Instacart customer Analysis
프로젝트 보고서

2팀 **2**조
한상우 김다혜 이지석 송재현

📌주제**: instacart customer** 분석

📌데이터 셋 **: Instacart Market Basket Analysis**

https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis/data

📌데이터 선정이유

팀원이 한개씩 데이터셋을 찾아왔고 그중 데이터 양이 많고 목적성 설정을 명확하게 할 수 있으며 데이터가 나눠져 있어서 연습적인 측면에서도 나쁘지 않다고 판단한 데이터 셋을 선정함.

📌참여자 역할 및 기여도**(%)**

참여자 이름	역할	기여도**(%)**
한상우	재구매를 많이 하는 사용자의 특징 분석 및 결과 도출,프로젝트 발표	25
김다혜	재구매를 많이 하는 사용자의 특징 분석 및 결과 도출,보고서 최종 수정	25
이지석	이탈 고객 분석 및 결과 도출,소개문서 작성(notion), 깃허브 업로드	25
송재현	이탈 고객 분석 및 결과 도출,프로젝트 발표	25

📌활용 기술 및 프레임워크

- Python(pandas,numpy,matplotlib)
- ChatGPT
- Colab

📌**instacart** 소개

온라인을 기반으로 식료품 배송 서비스를 제공하는 미국 기업
고객이 Instacart를 통해 주변 슈퍼마켓 및 식료품 매장의 식료품을 주문을 하면, Shopper가 대신 장을 봐준 뒤 배송을 시작한다.

📌컬럼특징 **(csv** 파일별**)**

csv file name	col name	explanation
aisle.csv	**aisle_id**	상품 종류의 ID
	aisle	상품 종류의 이름
department.csv	**department_id**	부서 ID
	department	부서 이름
order_products__prior.csv, order_products__train.csv	**order_id**	주문의 고유 ID
	product_id	상품의 ID
	add_to_cart_order	하나의 주문에 상품이 담긴 순서 (int)
	reordered	재주문 여부 (bool)
orders.csv	**order_id**	주문의 고유 ID
	user_id	사용자의 고유 ID
	eval_set	데이터셋의 유형, ['prior','train','test']로 나누어짐
	order_number	각 주문 내에서 사용자의 상품이 담긴 순서
	order_dow	주문이 이루어진 요일, 0~6의 정수가 일요일 ~ 토요일과 1대 1 매칭된다.
	order_hour_of_day	주문이 이루어진 시간, 0~23 사이의 정수로 표현
	days_since_prior_order	이전 주문부터 현재 주문까지의 일수. 첫 주문인 경우에 결측치로 표현.
product.csv	**product_id**	상품의 고유한 ID
	product_name	상품의 이름
	aisle_id	상품 종류의 ID
	department_id	부서 ID

📌프로젝트 세부결과

📋 분석방향 및 문제 정의

- 해결해야 하는 문제 : 매출 상승
- 매출증가를 위해 크게 두 종류의 분석을 진행 해보기로 하였다.
    A) 이탈 고객 분석 : 이탈 고객과 이탈하지 않은 고객의 특성을 분석하여 비교
    B) 재구매를 많이 하는 사용자의 특징 분석 : 재구매를 많이 하는 사용자와, 그렇지 않은 사용자의 특성을 분석 비교

📋 결론

A) 이탈 고객 분석 : 이탈 고객은 장바구니 사용률이 높고 한번에 대량 구매를 하는 경향이 있어 구매주기가 긴 것으로 확인되었다. 이러한 특징을 보았을때, 배송비에 부담을 느껴 대량구매를 한다고 결론 지었다. 문제 해결 방안으로 정기적인 배송 서비스를 할 시에 배송비 할인을 해주는 정책, 주기적인 배송비 할인 이벤트 등 을 고민해볼 수 있다.

B) 재구매를 많이 하는 사용자의 특징을 분석

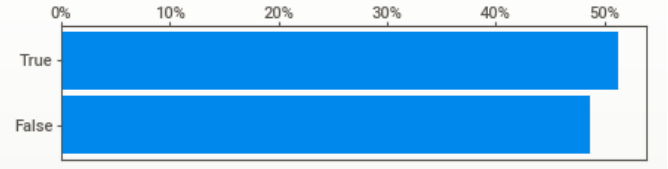
- milk와 egg, yogurt와 같은 카테고리가 재주문률이 높았다. 카테고리 특성을 이용한 정기배송 프로모션을 진행하면 매출 증가에 도움이 될 것으로 보인다.
- 하위 5%에게 마케팅을 한다고 가정한다면, morning segment에 해당하는 시간대에 진행하는 것이 유리할 것이다.
- 상위 5% 사용자를 대상으로 마케팅할때는 주말보다는 주중에 하는것이 매출상승에 도움이 된다.
- 하위 5%의 사용자들은 instacart를 사용하는 목적성이 “건강하고 신선한” 상품을 구매하기 보다는, “장보기”에 더 가깝다고 분석
- 오전 9시부터 오후 5시까지 주문량이 증가하므로 해당 시간대에 더 프로모션등을 진행하여 더 많은 주문을 유도할 수 있음

