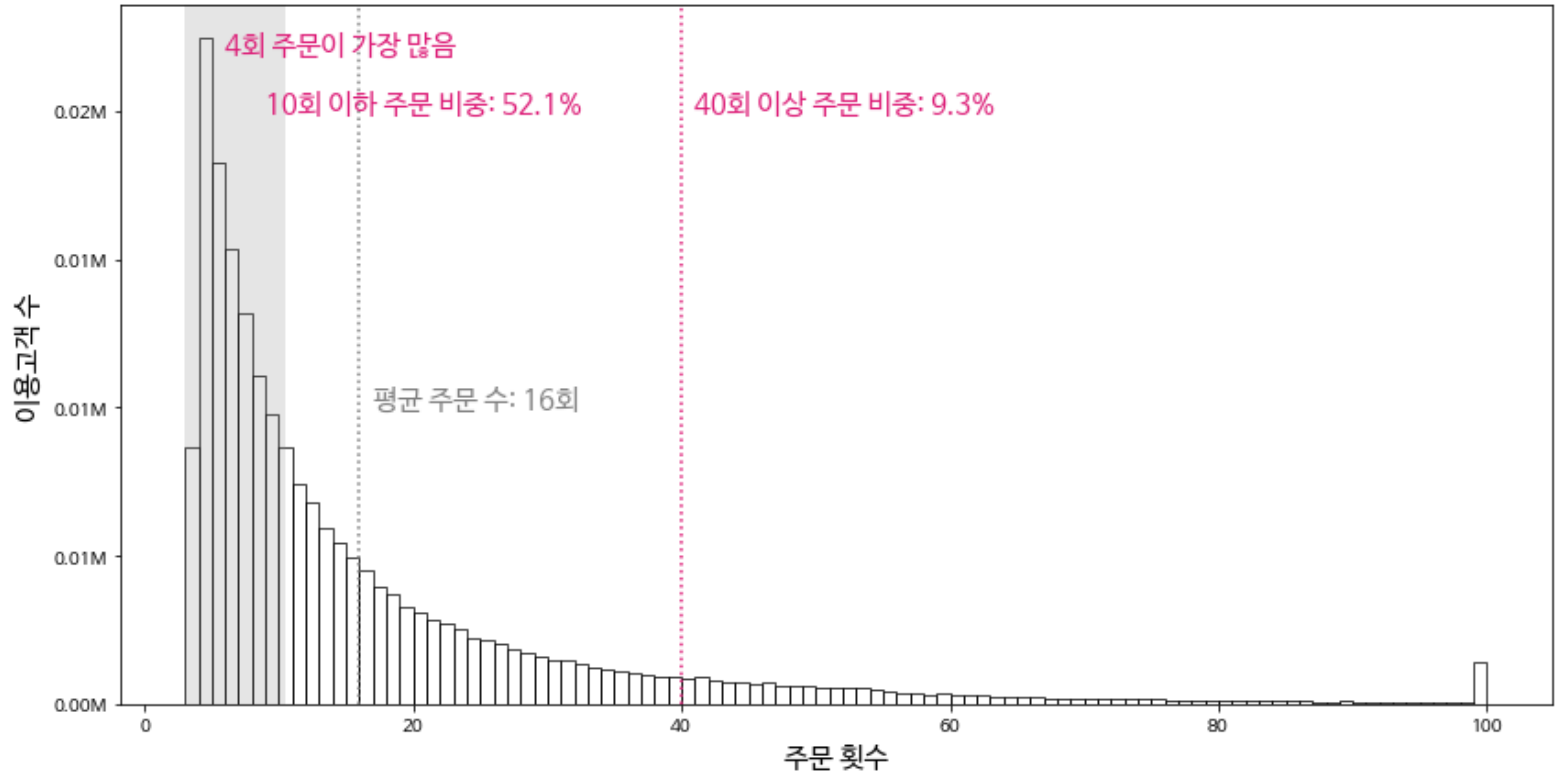
② 가장 주문이 많은 **요일**은?

- 가장 주문이 많은 요일은 **월, 화요일**입니다.



2. 주문 고객 분석

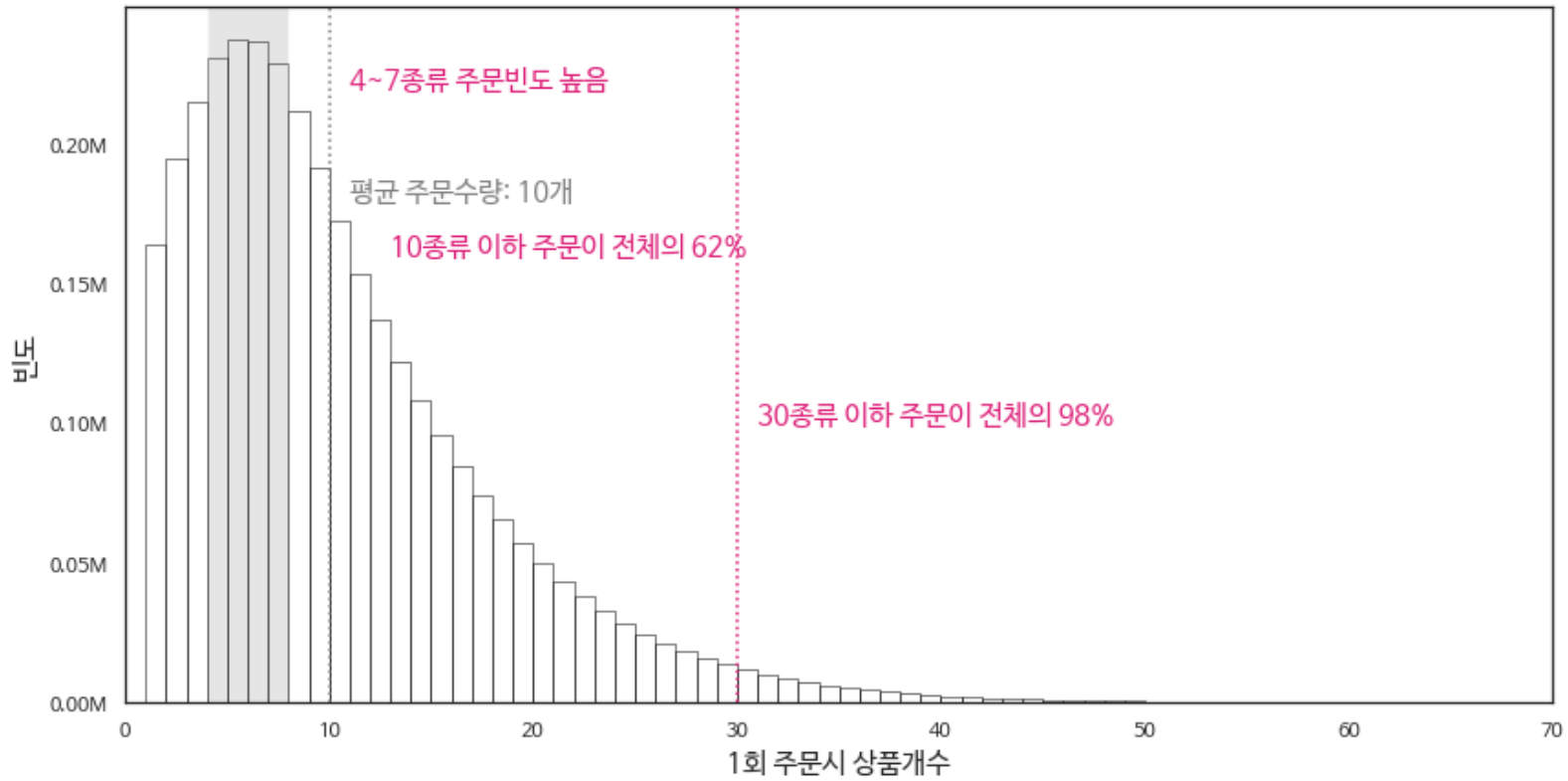
③ 고객별 **누적 주문** 횟수는? • **4번 이용**한 고객이 가장 많고, 5번 부터 감소합니다.



