CP1

DA Track Op. 1

AIB 15th 박건후

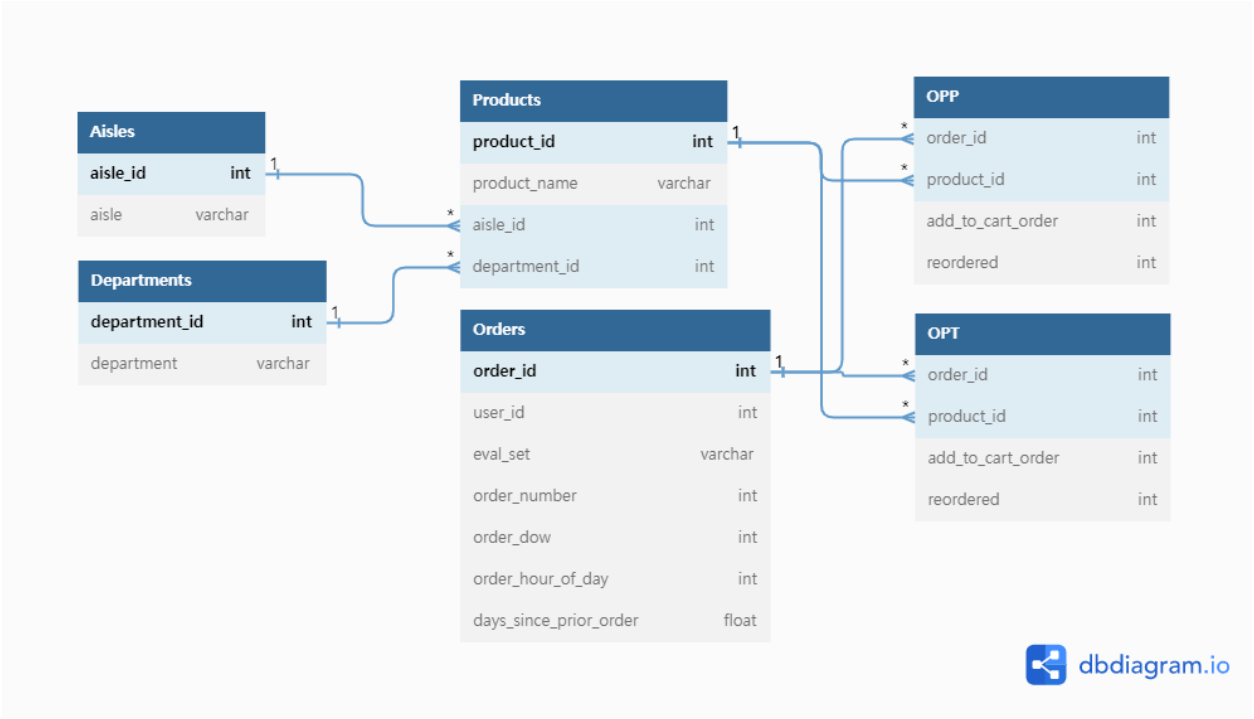
Instacart Market Basket Analysis

프로젝트 개요

1. 데이터 분석을 통해 인사이트를 도출한다.
2. 수요 예측 분석 및 주요 특성을 파악한다.
3. 주요 상품 및 고객 구매 패턴에 따른 예측 및 추천한다.
4. 고객 세그맨테이션을 통해 핸들링 전략을 제안한다.

데이터 소개

ERD



테이블 설명

- orders : 주문정보
- produscts : 제품정보
- departments : 제품 카테고리
- aisles_df : 제품 상세카테고리
- order products prior : 과거 구매자군의 제품주문내역
- order products train : 현재 구매자군의 제품주문내역

EDA 및 분석

※ 모든 EDA는 과거/현재 구분이 아닌 전체 구매자군에 대해 이루어졌다.

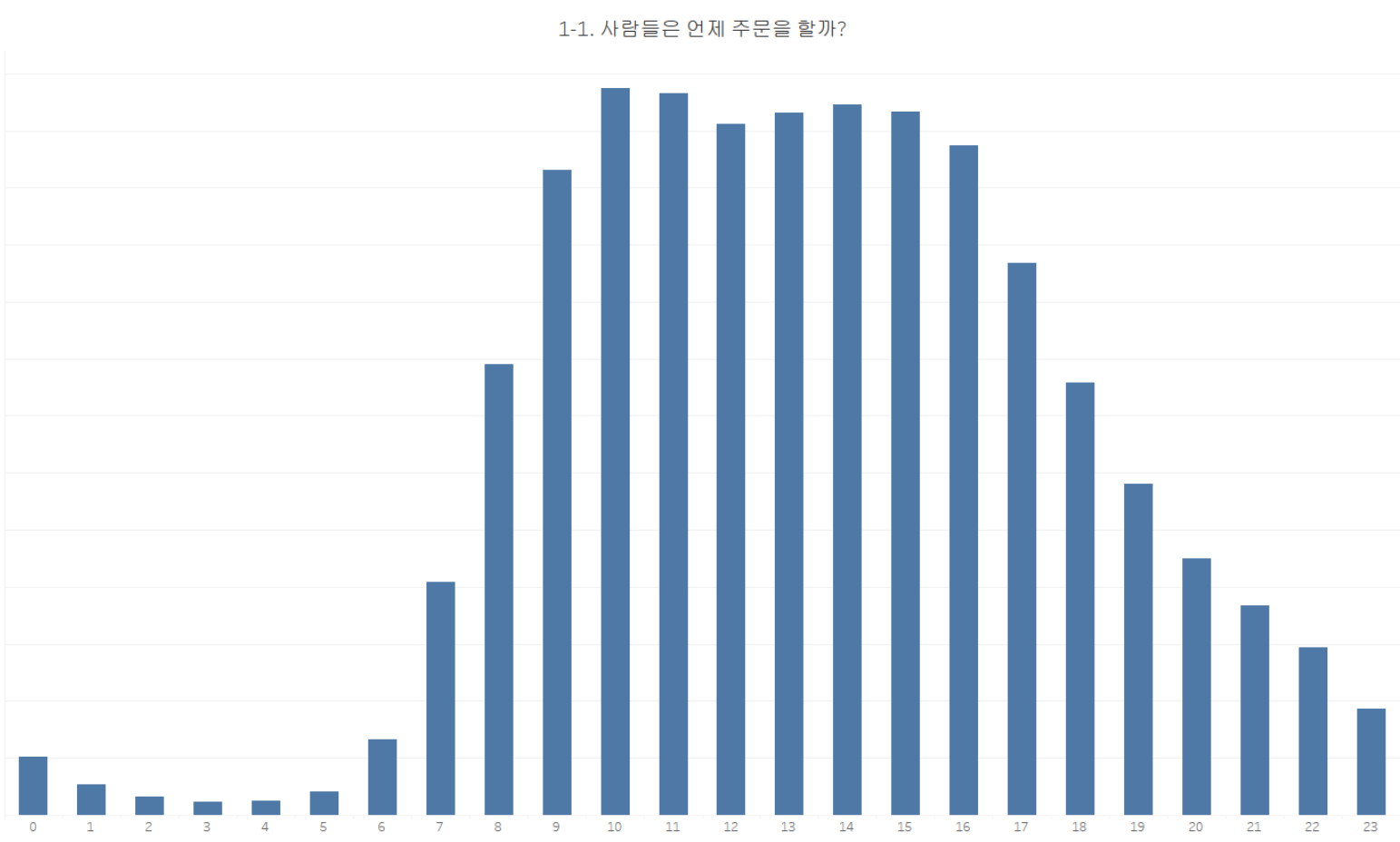
▼ EDA 목록

1. 사람들은 언제 주문을 할까?
    - 하루 중
    - 일주일 중

    - 요일별/시간대별 구매 추이
2. 얼마나 많은 항목을 샀을까?
3. 가장 많이 주문한 것은 무엇일까?
    - 카테고리
    - 제품
4. 언제 재주문을 할까?
5. 얼마나 재주문을 했을까?
6. 얼마나 자주 같은 제품을 주문했을까?
    - 그 중 어떤 카테고리가 재구매율이 높을까?
    - 그 중 어떤 제품이 재구매율이 높을까?
    - 시간대별 재주문 추이
    - 요일별 재주문 추이
    - 요일별/시간대별 구매 추이
7. 장바구니 추가와 재주문율 관계
    - 그 중 어떤 제품을 제일 처음 장바구니에 넣을까?
8. 연관성
    - 마지막 주문 날짜와 재주문
    - 주문 수와 재주문
9. VS
    - Organic vs Non-organic
    - Reordering Organic vs Non-organic
10. ~~카테고리별로 각각 어떻게 구성되어 있을까?~~ ⇒ 시간 부족으로 실패
    - ~~양도 나타낸다.~~