---

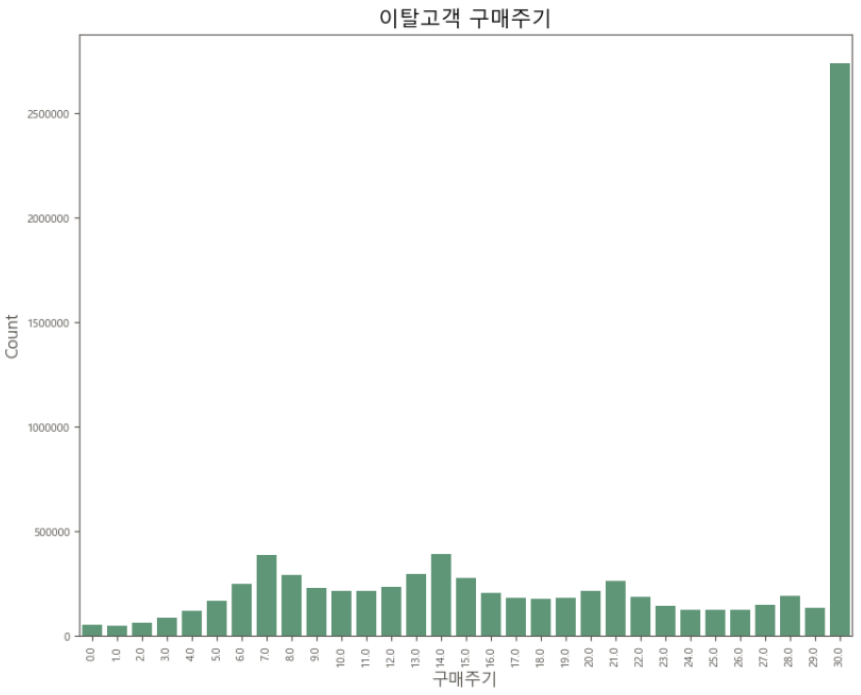
1. 이탈 고객 분석

이탈률이 높은 집단을 선별하여, 고객 이탈에 대한 몇 가지 가설 검증을 통해 이탈 방지에 도움이 될 수 있을 지 파악하고자 한다. 먼저 이탈 고객은 구매주기가 **15**일이 넘는 고객이며, 구매주기의 기준(15일)은 유저 별 평균 구매 주기를 구한 후에 중앙값을 사용하였다. 이탈률은 51.3%가 나왔으며 총 206,209명 중에 105,815명으로 집계되었다.

- 이탈고객 지표
    - user_id 별 구매주기의 평균값추출하여 15일 기준 지정
    - (유저 별 구매주기 평균> 15) - 이탈 고객으로 간주



True - 이탈고객



구매주기는
최대 30일까지이며,
이탈고객 기준
구매주기가 30일인
고객이 28%를 차지했다.