3. 주문 상품 분석

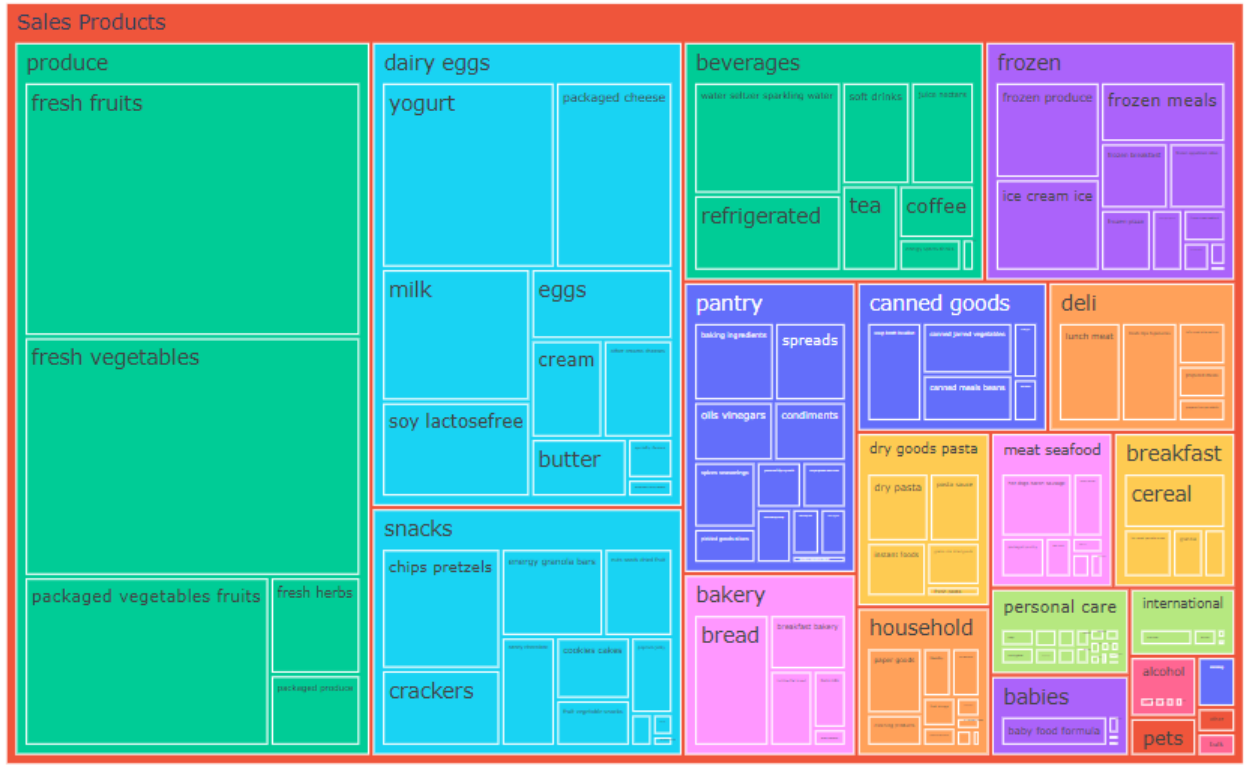
① 한번의 주문에 **주문한 상품의 개수**는?

- **4~7종류** 주문이 전체의 **약 30%**를 차지 합니다.
- **10종류 이하** 주문이 **전체의 62%**, **30종류 이하**는 **전체의 98%**



3. 주문 상품 분석

② 상품 **대분류**, **소분류** 시각화



- Treemap을 통해 상품을 대분류, 소분류를 함께 시각화 하였습니다.
- **박스의 크기가 매출된 수량 차이**를 나타냅니다. 크기가 큰 박스가 비중이 크다고 직관적으로 알 수 있습니다.
- **농산물**(야채와 과일) 비중이 크며, 그다음은 **유제품**(요거트,우유,치즈), **스낵**과 **음료** 순으로 판매수량이 많습니다.

3. 주문 상품 분석

③ 가장 **많이 판매**되는 상품 TOP 10

	product_name	product_cnt	%
0	Banana	491291	1.5
1	Bag of Organic Bananas	394930	1.2
2	Organic Strawberries	275577	0.8
3	Organic Baby Spinach	251705	0.7
4	Organic Hass Avocado	220877	0.7
5	Organic Avocado	184224	0.5
6	Large Lemon	160792	0.5
7	Strawberries	149445	0.4
8	Limes	146660	0.4
9	Organic Whole Milk	142813	0.4

- **바나나**가 1위이며, 상품종류가 많으므로 상위 판매상품도 각 상품이 전체매출에서 차지하는 비중은 1%대로 적습니다.
- 단일 품목이 수량 측면에서는 매출에 대부분을 차지하고 있지는 않습니다.

3. 주문 상품 분석

④ 카트에 **첫 번째**로 담기는 상품 TOP 10

- **물, 우유** 등 **생필품**이 많습니다.

	product_name	total_order	1stadd_rate
35469	Water Mineral	2282	0.47
11884	Sparkling Water, Bottles	1625	0.45
45003	White Multifold Towels	1590	0.44
6728	Cookie Tray	1281	0.40
20939	Organic Low Fat Milk	8806	0.40
40938	Drinking Water	5745	0.40
1728	2% Lactose Free Milk	1854	0.39
26404	XL Pick-A-Size Paper Towel Rolls	1319	0.36
17923	Seltzer Water	1982	0.36
45189	Vodka	5666	0.36

⑤ 카트에 **마지막**에 담기는 상품 TOP10

- **페이퍼타올, 비닐, 플라스틱 컵**처럼 **소모품**이 많았습니다.

	product_name	total_order	last_add_rate
26404	XL Pick-A-Size Paper Towel Rolls	1319	0.44
13947	Wastebasket Liners	1033	0.43
6728	Cookie Tray	1281	0.40
38299	Tall Kitchen Bag With Febreze Odor Shield	1818	0.39
15679	Red Plastic Cups	1680	0.39
43720	Wint-O-Green	1620	0.38
45003	White Multifold Towels	1590	0.38
1956	Fabric Softener Dryer Sheet Outdoor Fresh 160C...	1645	0.36
40938	Drinking Water	5745	0.36
30485	Organic Sweet Cherries	1131	0.34

구매 패턴 심층분석 (가설 검증)

가설

재 주문에는 구매 **패턴**이 있다.

- **어떤 패턴이 있는지?**
  주문 간격, 횟수, 수량, 주문 상품 구성, 주문 요일 및 시간
- **주문 횟수, 구매 상품 수, 재주문 비율**이 높은 상품 등을 확인합니다.
- 다양한 관점에서 어떤 특성들과 연관성을 가지는지 검증해봅니다.