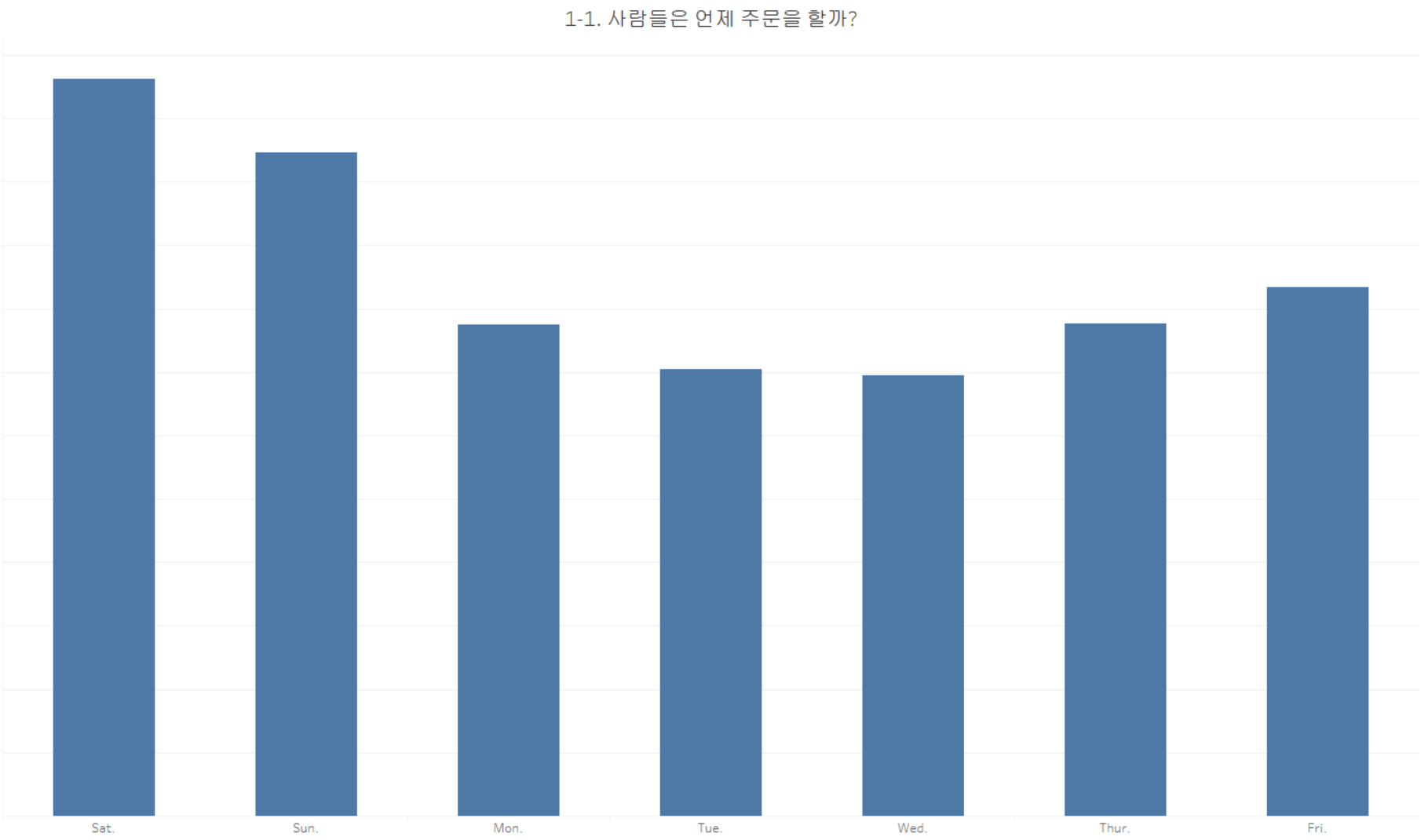
1. 사람들은 언제 주문을 할까?

▼ 하루 중



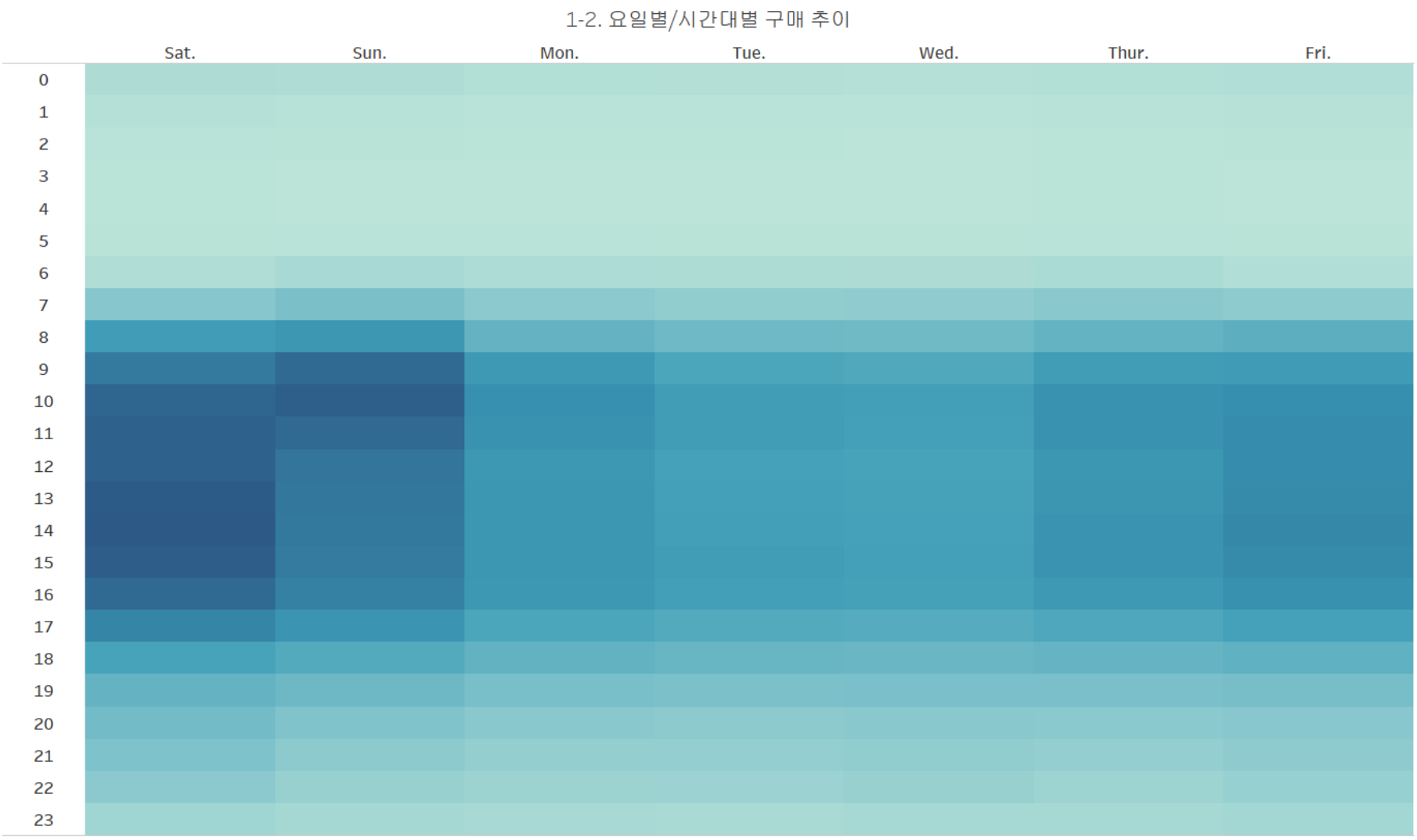
- 하루 중 대부분의 주문은 10~16시 사이에 일어나고, 그 중 아침 이후인 10~11시와 점심 이후인 14~15시에 가장 많이 일어난다.

▼ **일주일 중**



- 일주일 중 대부분의 주문은 주말에 일어난다.

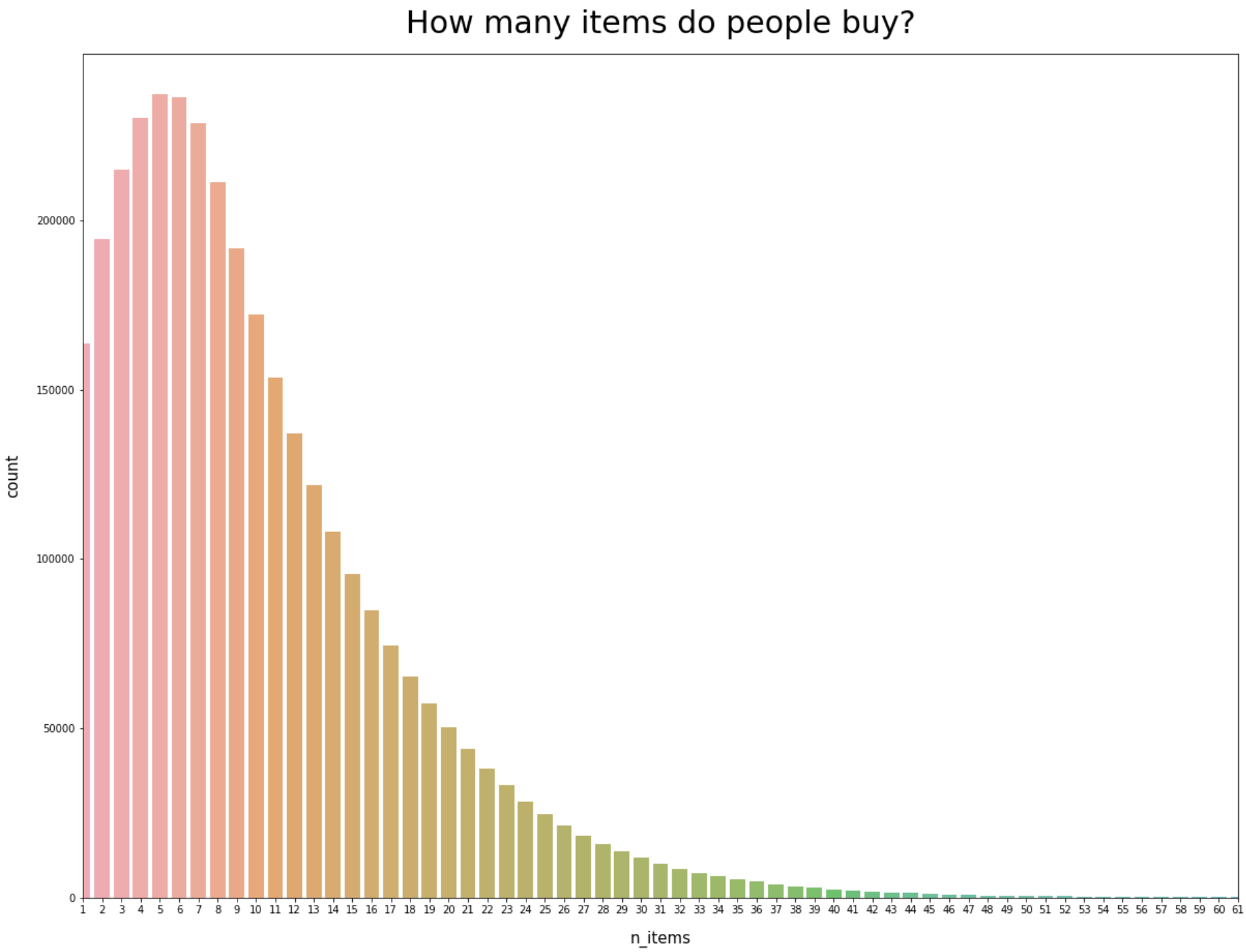
▼ **요일별/시간대별 구매 추이**



- 주말 중 토요일은 14~15시, 일요일은 9~10시에 주문이 많이 일어난다.

2. 얼마나 많은 항목을 샀을까?