- 이탈고객 가설

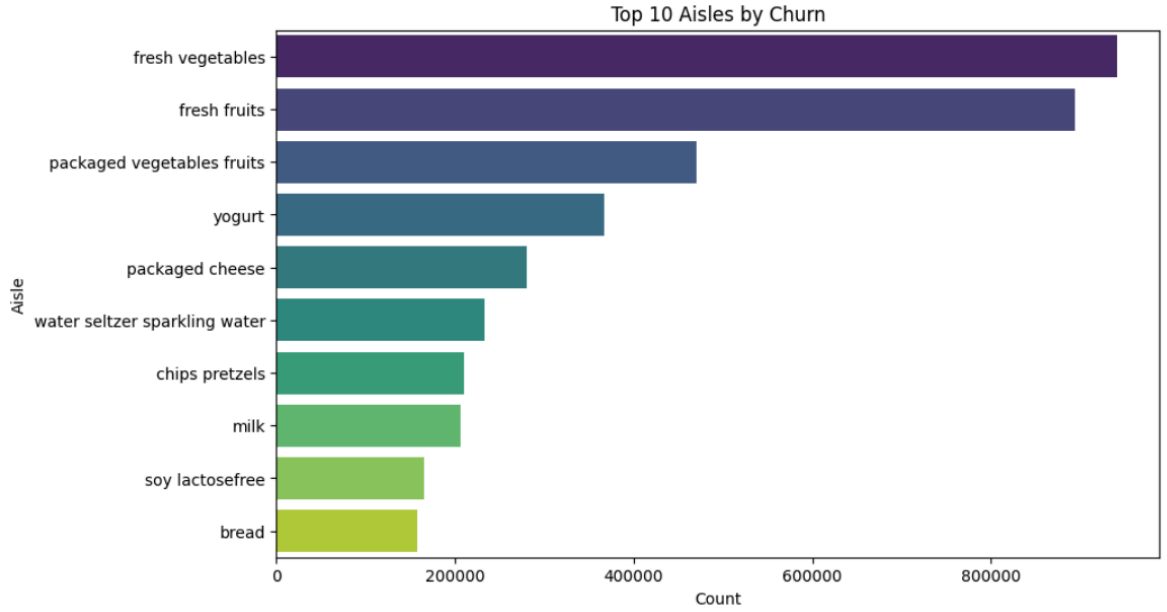
1. 이탈고객은 유통기한이 긴 제품의 구매가 많다.
    a. 제품의 유통기한은 이탈 고객 수치에 영향을 주지 않는다
2. 이탈고객이 발생한다면 첫 구매이후가 가장 많다.
    a. 고객의 구매횟수와 고객의 이탈은 비례하지 않는다.
3. 이탈 고객은 비이탈고객보다 새벽시간 구매비율이 높다.
    a. 이탈 고객, 비 이탈 고객의 구매시간대는 큰 차이가 없다. 이탈 고객은 오후 시간대, 비 이탈 고객은 주로 오전 시간에 구매하는 경향이 있다.
4. 장바구니의 사용률이 낮은 고객은 이탈고객이 될 확률이 높다.
    a. 이탈 고객는 한번에 대량 구매를 하기 때문에 구매주기가 긴 것으로 확인. 원인을 배송비라고 여기고 그에 따른 마케팅 전략 수립
5. 특정 aisle에서 이탈 고객이 많이 발생한다.
    a. baby, bulk 등 특정 aisle에서 이탈 고객과 비 이탈 고객 간의 차이가 많이 났다. 가족 구성원의 수와 관련되었다고 생각하였으나, user data가 없는 관계로 더 이상의 분석은 불가능했다.

- 가설 검증

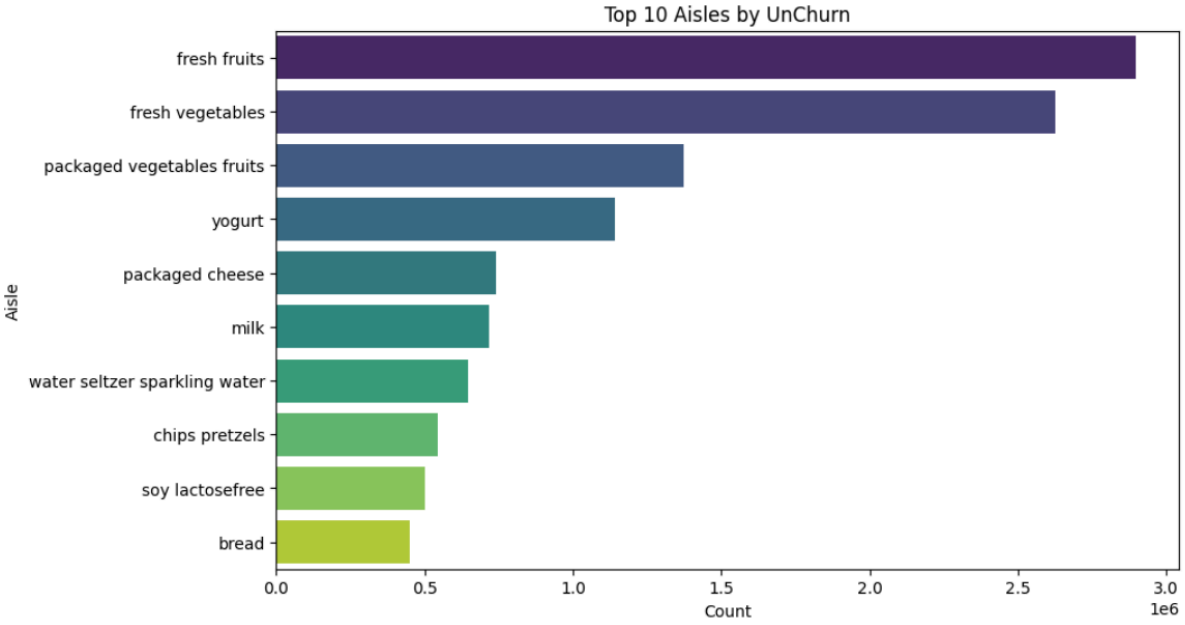
1) 이탈고객은 유통기한이 긴 제품의 구매가 많다**.**

➢ 유통기한이 긴 제품을 주로 구매하는 고객이라면 구매주기가 길어져 이탈고객으로 포함될 것이라는 가설을 세웠다. 이 고객층의 구매주기가 15일을 넘는다면 이탈률 영향을 줄 가능성이 있다.