가설 검증 후 활용 방안

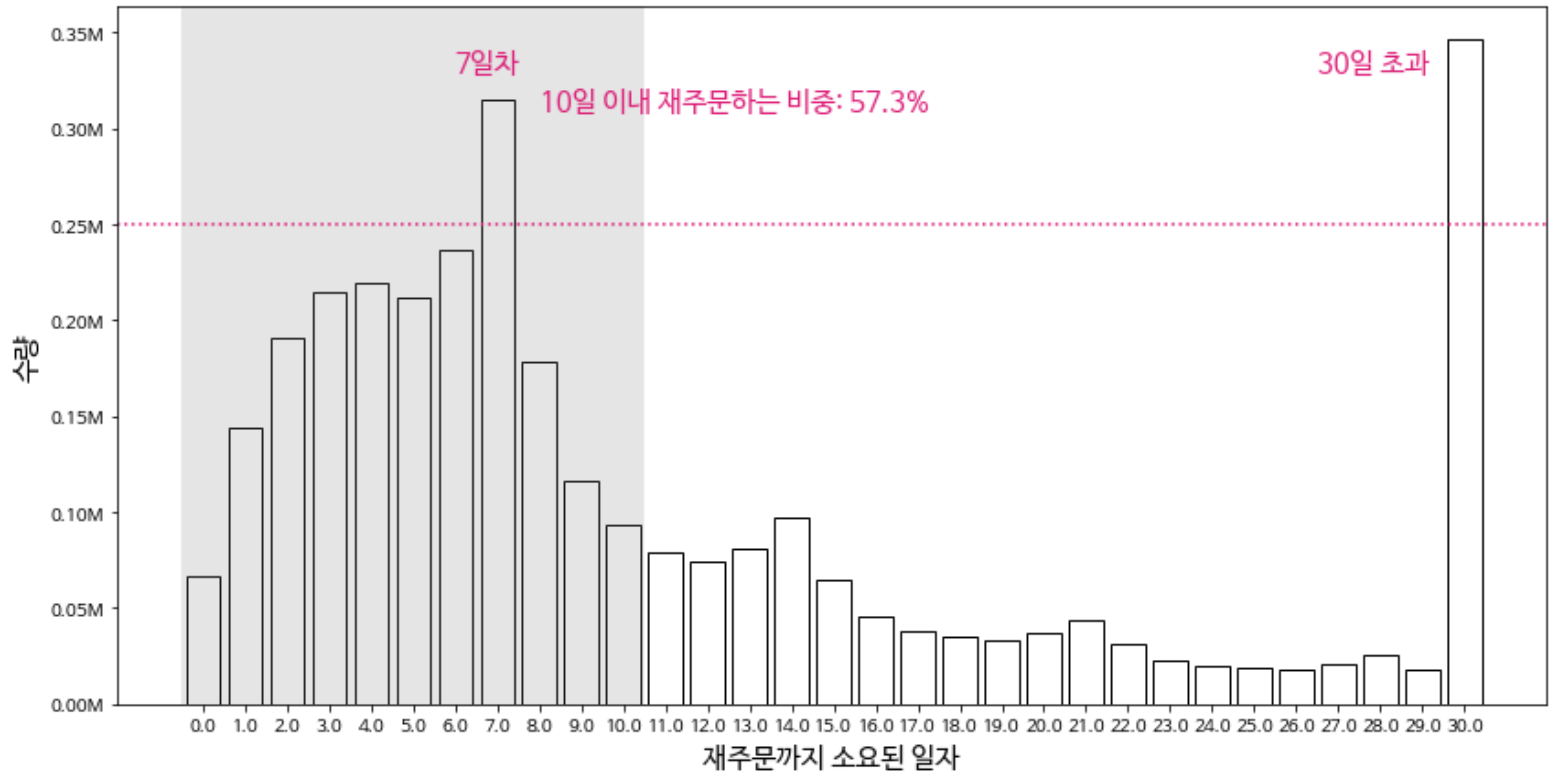
- **분석 결과 수치**를 토대로 고객 특성에 맞게 재주문을 독려하는 방안을 마련합니다.
  예) **주문 간격: 7일** – 메일링,문자로 판촉내용을 알리는 주기를 7일로 합니다.

- 고객의 **타겟팅을 강화**하여 마케팅 활동을 결정 할 수 있습니다. 구매한 상품 목록을 분석하여 **상품추천**을 향상 시킵니다.
  예) 평균적으로 4회 주문이 가장 많고, 5회부터 감소 추세 → **4회 이상**과 **이하** 이용 고객을 분리하여 관리합니다.

1. 주문 간격

① **재주문**까지 어느 정도 **간격**이 있을까?

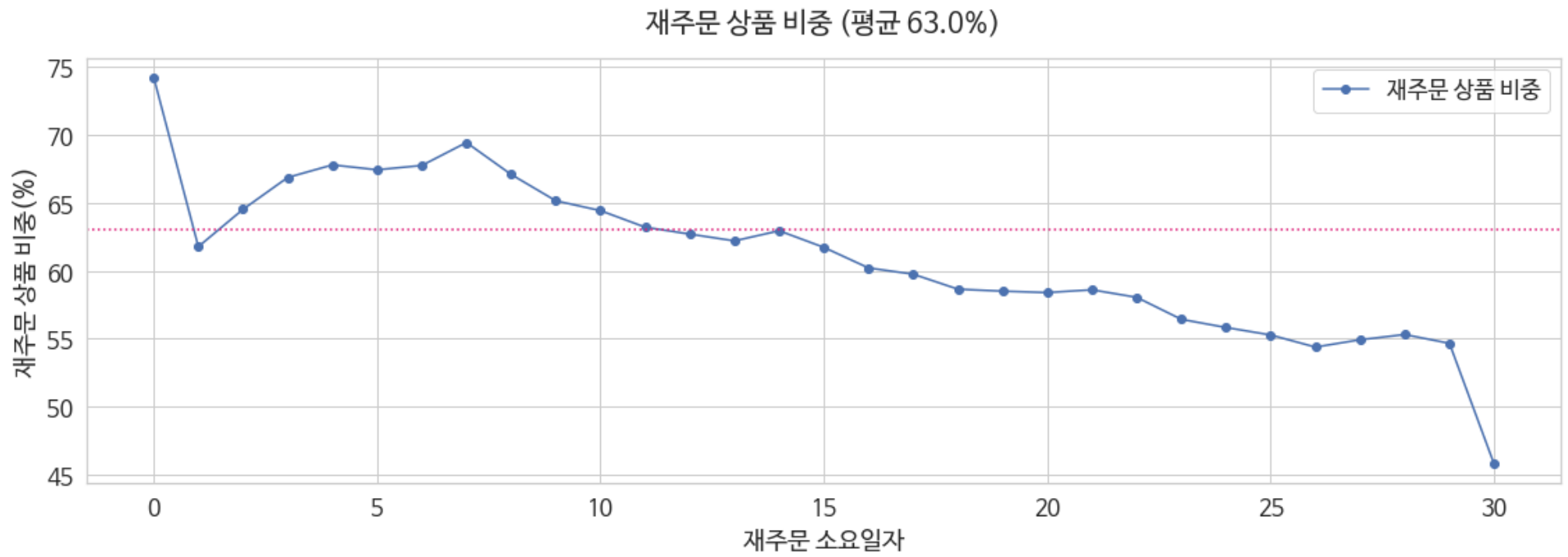
- 보통 10일 이내에, 특히 **일주일** 간격으로 재주문이 많습니다.