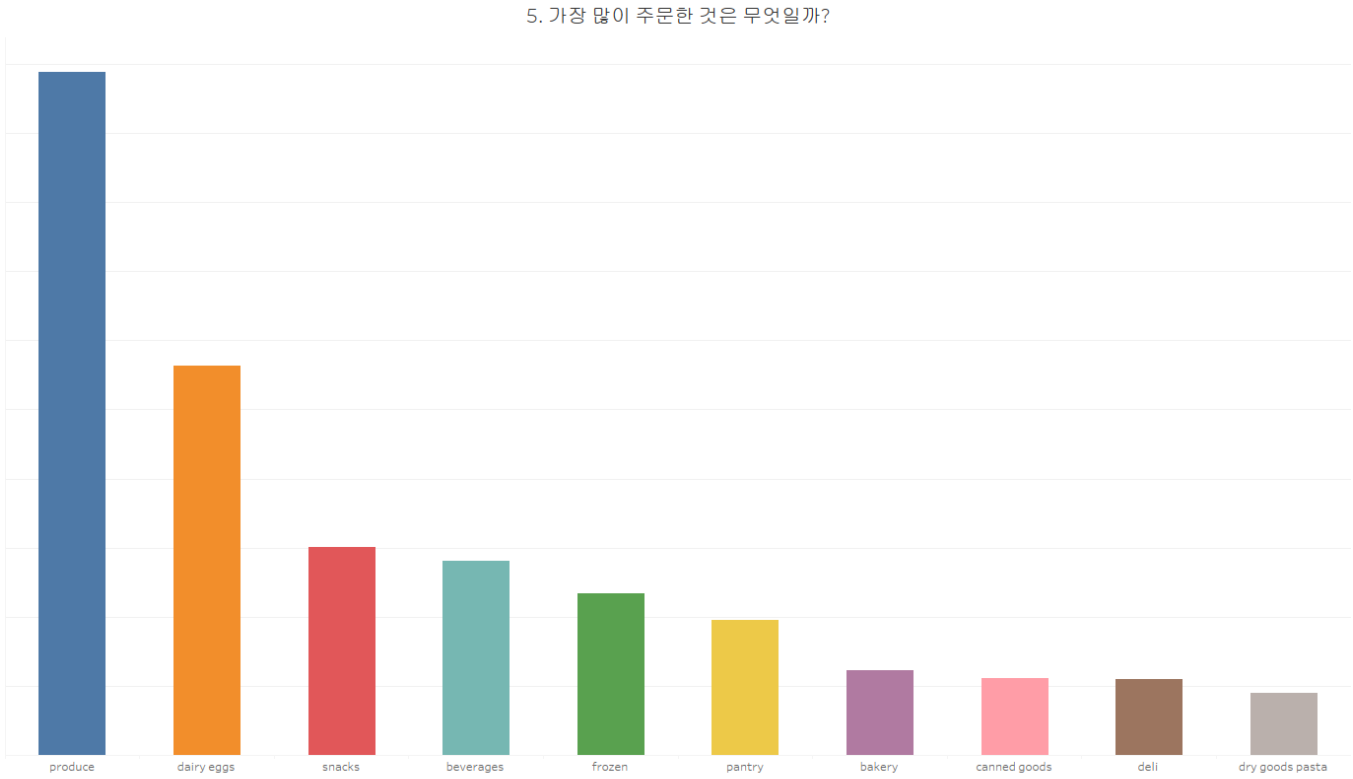
▼ 사람들은 한 번 구매를 할 때 보통 몇 개의 항목을 구매할까?



- 사람들은 구매할 때 보통 한 번에 5가지 항목을 구매한다.

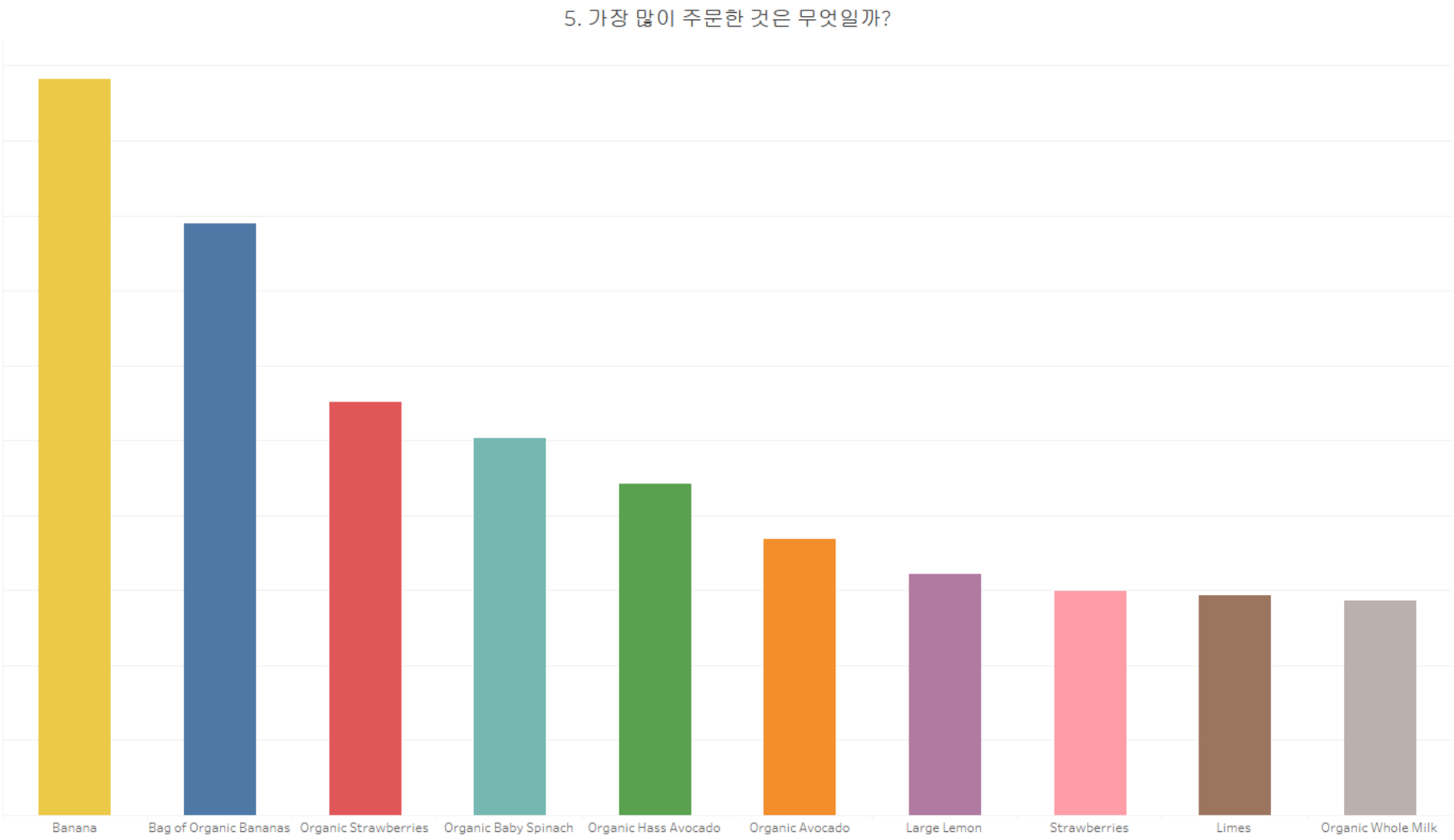
3. 가장 많이 주문한 것은 무엇일까?

▼ 카테고리



- 농산물(produce)가 압도적으로 가장 많이 판매되었고, 그 다음은 유제품(dairy eggs)이 많이 팔렸다.

▼ **제품**



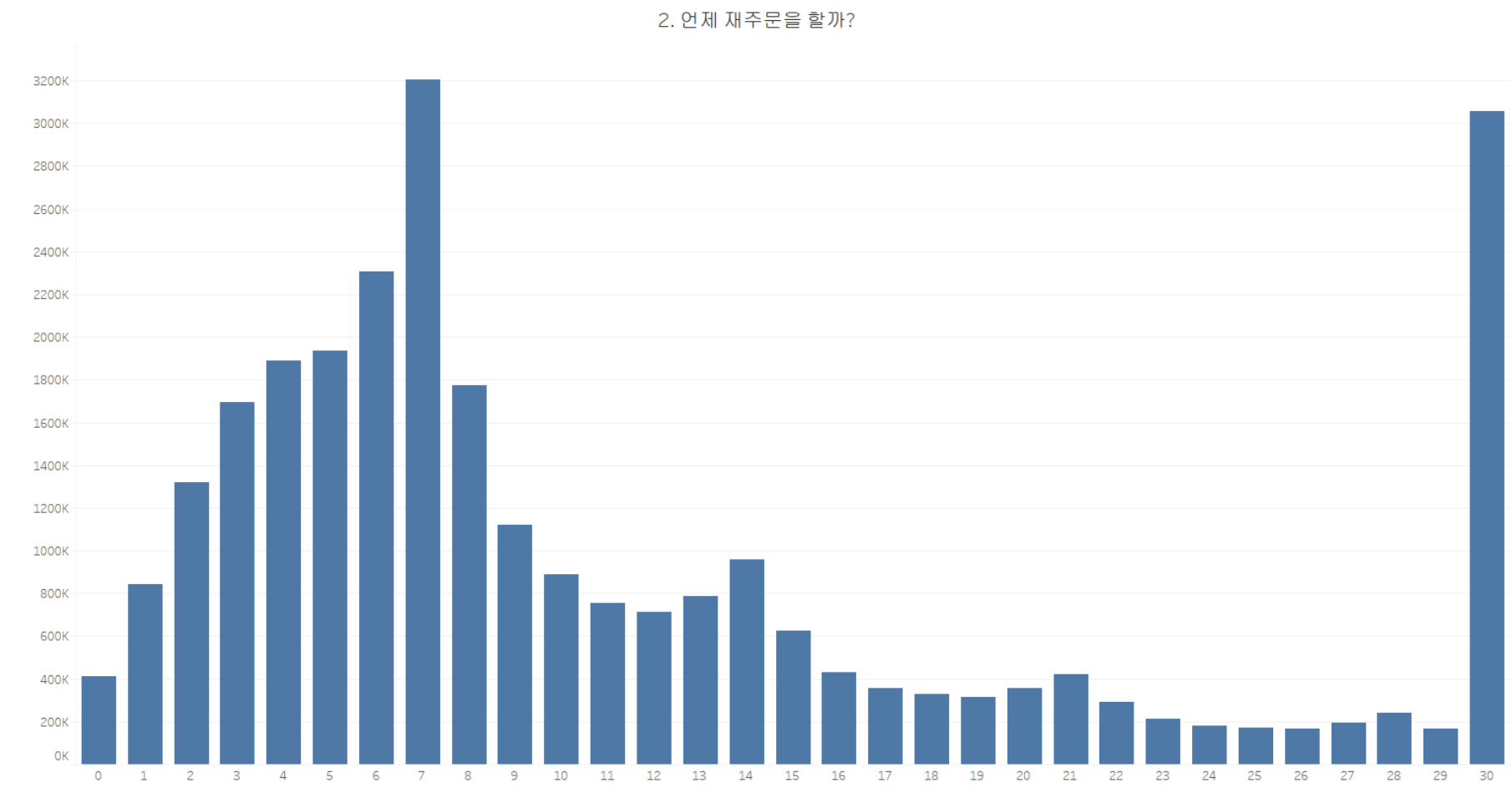
- 바나나가 가장 많이 팔렸고, 그 다음도 organic 바나나이다. 그리고 그 이후로도 organic 농산물들이 우세한 판매량을 보여준다.