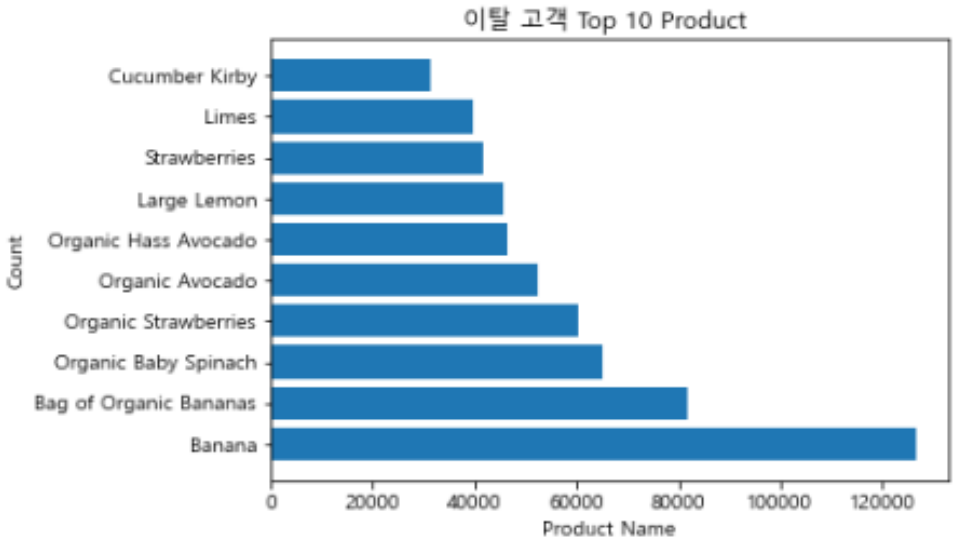
➢ 유통기한과 이탈률의 관계를 알아보기위해 이탈고객을 기준으로 하여 제품 판매량 분석을 진행하였다.



<이탈 고객들이 많이 구매한 aisle 순위>



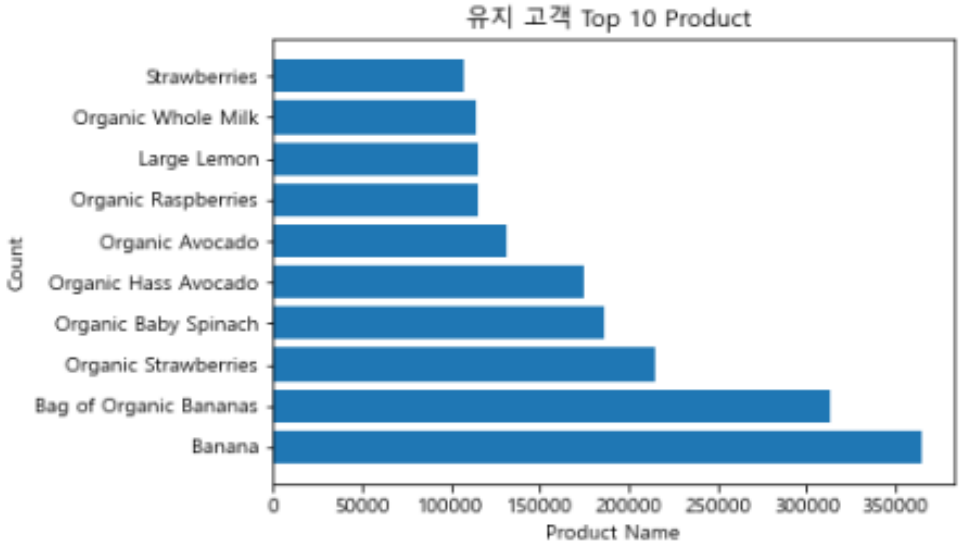
<비 이탈 고객들이 많이 구매한 aisle 순위>



<이탈 고객들이 많이 구매한 Product 순위>

➢ 이탈 고객을 대상으로 aisle(진열대), product 판매량 분석을 진행하였다.

➢ aisle에서는 채소와 과일에 순위가 높았고, product 또한 상위 10개 항목 모두 채소와 과일이었다.