1. 주문 간격

② **재주문 기간**이 짧을 수록 **이전 주문**에서 구매한 상품을 **재주문**할까?

- 재주문 상품 비율이 평균대비 높습니다.
- 평균적으로 10일 이내 주문시 재주문 상품비중이 평균대비 높지만, 큰 차이는 아닙니다. (5%미만 차이)



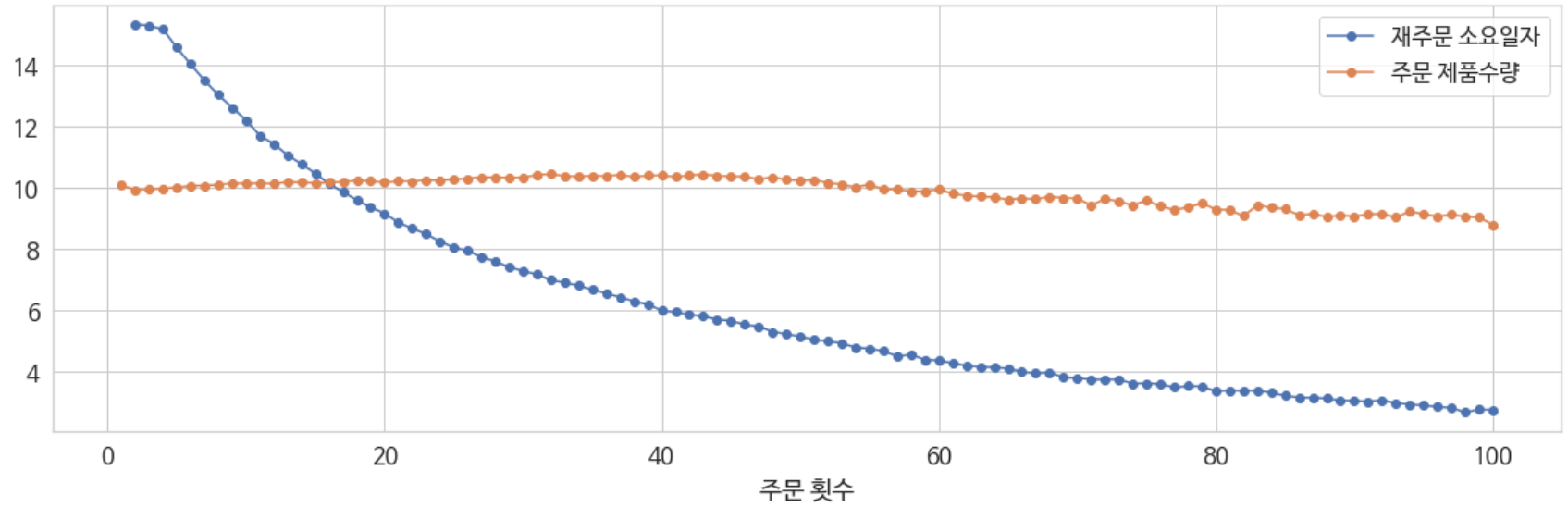
2. 주문 횟수

① **주문 횟수**가 많아 질수록 **재주문 시점**이 빠를까? (그래프: 파랑)

- 주문 횟수가 많다는 것은 고객충성도가 높다고 볼 수 있으므로, 재주문이 빨리 이루어질 것으로 예측 할 수 있습니다.

② **주문 횟수**가 많아 질수록 주문당 **구매하는 상품수**가 많을까? (그래프: 노랑)

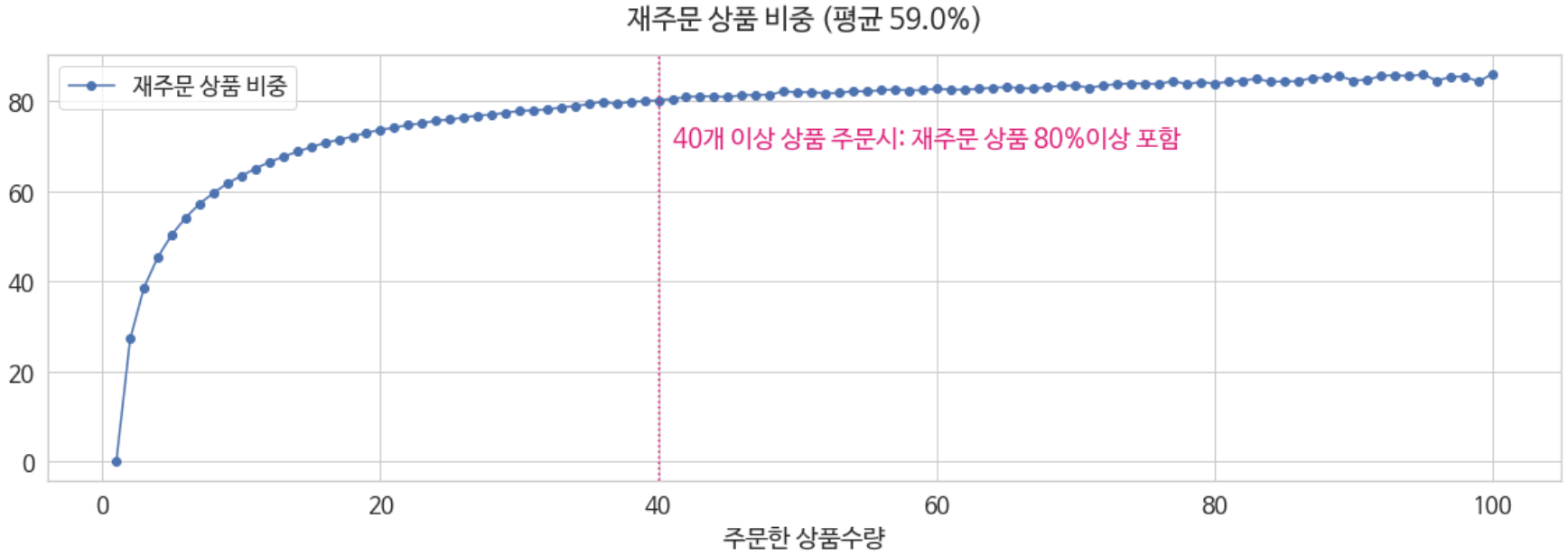
- 상관없이 주문 수량은 유사한 패턴을 보입니다.



3. 주문 수량

① **많은 수량**을 구매한 오더에는 **재주문 상품**이 많을까?