4. 언제 재주문을 할까?

※ 주문율 : 사람들이 얼마나 주문을 하는지

※ 재주문율 : 같은 상품을 얼마나 다시 주문을 하는지

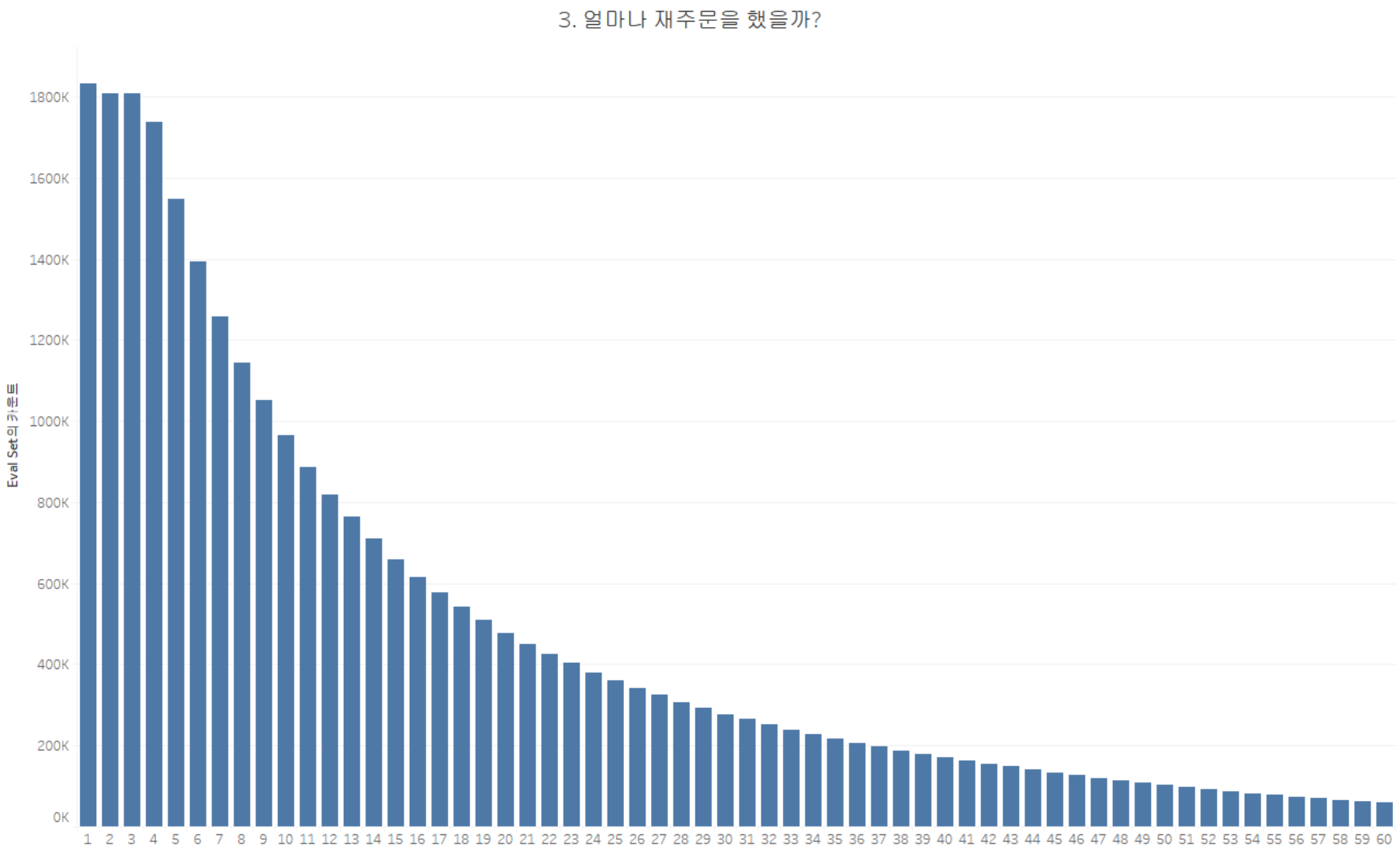
▼ **1달 중 언제 다시 주문을 할까?**



- 사람들은 보통 일주일 안에 다시 주문을 하고, 마지막 날에 주문을 한다.

5. 얼마나 재주문을 했을까?

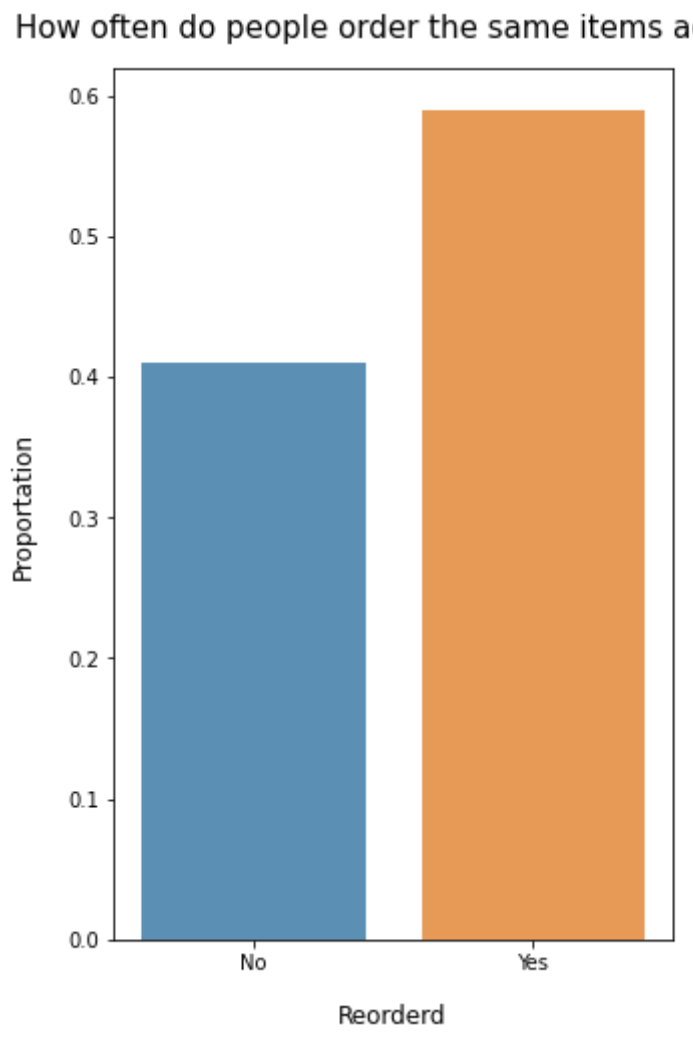
▼ 그렇다면 보통 몇 개 정도 재주문을 할까?



- 보통 3개 정도 재주문을 한다.

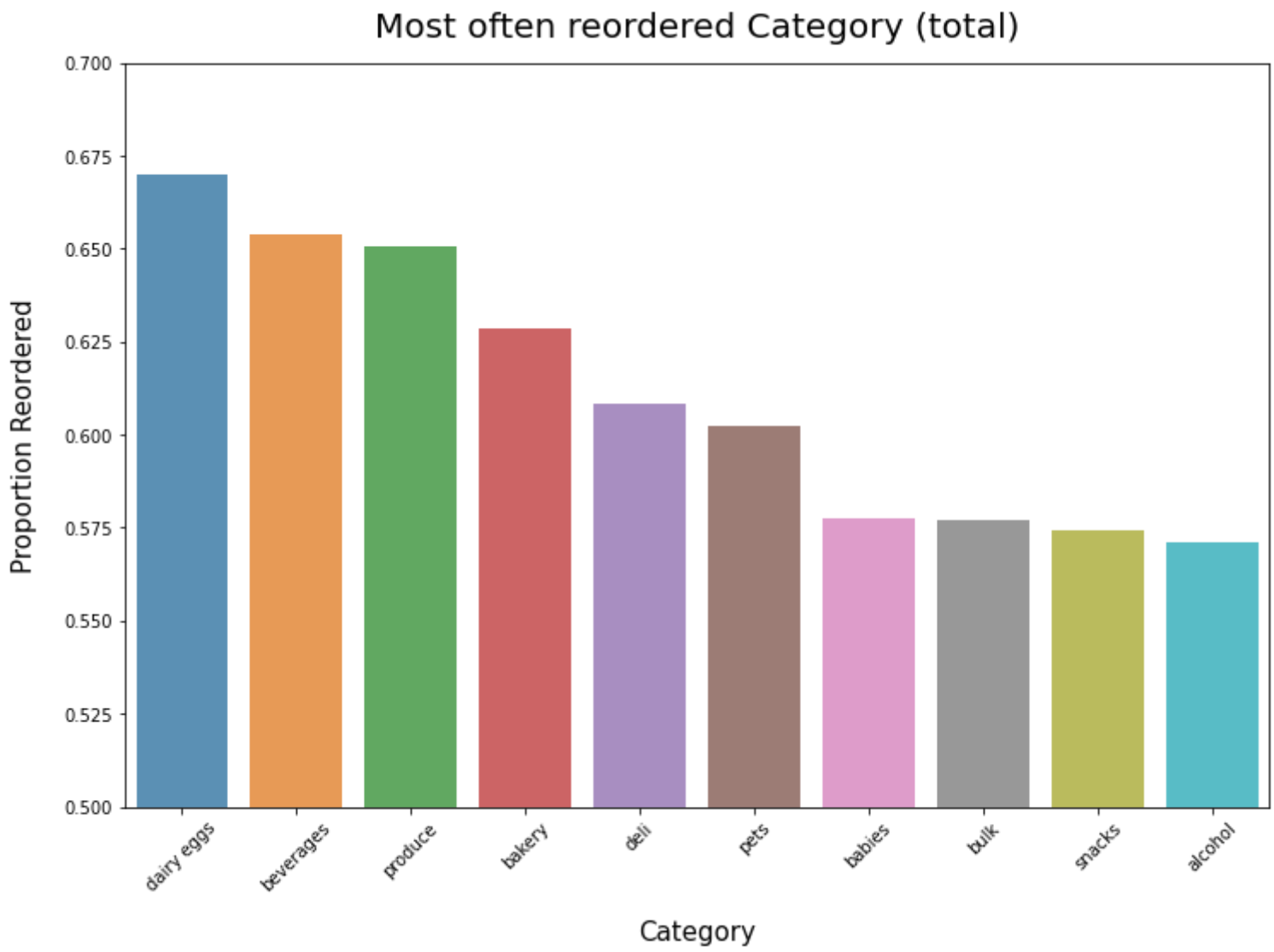
6. 얼마나 자주 같은 제품을 다시 주문했을까?