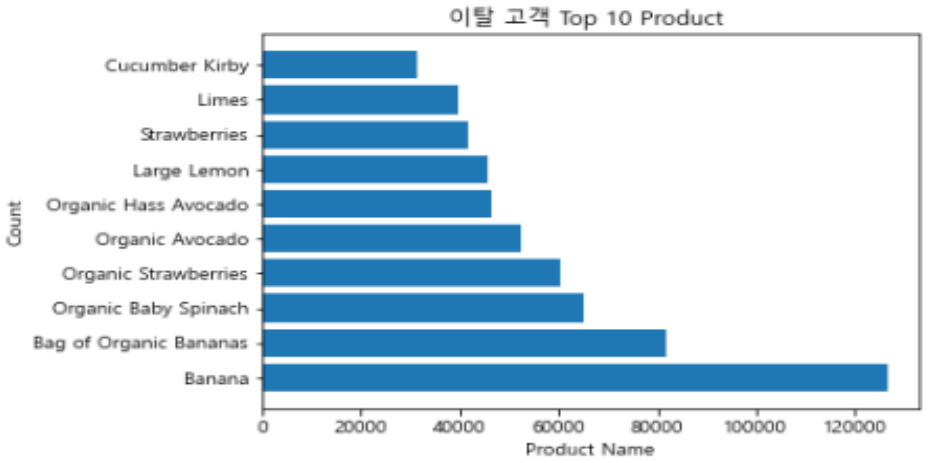


<비 이탈 고객들의 판매 순위>

➢ 비 이탈 고객 또한 과일, 채소 구매가 많이 이루어진 것을 확인 할 수 있었다.

➢ 품목 별 판매량을 확인했을 때, 유통기한이 긴 제품의 결과를 찾아 볼 수 없었다.

➢ 이탈 고객과 비 이탈 고객이 선호하는 품목도 크게 다르지 않았다.



- 확실한 비교를 위해 이탈주기가 30일(최댓값)인 그룹으로 나누어 품목을 비교해 보았다.
- Banana, Bag of organic Bananas, Organic Baby Spinach의 구매가 많이 발생한 것을 알 수 있다.

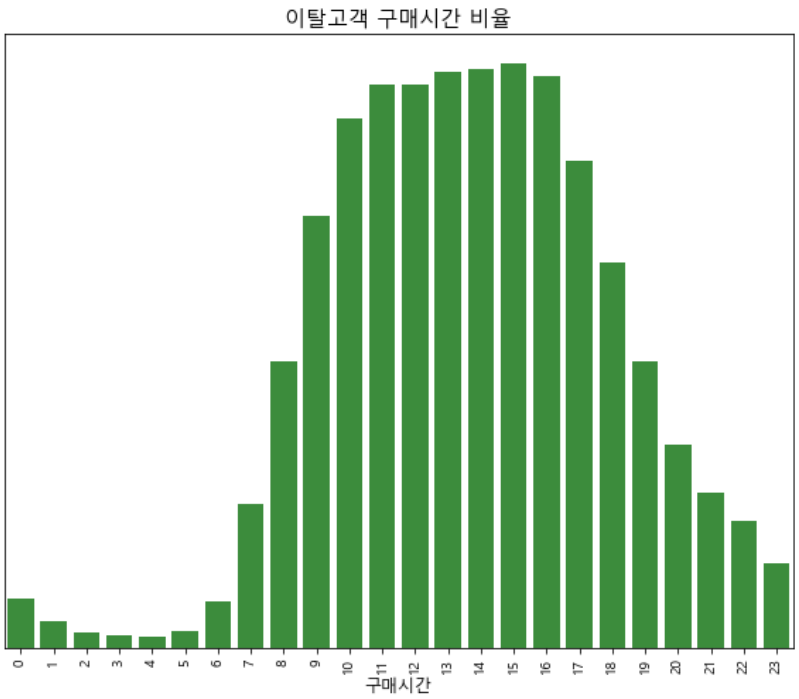
  💡이탈, 비 이탈 고객 제품 구매량 비교 결과
- 이탈 고객과 비 이탈 고객의 제품 선호차이는 없다.두 그룹 모두 fresh fruits, fresh vegetables를 선호한다.
- 두 그룹 모두 Banana, Bag of organic Bananas, Organic Baby Spinach, Organic Strawberries 의 구매가 많다
- 제품의 유통기한과 고객 이탈의 관계는 없다.

2) 첫 번째 구매 이후 이탈 고객이 가장 많이 발생할 것이다.

첫 구매이후 재주문을 한 고객은 품질이나 서비스에 만족한 고객이며, 반대로 이탈고객이 생긴다면 첫 구매 이후에 가장 많이 생길 것으로 생각했다. 첫 구매 이후 재주문에 대한 분석을 해본 결과, 해당 데이터에 있는 모든 고객이 재주문을 한 것으로 확인되었다.