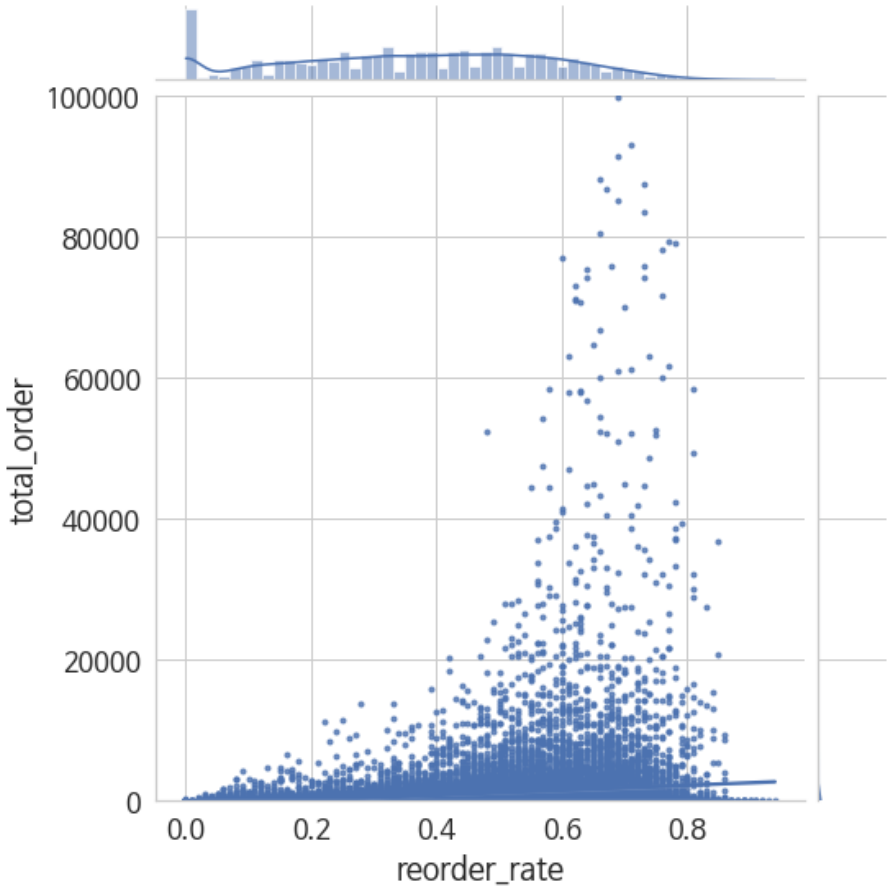
- 그렇습니다. 많은 물품을 주문하는 경우에는 기존 상품을 살 확률일 높습니다.
- **40개 이상 주문**한 경우 **재주문 상품비중이** 80% **이상**입니다.
- 많은 종류의 상품을 구매하는 사람들은 자주 구매하는 상품들이 고정적으로 있는 것으로 추정할 수 있습니다.



3. 주문 수량

② 주문이 많은 **인기상품의 재주문율**이 높은지?

- 당연합니다. 수치로 확인해보아도 양의 상관관계가 나옵니다.
- SpearmanrResult(correlation=0.58)
- spearman 상관계수로 단조성
  (한 변수 값이 커지면 다른 변수도 커지는지) 확인



4. 주문 상품 구성

① **재주문 비율**이 높은 제품은?

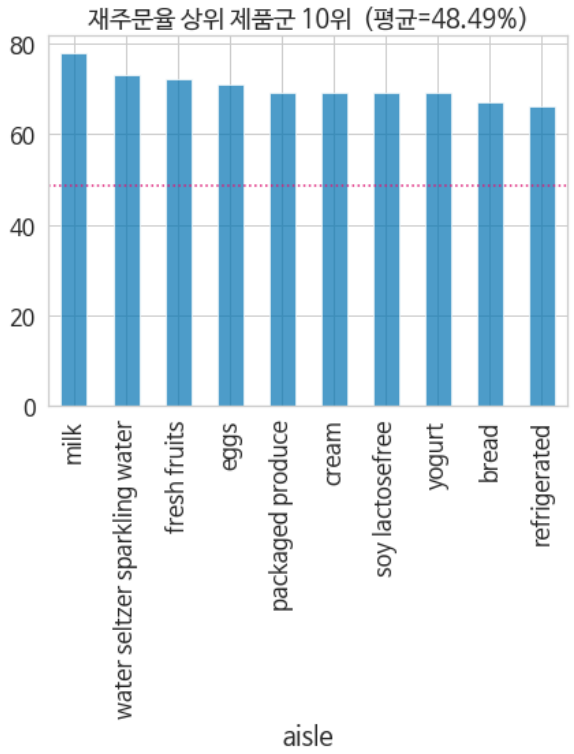
	product_name	total_order	reorder_rate
9291	[1] Half And Half Ultra Pasteurized	2995	0.86
5513	Organic Homogenized Whole Milk	4095	0.86
43393	Organic Lactose Free Whole Milk	8742	0.86
47230	Ultra-Purified Water	1524	0.86
45503	Whole Organic Omega 3 Milk	9410	0.86
38688	Organic Reduced Fat Milk	36869	0.85
24851	Banana	491291	0.85
29446	Milk, Organic, Vitamin D	20770	0.85
34196	Goat Milk	5353	0.85
39179	Organic Lowfat 1% Milk	15352	0.84

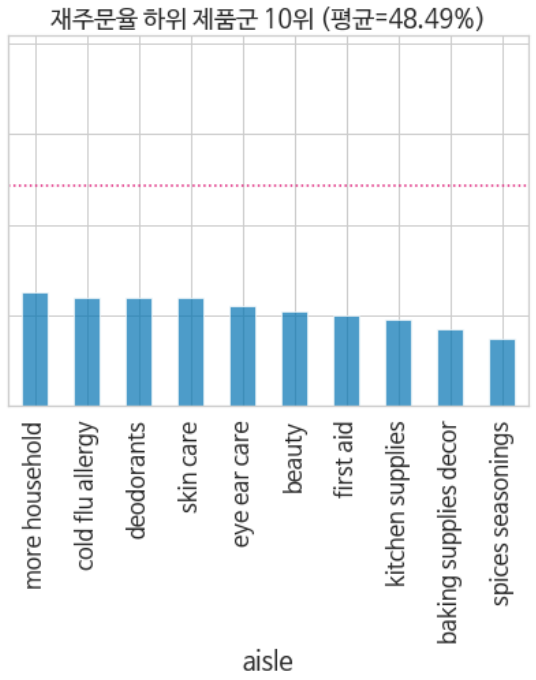
- 동일 상품에서 총 주문량에 비해 재주문량이 많은지 비교하면 **유제품** 비중이 높습니다.
- **우유 제품(84구역)**이 대부분을 차지하며, **83%의 높은 재주문** 비율이 나타납니다.
- 우유 이외는 바나나, 물이 있습니다.