▼ 얼마나 자주 같은 제품을 다시 주문했을까?



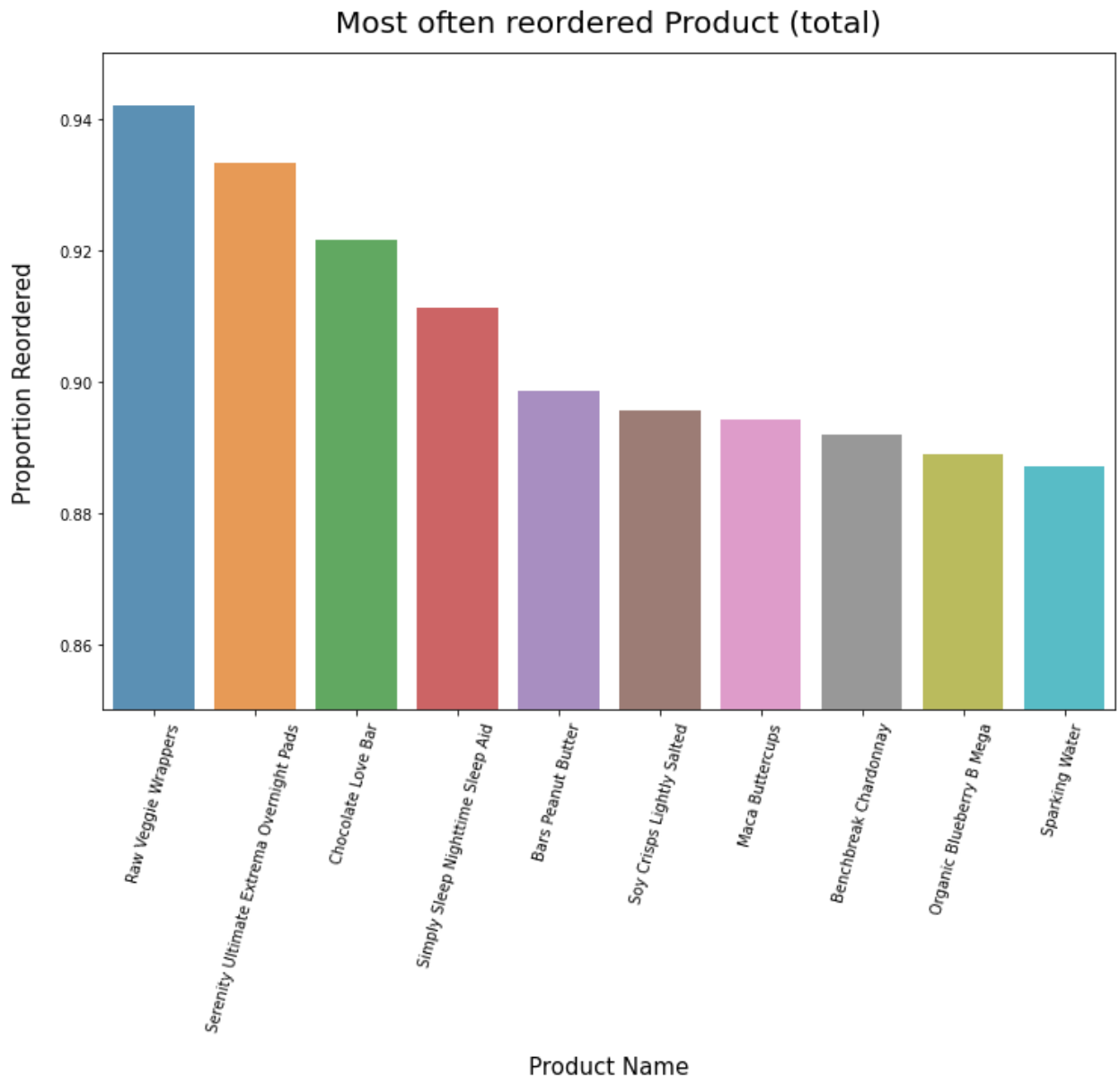
- 주문한 품목의 약 59%가 재구매된다.

▼ 그 중 어떤 카테고리가 재구매율이 높을까



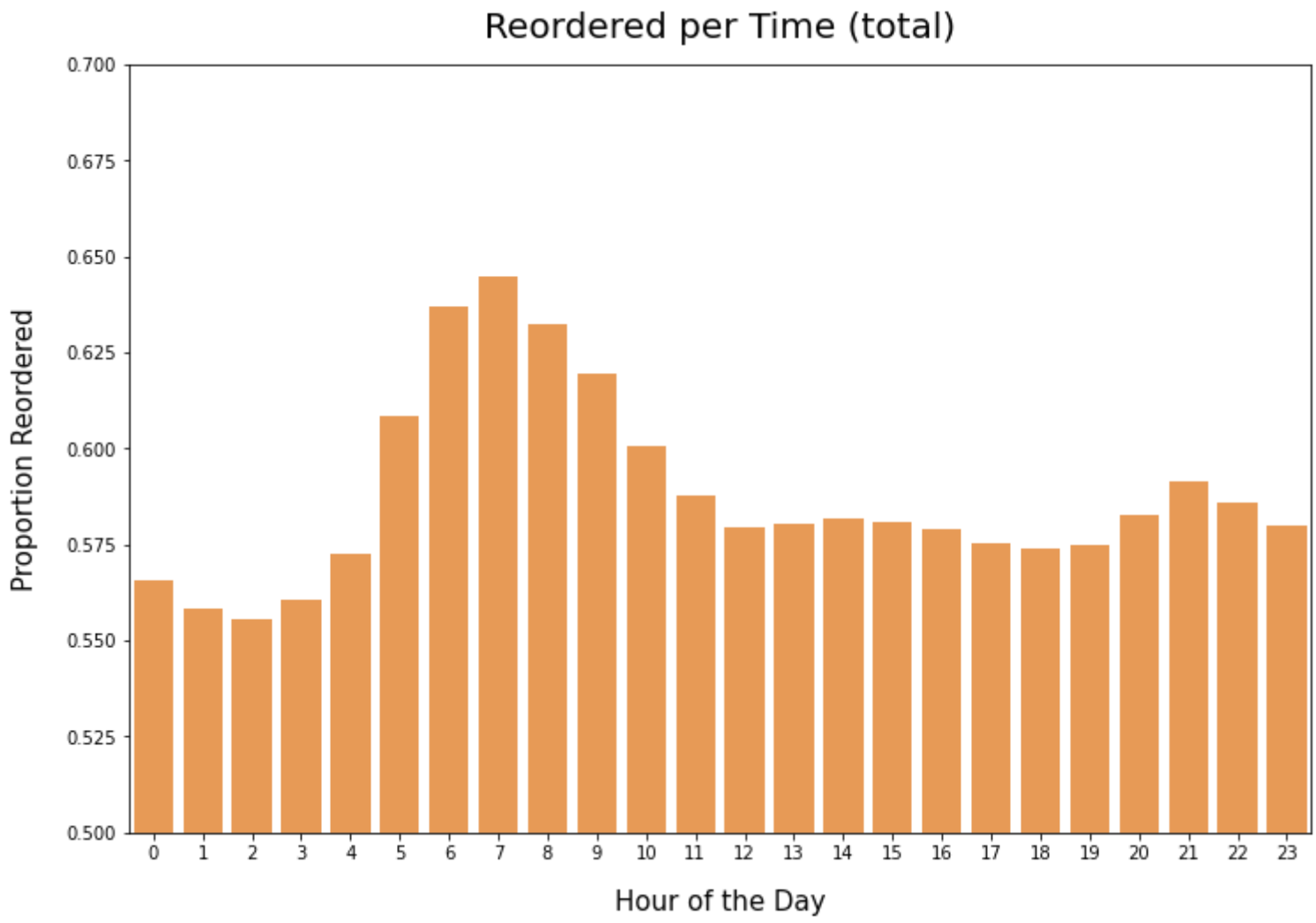
- 이 10개 카테고리가 재주문 가능성이 가장 높다.