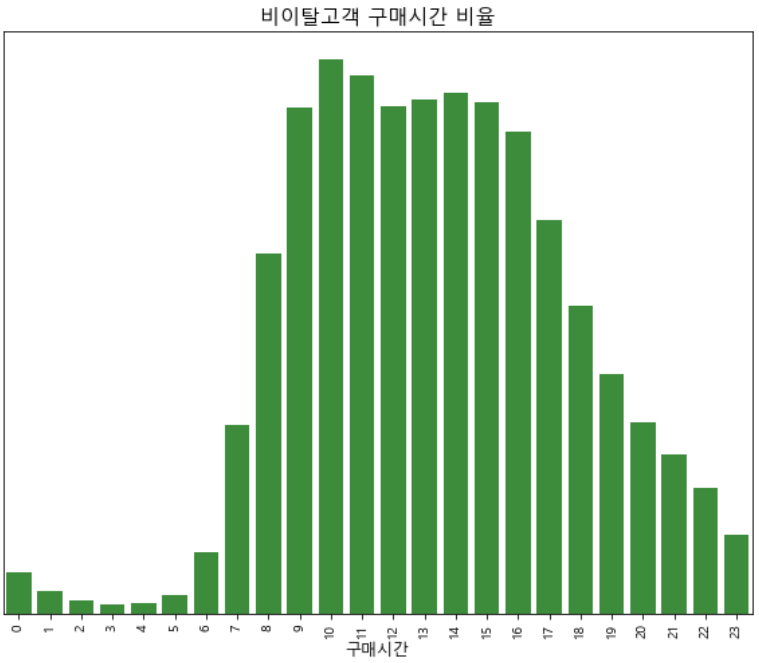
3) 이탈 고객은 비이탈고객보다 새벽시간 구매비율이 높아 이탈률에 영향을 줌**.**

- 주로 오프라인 구매를 선호하는 고객이라고 가정했을때, 주변 마트가 닫은 늦은 시간에만 구매한 경우가 있을 수 있다. 이 고객층은 사용 빈도가 낮아 이탈 고객일 것이다라는 가설을 세웠다.
- 가설을 토대로 한다면 새벽 시간구매비율 **(**이탈 고객 **>** 비 이탈 고객**)**
- 검증을 위해 고객 구매시간 데이터를 분석해보았다.
- 오프라인 구매가 힘든 시간대는 구매 데이터가 크게 감소하는 **0**시**~6**시로 기준을 잡았다.



➢ 고객이 가장 활발하게 구매하는 시간대는 **13~16**시이다.

➢ 이탈고객 0시~6시의 구매비율은 **0.19%**을 차지하였다.



➢ 비이탈고객 0시~6시의 구매비율은 **0.2%**을 차지하였다.

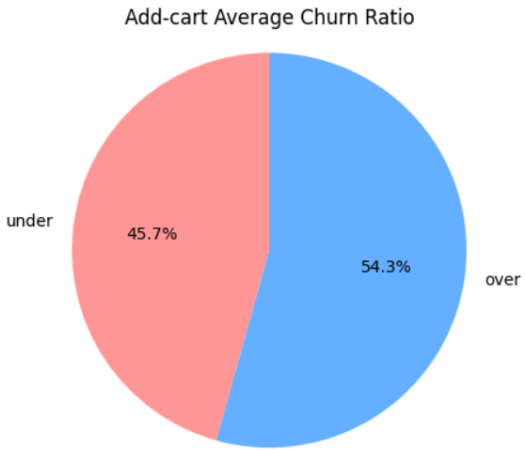
➢ 새벽 구매시간 비율만 비교해보았을 때 이탈고객은 **0.19%,** 비 이탈 고객은 **0.2%**로 비 이탈 고객이 더 높았다**.**

➢ 이탈 고객과 비 이탈 고객의 구매시간을 t검정 수행하였을 때, p_value 값은 0에 수렴하여 귀무가설을 기각하였다.

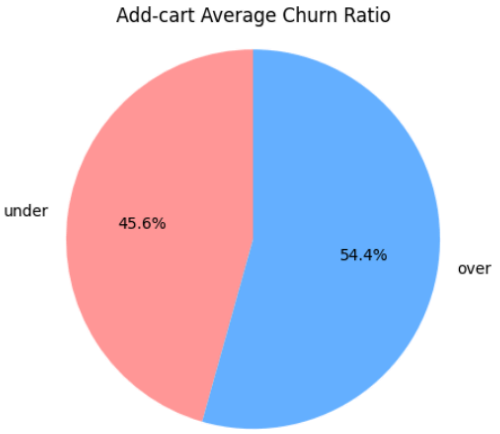
- ➢ 시간대 특징으로는 이탈 고객은 오전 시간보다 오후 시간 구매가 많았고, 비 이탈 고객은 오전시간에 더 많은 구매가 발생했다.
- ➢ 고객의 구매시간을 파악하여 맞춤 알림 서비스를 진행하는 것이 매출에 도움이 될 수 있다.