1) milk + organic cream

4. 주문 상품 구성

② 소분류 내에서 **재주문 상품 비율** 상위 항목 10개, 하위 10개



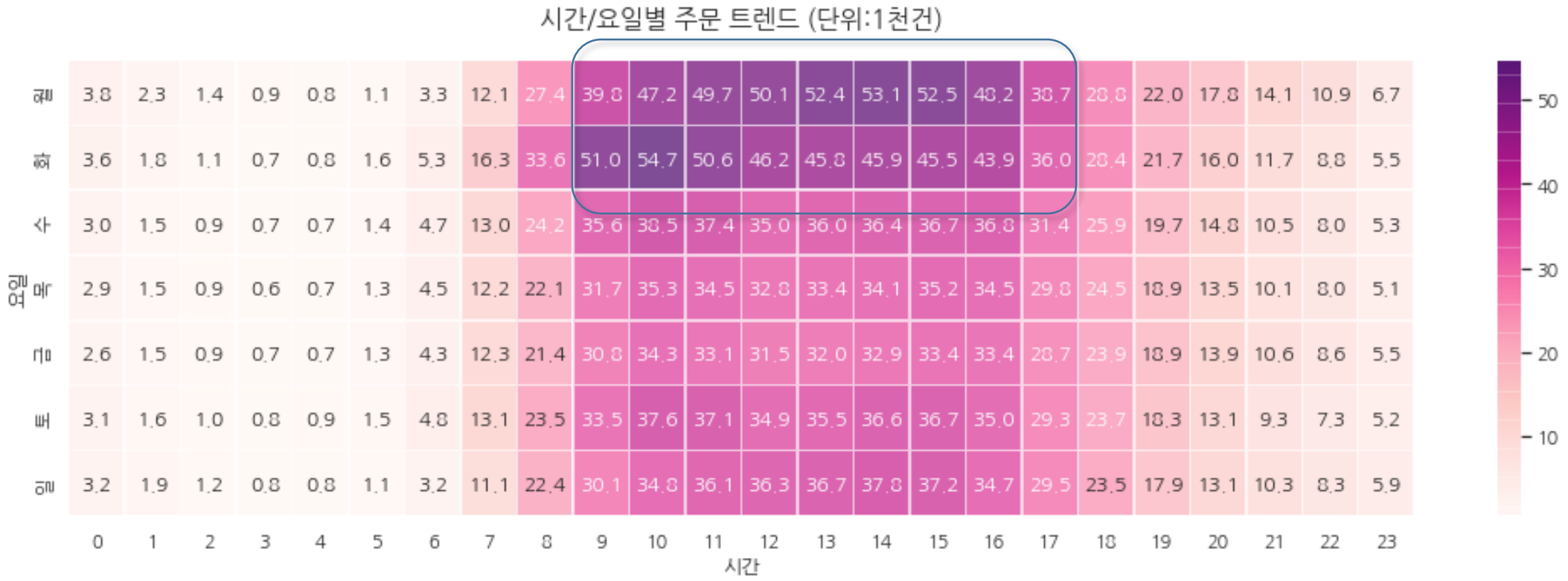


- **제품군에 따라 재주문 비율**이 다른지 살펴보면 **유통기간**에 따라 차이를 발견할 수 있습니다.
- 재주문 비율이 높은 상품: **유통기간이 짧은** 우유, 계란, 과일, 야채 등 신선 농산물 위주
- 재주문 비율이 낮은 상품: **유통기간이 긴** 상품 위주

5. 주문 요일 및 시간

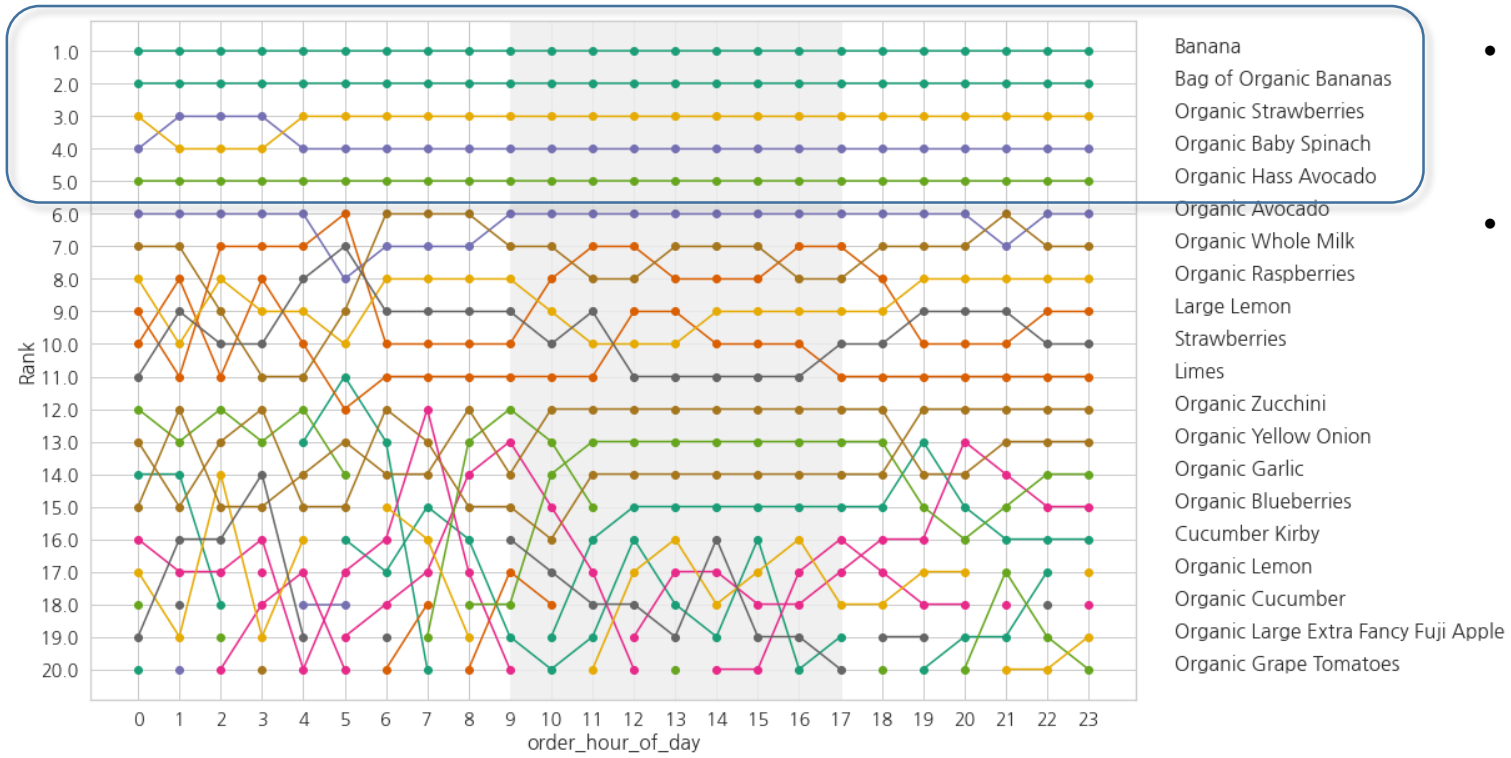
① **주문 트렌드**는 어떤지?

- 모든 요일에 **09~17시.** 특히 **월요일**과 **화요일** 오전까지 주문이 많습니다.
- **재주문 수**와 **재주문 상품의 주문 갯수**로 비교해 봐도 패턴이 동일합니다.
- 전체 주문이 증가하는 것이며, 고객당 주문건수는 증감 차이가 적습니다.



5. 주문 요일 및 시간

② 재구매 상품 중 **구매 시간에 따른 판매 상위모델에 차이**가 있는가?