▼ 그 중 어떤 제품이 재구매율이 높을까



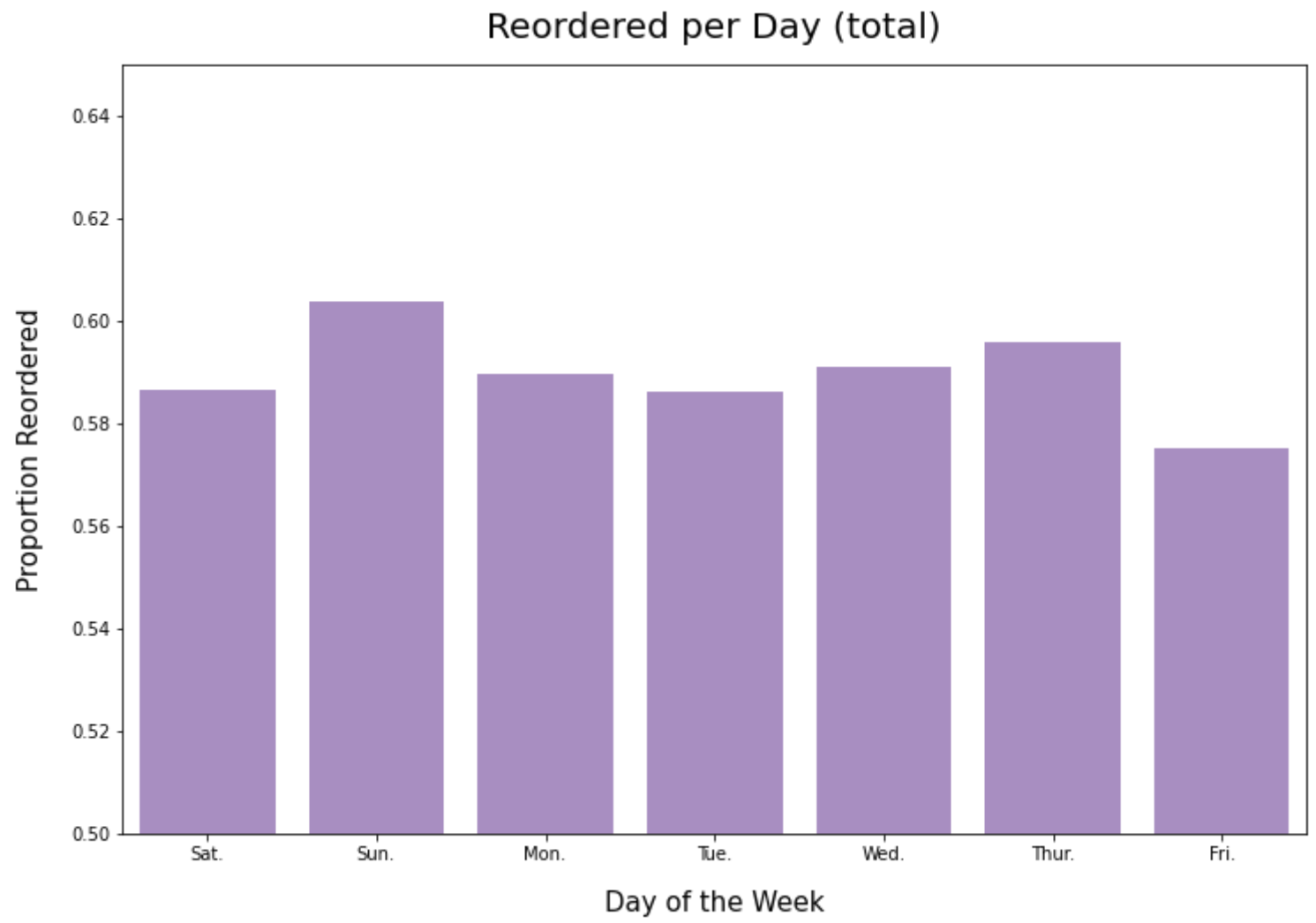
- 이 10개 제품은 재주문 가능성이 가장 높다.

▼ 시간대별 재주문 추이



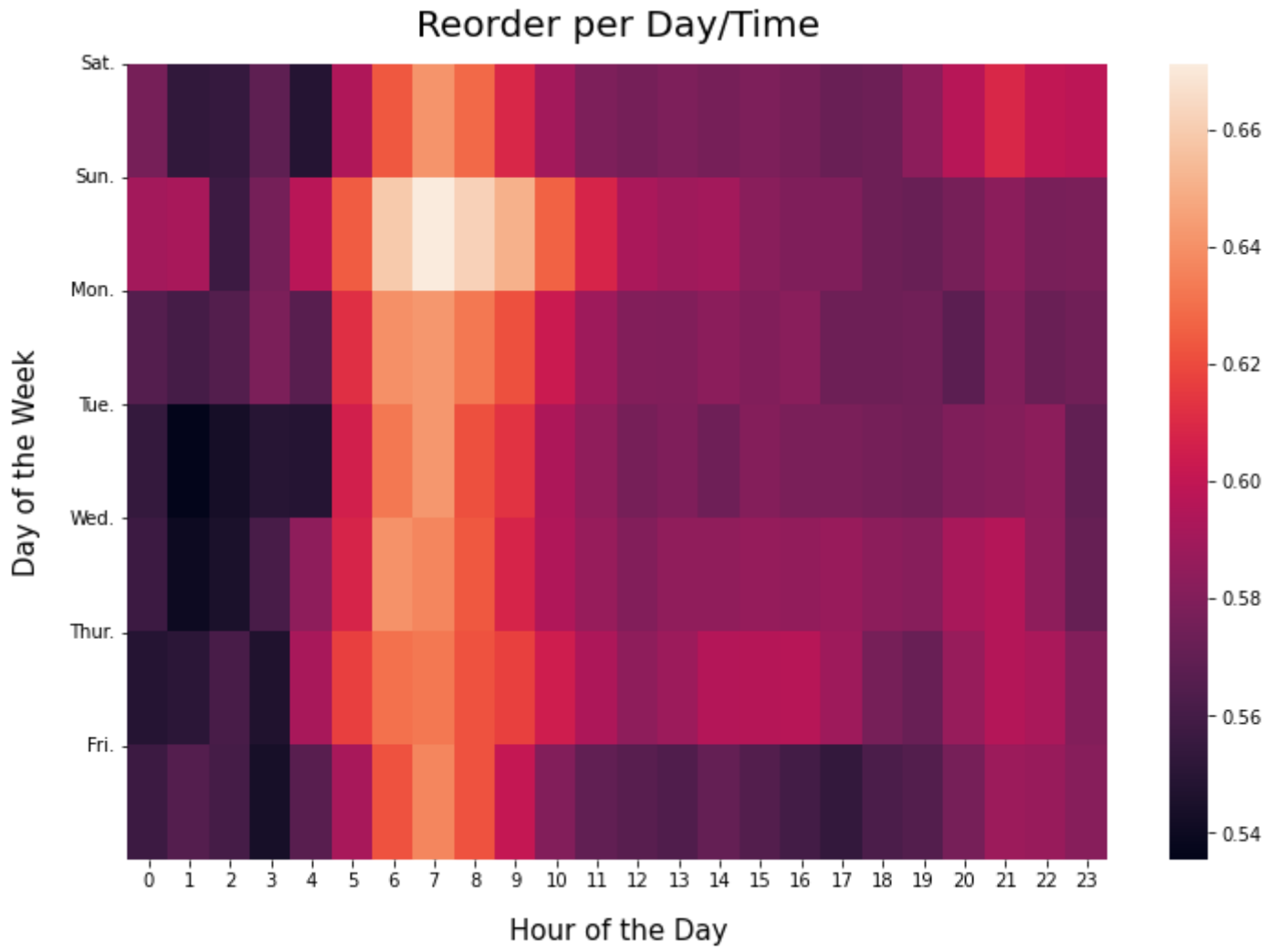
- 6~8시 사이에 재주문이 많이 이루어진다.
- 보통 주문이 10~17시에 이루어지는 것과 다른 시간이다.

▼ 요일별 재주문 추이



- 요일별로 보아도 보통 주문이 주말에 가장 많이 일어났던 것에 반해 재주문은 일요일 다음으로 목요일이 높다.

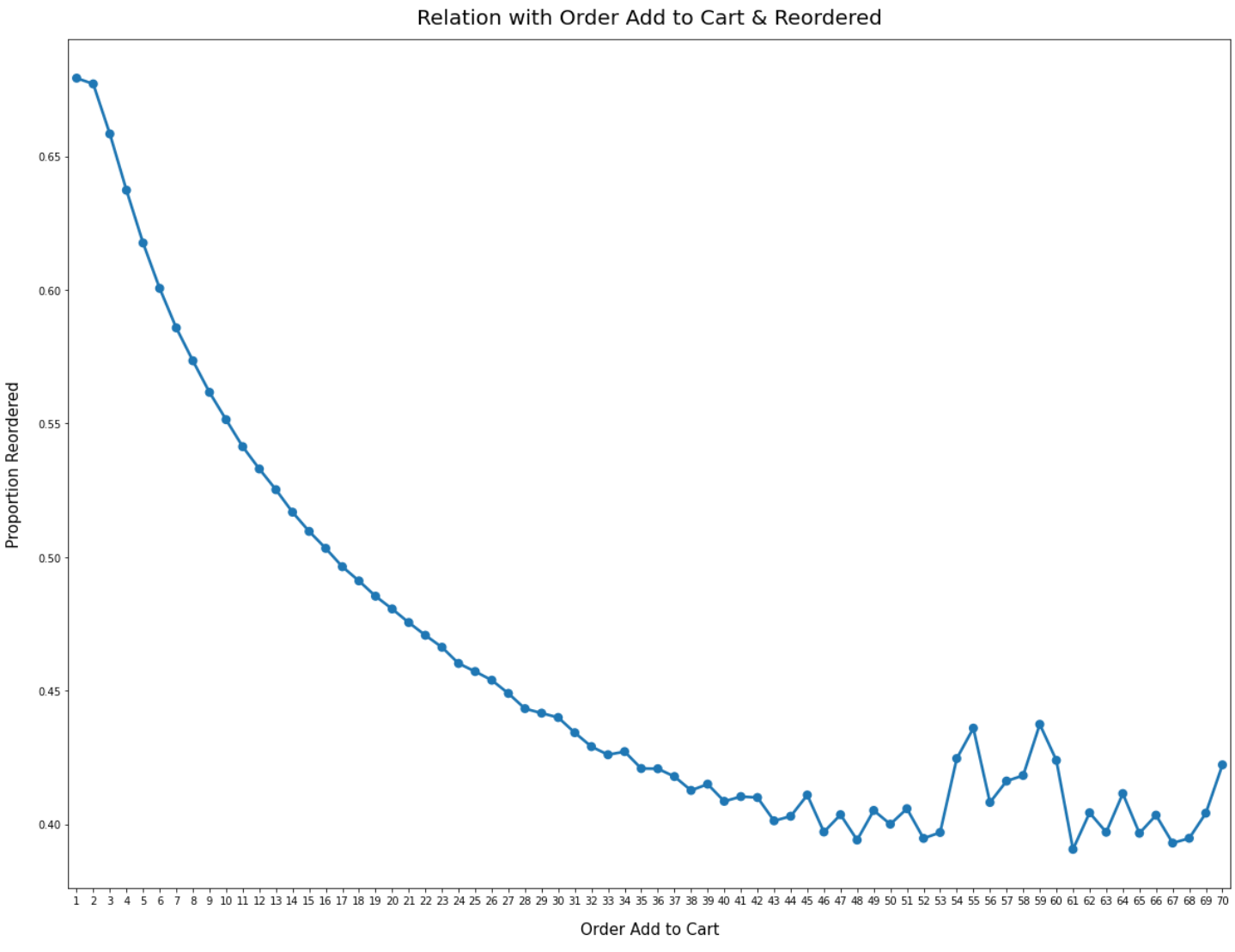
▼ 요일별/시간대별 재주문 추이



- 특히, 일요일에 압도적으로 재주문율이 높다.