4) 장바구니 이용이 평균보다 낮은 고객의 이탈률이 높다

장바구니 이용에 따른 이탈률을 검증해보고자 한다. 장바구니에 상품을 덜 담는 사용자일수록 이탈 고객일 확률이 높다고 가설을 세웠다. 그래서 이탈 고객들의 평균 장바구니 물품 개수를 구해보았다. 평균 장바구니 개수인 **10**개를 기준으로**,** 평균 이하와 평균 이상으로 나누었다**.** 먼저 이탈 기준을 평균 구매주기 값인 15일로 구했을 때 비율 차이이다.

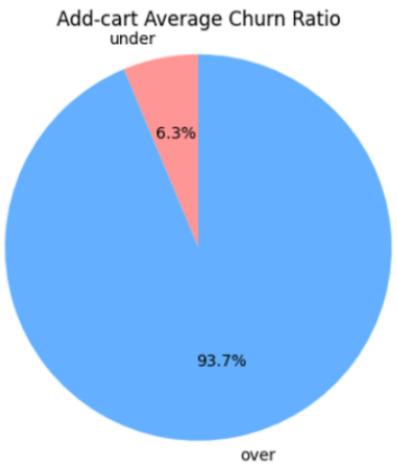


<이탈 고객 장바구니 개수 평균 이하/이상 비율>

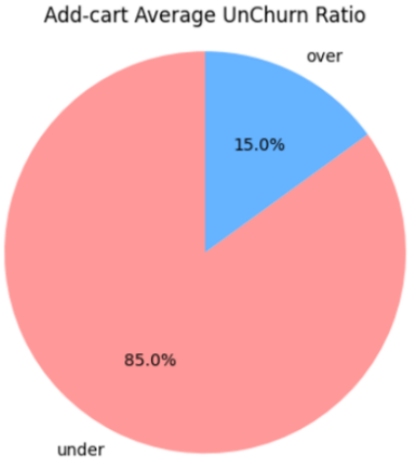


<비이탈 고객 장바구니 개수 평균 이하/이상 비율>

평균으로 구했을 땐, 비이탈 고객와 이탈 고객 간의 장바구니 이용률이 크게 다르지 않았다. 하지만 이탈 주기를 27일로 설정했을 때는 달랐다. 이탈률은 11% 정도로 집계되었으며, 장바구니 이용률에서 유의미한 차이를 보였다.



<이탈 고객 장바구니 개수 평균 이하/이상 비율>



<비이탈 고객 장바구니 개수 평균 이하/이상 비율>

이렇듯 이탈 고객의 **93%** 정도가 평균 이상이고**,** 비이탈 고객의 **85%** 정도가 평균 이하임을 보였다**.**

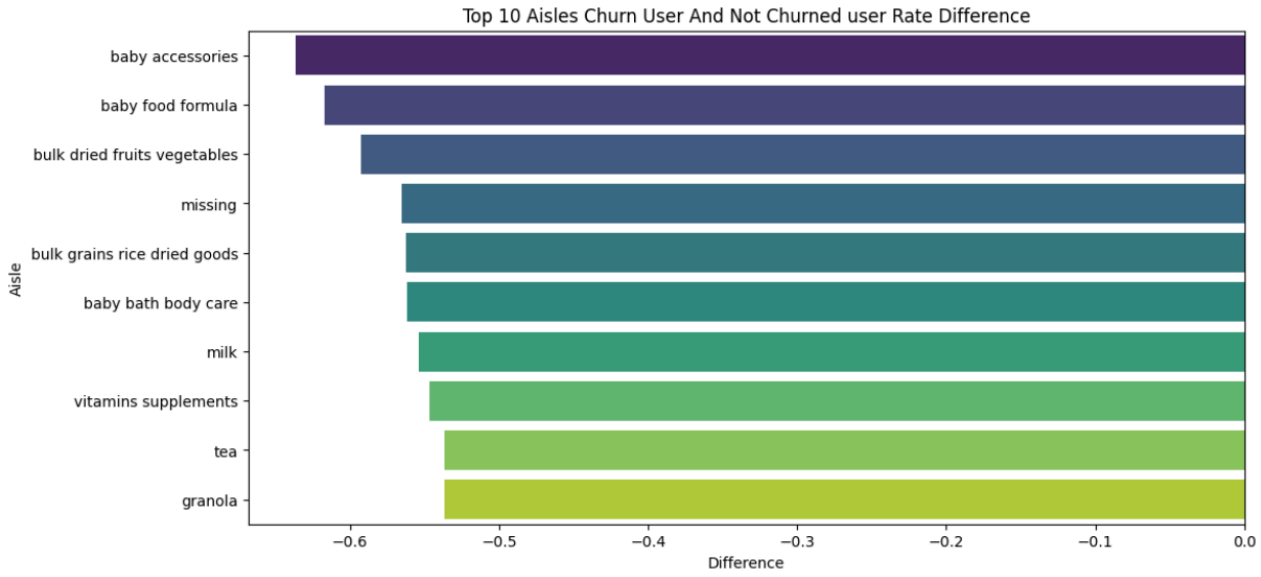
이러한 차이는 이탈 주기가 평균 값인 15일부터 시작해 30일까지 갈수록 두드러지게 나타났다.