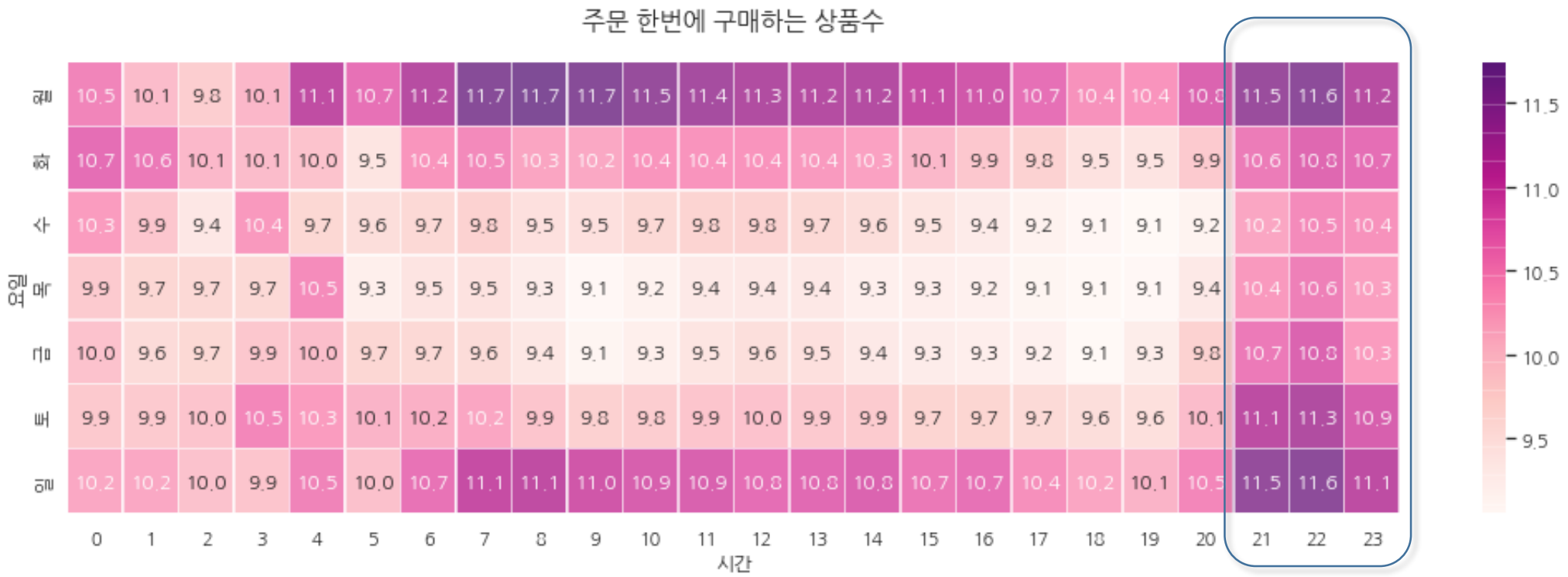


- **1~5위** 까지는 거의 변화가 없고, **6위 아래**는 변화가 있습니다.
- 구매가 집중된 9시~17시 구간과 그 외 시간의 구매 상품이 많이 다릅니다.

5. 주문 요일 및 시간

③ 재주문 한번에 **구매하는 상품이 많은** 시간/요일은?

- **늦은 밤**(21~23시) 구매와, 주문량이 많은 **월,화**, 주말인 **일**에 주문 수량이 많습니다.



5. 주문 요일 및 시간

④ 재주문까지 **걸리는 시간**은?

- 수치가 작을 수록 금방 재주문을 했다는 의미이고, 자주 주문함을 의미합니다
- **새벽시간** 구매는 이전 주문 이후 **오랜만에 구매**한 경우가 많습니다.

가설 검증 결과: 재주문에 패턴이 존재한다. → Yes!

- **패턴**(주문 간격 / 횟수 / 수량 / 상품 구성 / 요일 및 시간)에 따른 **마케팅 인사이트**

1. 주문 간격: 7일 (~10일 이내)
고객 타겟 판촉일정을 맞춥니다. 메일링,문자로 판촉내용을 알리는 주기를 7일로 합니다.

2. 주문 횟수: 주문 빈도에 따른 **고객 관리**를 합니다. (평균적으로 **4회** 주문이 가장 많고, 5회부터 감소 추세)
→ **4회 이상과 이하 이용 고객을 분리**. 초기에 지속적인 구매를 독려하는 것이 중요합니다.
→ 이용 횟수 **10회 단위**로 고객을 구분하여, 구매 적립포인트를 다르게 하는 등 차별화된 이점을 줍니다. **상위 10% (40회 이상 구매고객)**는 밀착 관리 합니다.

3. 주문 수량: 구매수량↑ → 재주문 상품 비중↑
주문이 많고 적음에 따라 추천상품을 다르게 합니다.

4. 상품 구성: 유통기간 짧은 상품 → 재주문↑
유통기간에 따라 한번에 구매하는 양이 달라집니다.
→ 보관이 편한 상품은 대량 패키지 구성
→ 조금씩 자주 구매 해야 하는 상품은, 기존 구매 상품을 먼저 추천합니다.

5. 요일 및 시간: 월,화, 9~17시↑
주문시간에 따른 구매패턴이 다르므로 상황에 맞는 판촉을 계획합니다. 이전 주문 간격, 주문 시간에 따른 구매 수량, 구매 품목 등을 반영 합니다.

연관상품 분석

- 고객들의 상품 **구매 데이터**를 이용하여 **품목간의 연관성**을 분석하고자 합니다.
- 구매한 상품간 연관성이 있는 제품군을 묶어서 살펴보고, **상품 추천**에 활용하고자 합니다.

Word2Vec이용한 **Product2Vec**

- Word2Vec은 구글에서 발표한 자연어 처리 기술로, 대량의 학습 데이터셋을 빠르게 학습할 수 있어서 **대량의 컨텐츠를 이용** 해야하는 추천 시스템에서 활용하기 적합합니다.
- 추천 시스템에서 주로 Word2Vec은 **상품을 벡터화** 시킬 때 사용합니다. 이렇게 벡터화 시킨 결과는 유사한 상품 추천, 마이크로 카테고라이징, 의미 벡터로써 활용할 수 있고, **기업의 활용사례가 많습니다**.
  예) 아프리카 TV(live2vec), Sportify(song2vec), Criteo(meta-pro2vec)등

- 이번 프로젝트에서는 **User_id당** 구매한 **Product 수가 여러가지** 인데, 이를 벡터화 시켜 분석합니다.
  (파이썬의 gensim 패키지에 구현된 Word2Vec(클래스) 활용하여 구현하였습니다.)

상품 id에 따라 유사도 측정

- 학습된 Word2Vec을 통해 **상품 id**를 넣으면 **유사한 제품을 추천**받을 수 있게 구성했습니다.
- 대표적인 상품 id 3가지(샘플링) 결과를 보면, 입력한 단어와 유사하게 추출함을 알 수 있습니다.

	product	similarity
0	Banana	1.000000
1	Bag of Organic Bananas	0.704021
2	Seedless Red Grapes	0.579649
3	Bartlett Pears	0.578962
4	Organic Banana	0.530327
5	Granny Smith Apples	0.519383
6	Organic Fuji Apple	0.513325
7	Total 0% Greek Yogurt	0.510567
8	XL Emerald White Seedless Grapes	0.505114
9	Total 2% All Natural Plain Greek Yogurt	0.490714
10	Almond Breeze Original Almond Milk	0.473399

	product	similarity
0	Organic Whole Milk	1.000000
1	Organic Reduced Fat Milk	0.789070
2	Organic Lowfat 1% Milk	0.726632
3	Whole Organic Omega 3 Milk	0.715100
4	Organic Lactose Free Whole Milk	0.633504
5	Organic Multigrain Waffles	0.625995
6	Organic Whole Grassmilk Milk	0.613493
7	Organic Whole String Cheese	0.610874
8	1% Lowfat Milk	0.608748
9	Organic Mini Homestyle Waffles	0.604269
10	Organic Yokids Lemonade/Blueberry Variety Pack...	0.599768

	product	similarity
0	Cookie Tray	1.000000
1	Vegetable Tray With Low Fat Dressing	0.910919
2	Red Plastic Cups	0.907853
3	100% Juice, Variety Pack	0.906861
4	Cheesecake	0.905112
5	Entertainment Crackers	0.900725
6	Butter Pound Cake	0.899281
7	Premium Lots of Pulp Orange Juice	0.896205
8	Madeleines	0.890736
9	Cola Cans	0.888052
10	Ice Cream Variety Pack	0.880662