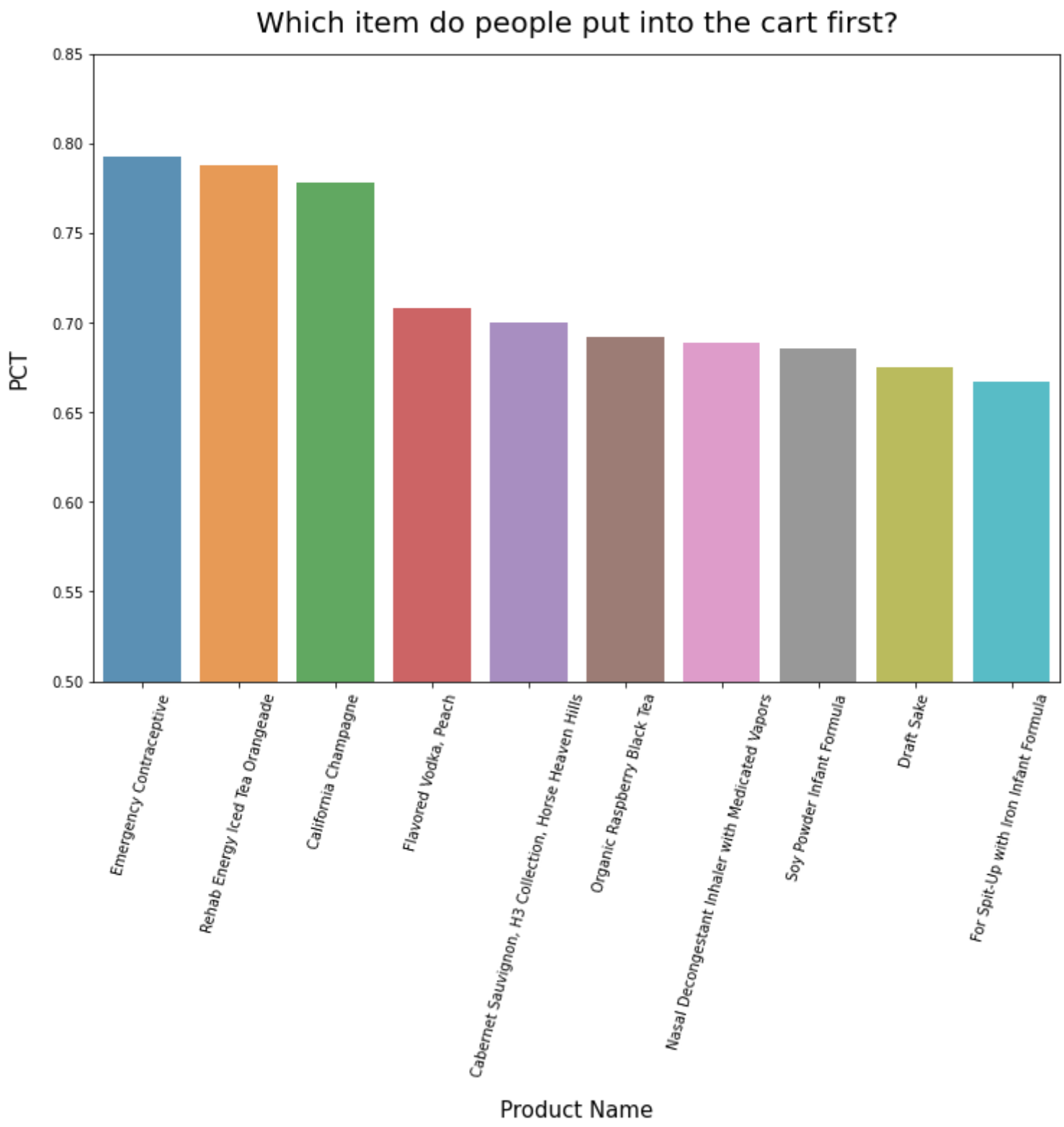
7. 장바구니 추가와 재주문율 관계

▼ 장바구니에 추가하는 순서와 재주문은 관계가 있을까?



- 제일 처음에 장바구니에 추가된 제품이 나중에 추가된 제품에 비해 다시 주문될 가능성이 높다.
- 따라서, 사람들은 자주 사용하는 제품을 먼저 주문한 후에 새로운 제품을 찾는 경향을 가지고 있음을 알 수 있다.

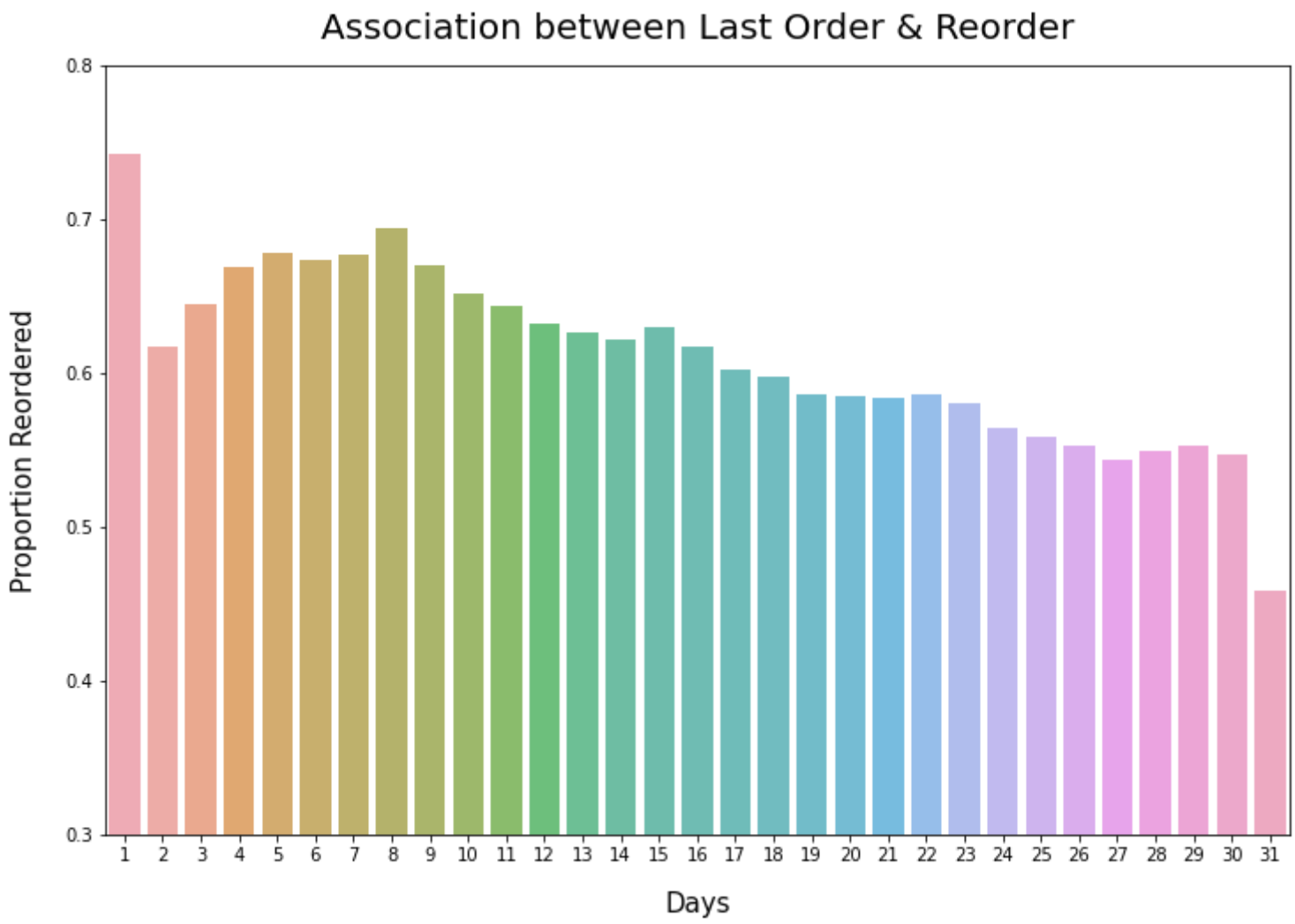
▼ 그 중 어떤 제품을 제일 처음 장바구니에 넣을까?



- 이를 통해, 사람들은 **응급피임약(Emergency Contraceptive)**을 자주 찾는다는 것을 알 수 있고, 이를 약 80%의 확률로 제일 처음에 카트에 넣는다.
- 그 이후로는 차(Tea)나 술(California Champagne, Flavored Vodka 등)을 많이 찾는다.

8. 연관성

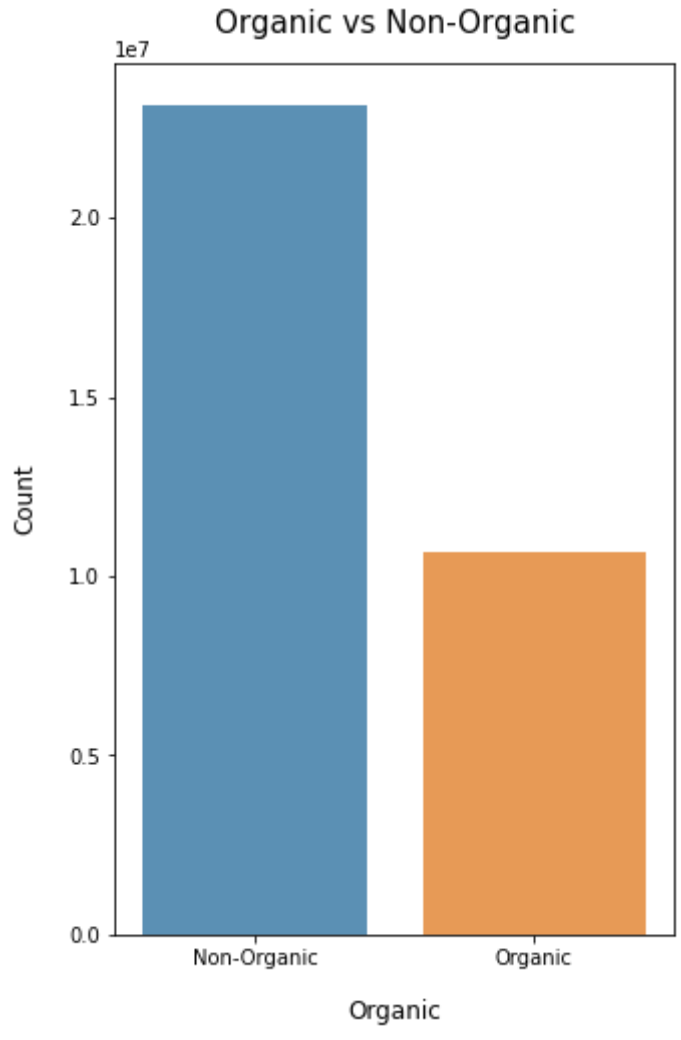
▼ 마지막 주문 날짜와 재주문



- 사람들이 같은 날에 다시 주문하면 같은 제품을 더 자주 주문한다는 것을 알 수 있다.
- 반면, 30일이 지나면 새로운 것을 순서대로 시도하는 경향이 있다.

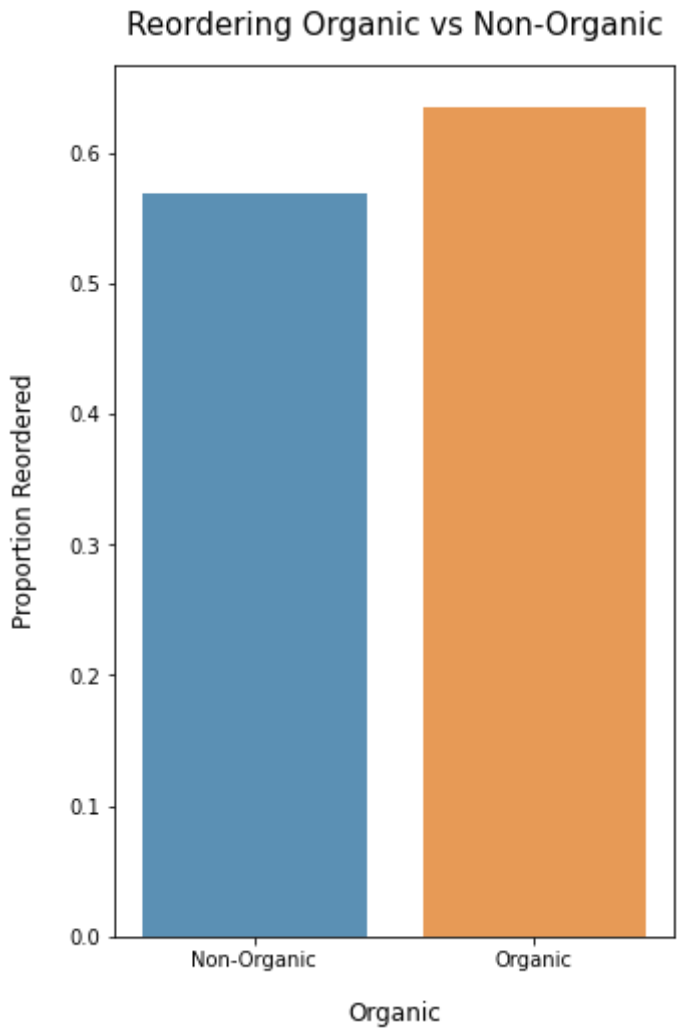
9. VS

▼ **Organic vs Non-organic**



- organic의 제품 수가 적으므로 당연히 organic의 주문 수가 적다.

▼ **Organic vs Non-organic 의 재주문**



- 하지만, organic 제품의 재주문율이 더 높다.

인사이트 도출 및 전략

- 위 EDA에서 파악한 분석을 종합해본다.

인사이트 도출

▼ **1. 수요 예측 분석 및 주요 특성 파악**

- **시간에 따른 수요 분석 및 특성 파악**
  - 주문율은 일주일 중 주말이 가장 높고, 그 중 10~14시 사이가 높다. 이는 시간이 가장 여유로운 주말을 이용한다는 것을 알 수 있다.
  - 재주문율은 일요일과 목요일이 높고, 6~8시 사이가 높다. 일요일은 일주일 동안 필요한 물품을 주문하는 것이고, 목요일은 금토일 주말에 필요한 물품을 주문하는 것임을 예측해볼 수 있다.
  - 재주문율은 한 달 시작 후 일주일 동안 상승하고 그 후 하락한다. 또한, 월 말일에 급격히 높아진다. 이를 통해, 사람들은 월 말에 익월에 필요한 필수 물품들을 구매하거나 해당 달이 시작되었을 때, 일주일 간 구매를 한다는 것을 알 수 있다.
- **수요에 따른 수요 분석 및 특성 파악**
  - 보통 한 번에 5개의 항목을 구매한다.
  - 주로 식료품의 수요가 제일 많고, 그 중에서 농산물, 유제품이 가장 수요가 많다.
  - 재주문 상품 중에 잠에 관련된 상품들이 상위권에 있다. (ex. Serenity Ultimate Extrema Overnight Pads, Simply Sleep Nighttime Sleep Aid)

▼ **2. 주요 상품 및 고객 구매 패턴에 따른 예측 및 추천**

- 사람들의 수요가 가장 많은 것은 식료품이고, 그 중 농산물, 유제품 등 인 것을 파악하였다.
- 그리고 그 중에서도 organic 제품들이 대다수의 판매 상위권을 차지하였다. (3번 EDA의 제품 그래프)
- 그 증거로 organic 제품의 재주문율을 보면 알 수 있다.

▼ **3. 고객 세그맨테이션을 통한 핸들링 전략 제안**

- 제일 처음에 장바구니에 추가된 제품이 나중에 추가된 제품에 비해 다시 주문될 가능성이 높다.
- 따라서, 사람들은 자주 사용하는 제품을 먼저 주문한 후에 새로운 제품을 찾는 경향을 가지고 있음을 알 수 있다.

- 그 중, 응급피임약을 자주 찾는다는 것을 알 수 있고, 그 다음으로는 차나 주류를 많이 찾는다.
- 이는 고객 분류군에서 연인을 생각해 볼 수 있다.

종합 전략

▼ 전략

- 주말에는 평일에 필요한 일상 물품을 노출시키고, 평일에는 주말에 필요한 휴식, 주류 물품을 노출시킨다.
- 식료품이나 재주문 상위의 상품들을 선정하고, 자체 제품을 개발하여 직접 판매를 해 이익을 올린다. (ex. 이마트의 'No Brand')
- 재주문율이 높은 제품들을 최대한 우선 입점시킨다. 이러한 제품들은 사람들이 자연스럽게 찾기 때문에 마케팅이나 광고는 필요가 없다. 하지만 할인을 하게 된다면 최우선 순위로 마케팅을 한다.
- 평상시에는 재주문율이 높은 제품들이 속해있는 카테고리의 다른 제품들을 앱/웹 메인이나 상단에 노출시킨다.
- 위의 분석을 통해 잠에 관련된 상품들이 상위권에 속해있는 것을 보아 사람들이 휴식이 필요한 것으로 보인다. 잠, 휴식에 관한 영양식품들을 연관 상품에 노출시킨다.
- 회원가입을 할 때 혼인 유무를 파악하여 관련된 상품을 노출시키거나, 연인과 관련된 이벤트를 연다.

회고 및 아쉬운 점

▼ 태블로

- 프로젝트 첫 날에 태블로에 관한 달콤한 말에 넘어가 처음 듣는 프로그램이었나 태블로를 공부했다.
- 무려 4일이라는 시간을 투자했지만 그만한 성과를 내지 못했다.
  - 데이터 연결에 시간을 많이 쏟았다.
    - 태블로에서 연결하는 것보다 SQL이나 pandas를 통해서 연결해서 하나의 파일을 태블로에 적재하는 것이 좋다.
  - SQL을 사용하여 데이터를 적재하지 못 했다. 그냥 csv 파일 자체를 사용했다.
  - 태블로 강의를 들었음에도 기능을 적절히 사용하지 못했다.
  - 처음에 태블로 퍼블릭에서 시작했고, 데이터를 연결하여 워크시트로 가는데 15,000,000개의 데이터까지만 가능했다. 옵션 1의 데이터는 34,000,000개로 불가능했다.
    - 그래서 결국 태블로 데스크탑 평가판을 이용했다.
  - 태블로는 집계에 집계가 안된다. (ex. max에 count가 안된다.) 그래서 방법을 찾느라 시간을 많이 소모했고, 결국 matplotlib과 병행하여 그래프를 그렸다.
  - 이제껏 태블로로 만든 소수의 그래프를 추출할 때, 태블로 테스크탑은 url로 공유할 수 없었고, 서버나 퍼블릭을 통해서 밖에 할 수 없었다. 하지만 서버는 비용을 지불했어야 했고, 퍼블릭은 데이터 개수에 한계가 있어 결국 이미지로만 내보내기를 할 수 밖에 없었다.
    - 그래프가 동작하거나, 대쉬보드를 전혀 사용할 수가 없었다.
- DA 트랙을 하면서 태블로로 끙끙댔는데 DE 트랙을 하는 동기들이 태블로를 사용하는 것을 보면서 가슴이 쓰라렸다. '쟤네는 왜 저걸 쓰는거지?' DA가 태블로를 사용하는 이유가 사라지면서 의욕을 상실했다.
- 여러 기업, 업계에서 태블로를 사용하는만큼 좀 더 공부를 하여 실력을 향상시켜야겠다고 생각했다.

▼ 모델 예측

- 태블로에 대한 시간 소모로 모델 예측은 손도 못 댔다.
- 다른 DA 트랙 동기들은 모델도 하고 했는데 나는 뭐 한 건가 싶다. 미래가 없다.

▼ 전략의 한계

- 모델 예측을 하지 못 해 제대로 된 분석이 나오지 않았다.
- 시도도 하지 않았던 이커머스 부분이라 없는 도메인 지식을 끌어와 생각을 하느라 전략이 제대로 짜여지지도 않았고 (보면 누구나 생각할 수 있는 그런 전략들이다. 사실 전략이라고 말하기도 뭐하다.), 뭘 해야 할지 한계가 있었다.
- 시간이 부족해 이커머스에 대한 공부나 정보를 알아볼 수 없었다. 너무 그래프 그리는 것에만 급급했다.

▼ 실력

- 태블로는 물론이거니와, 파이썬을 통해 그래프를 그리는 것도 쉽지 않아, 동기의 도움을 많이 받았다.
- EDA에 관한 주제들은 100% 내 아이디어가 아니다. 대부분 인터넷에서 가져왔다.

- 실력이 다른 동기들에 비해 너무 많이 떨어진다. 내가 기수 중에서 꼴등이 아닐까 싶다.

▼ 총 평

- 점수 : 2/10
- 지금까지의 프로젝트들은 자신이 있었고, 만족했지만 이번 프로젝트는 망함 그 자체다.
- 과연 내가, DE나 DS를 할 수 있을까라는 의문이 들었다.
- 취직 못 할 것 같다. 가망이 없다.