KMeans clustering

- KMeans clustering은 주어진 데이터를 k개의 클러스터로 묶는 알고리즘입니다.
- Word2Vec 학습된 **상품들을 그룹화**(293개)하여 **추천시스템**에 활용하고자 합니다.
  - 상품수(약 4.9만개)가 매우 많으므로 우선 추가 분석이 가능한 300개 수준으로 구분

```
printClusterMembers(2, topn=10) # 우유,야채,치즈 등
```

```
84 / 1% Low Fat Milk  https://www.google.co.kr/search?tbm=isch&q=
123 / Asian Chopped Salad with Dressing  https://www.google.co.kr/
83 / Red Onions  https://www.google.co.kr/search?tbm=isch&q=Red+On
83 / Asparagus Spears  https://www.google.co.kr/search?tbm=isch&q=
123 / Southwest Chopped Salad Kit with Dressing  https://www.goog
129 / Sweet Potato Fries with Salt  https://www.google.co.kr/searc
91 / Dark Chocolate Almond Milk  https://www.google.co.kr/search?t
123 / Organic 50/50 Blend Salad  https://www.google.co.kr/search?t
96 / Select Natural Applewood Smoked Turkey Breast  https://www.gc
86 / 100% Liquid Egg Whites  https://www.google.co.kr/search?tbm=
```

Great Value 1% Low Fat Mil..

Asian Chopped Salad (Qui...

Grilled Asparagus Spears F

Dark Chocolate Almond Be..

Bob Evans® 100% Liquid Egg .

```
printClusterMembers(80, topn=10) # 설탕, 휘핑크림, 베이킹재료
```

```
17 / Pure Baking Soda  https://www.google.co.kr/search?tbm=isch&q=
17 / Pure Granulated Cane Sugar  https://www.google.co.kr/search?t
17 / Pure Cane Granulated White Sugar  https://www.google.co.kr/se
119 / Original Whipped Topping  https://www.google.co.kr/search?tb
17 / Granulated White Cane Sugar  https://www.google.co.kr/search?
17 / Pure Dark Brown Cane Sugar  https://www.google.co.kr/search?t
17 / Pure Cane Confectioners Powdered Sugar  https://www.google.cc
17 / Toll House Semi Sweet Chocolate Mini Morsels Chips  https://w
97 / Non-Stick Parchment Paper  https://www.google.co.kr/search?tb
17 / Pure Cane Golden Brown Sugar  https://www.google.co.kr/search
```

Arm & Hammer Pure Bakin...

C & H Pure Granulated White C..

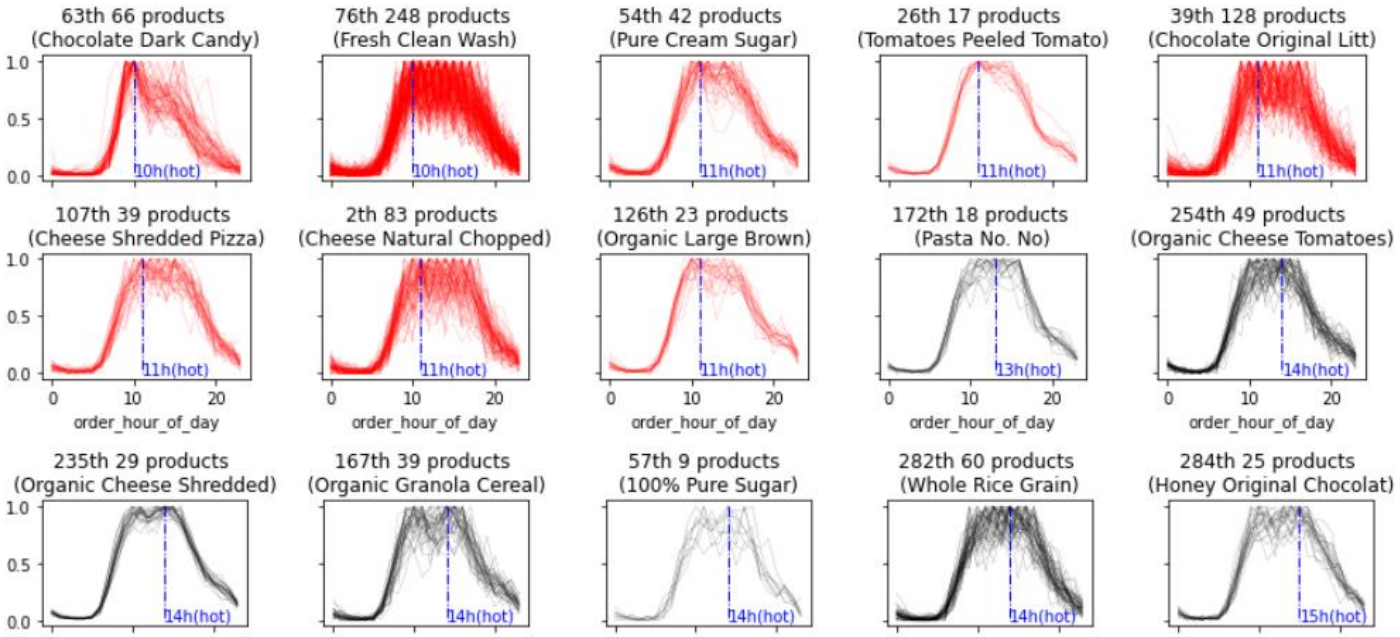
Kraft Cool Whip Original Whi...

C&H® Pure Cane Dark Brown S

Buy Flair Non Stick Baking Paper ,

KMeans clustering

- 같은 클러스터내 상품들(샘플링)의 시간별 **구매 패턴**이 비슷한지 추가로 검증해보았습니다.
- 오전, 오후에 따라 그래프 색상이 다릅니다. (오전: 붉은색, 오후: 검은색)
- 같은 클러스터내에 상품들의 **구매가 가장 많은 시간대가 유사**하게 잘 묶였습니다.



추천 시스템 활용방안 및 추후 발전방향

추천 시스템 활용방안

- 클러스터링으로 나눠진 상품들로 **타겟 마케팅**을 진행할 수 있습니다.
- EDA를 통해 발견한 **마케팅 인사이트에 위의 추천시스템을 함께 적용할 수 있습니다.**

- 특정 타이밍, 필요 수량, 어떤 상품인지 등 고객의 상황에 맞게 상품을 추천해야 하는데
① 기존 구매목록 기반으로 구매했던 상품과
② 같은 클러스터로 분류된 유사한 제품을 적절히 조합하여 추천합니다.

한계점 및 추후 발전 방향

- 제한된 수로 **클러스터**를 나누었는데 더 **세분화** 하여 상품이 구분되면 정교한 추천시스템을 구현할 수 있을 것 같습니다.
- **상품 추천 시스템**을 구성하는 방식은 여러가지가 있는데, Product2Vec외에 다른 방식으로 진행했을 때 장단점도 더 파악하여 혼합하여 사용하거나, 구매성공률을 높이고 싶습니다.
- 이번 분석데이터에는 상품당 **가격** 자료나 동일 제품당 구매 수량정보 등이 없습니다. 수량만큼 가격정보가 중요한데, 이후에는 가격 정보가 있는 데이터로 **매출 분석**도 진행하고 싶습니다.

Thank you for your attention!