결론적으로 이탈 고객는 한번에 대량 구매를 하기 때문에 구매주기가 늘어남을 알 수 있었다**.**

이를 통해 우리는 배송비가 비싸기 때문에 한번에 많이 구매하는 고객이라고 여겼고, 그에 따른 마케팅 전략을 세워보았다. 전략은 **1)** 정기적인 배송 서비스를 할 시에 배송비 할인을 해주는 정책 **2)** 주기적인 배송비 할인 이벤트 등 두 가지를 세울 수 있었다.

5) 특정 **aisle**에서 이탈 고객이 많이 발생한다**.**

특정 aisle에서 이탈 고객들과 비 이탈 고객들 간의 구매 비율이 차이가 나는지 확인한다. 차이가 큰 순서대로 10개를 정렬해봤을 때, baby나 bulk 등 품목이 많은 것으로 보여진다. 가족 구성원의 수의 차이가 아닐까, 생각할 수도 있지만 user 관련 data 부족으로 더 이상 분석을 진행할 수 없었다.



<이탈 고객/비이탈 고객 aisle 구매 비율이 차이나는 10개>

---

2. 재구매를 많이 하는 사용자의 특징 분석

[분석 전 지표 설정 및 분석 해보고자 하는 내용 설정**]**

- ➢ 첫 주문 이후 재주문 비율
  - 분석결과 해당 데이터는 모든 사용자가 재주문을 하였기 때문에 의미가 없음
- ➢ 사용자별 첫 주문 이후 재주문 여부 확인 (위와 동일)
- ➢ 재주문에 걸리는 평균 시간 (약 15일)
- ➢ 재주문을 많이 하는 사람을 파악 (지표설정)
  - 재주문 비율 지표: 주문시마다 다시 구매하는 상품이 몇개인지의 비율로 정함
    - order_id와 해당 주문의 reordered 비율 (order_id별 reordered 갯수)
    - order_id를 그룹화하고 각 order_id별로 reordered의 평균을 계산
  - 높은 재주문률 : 재주문율 >= 임계치(70-80%)

Reorder Rate:
Calculate the reorder rate for each user, defined as the ratio of the number of orders with reordered items to the total number of orders for that user.
Reorder Rate=Number of Orders with Reordered ItemsTotal Number of Orders
Reorder Rate=
Total Number of Orders / Number of Orders with Reordered Items

You can set a threshold (e.g., 70%, 80%) to identify users with a high reorder rate.
High Reorder Rate=Reorder Rate≥Threshold
High Reorder Rate=Reorder Rate≥Threshold
This threshold will help you classify users who reorder a significant portion of their items.

➢ 상위 5%와 하위 5% 사용자의 기준을 정하기 위한 **metric** 설정

- 분석하는 데이터에서는 모든 사용자가 재주문을 했기 때문에 한 **user**가 하루에 몇번의 구매를 했는지(orders_per_day)를 metric으로 삼았다,
- orders_per_day를 계산한 방법은 다음과 같다.
    1) orders.csv에서 unique한 user_id 별로 그룹을 만들어 준다.
    2) 하나의 그룹(user_id 하나)에 대해 unique한 “order_id”를 count 하여 이를 “total_orders”라는 col로 설정한다
    3) 하나의 그룹(user_id 하나)에 대해 모든 “days_since_prior_order” 값을 더하여 “total_days”라는 col로 설정한다
    4) “total_orders” / “total_days”를 orders_per_day로 설정한다. 이때, divide by zero를 해결하기 위해 모든 “total_days”의 값에 1을 더해준다.
- **orders_per_day**값이 **"95-th percentile"** 이상인 사용자들을 재주문을 많이 한다고 정의하고**, “5-th percentile”**에 해당하는 사용자들을 재주문을 많이 하지 않는 사용자로 정의한다. 이후로는 각각 **“**상위 **5%”, “**하위 **5%”**라는 표현으로 두 그룹을 칭하도록 한다.
- orders_per_day를 metric으로 잡으면 ”total_days” 값이 작은 사용자들에게 유리하다. 하지만, “total_days”값이 작은 사용자의 경우 표본이 충분하지 않을 가능성이 높다.

user_id	total_orders	total_days	orders_per_day
164320	4	1	4
109010	4	1	4
80567	4	1	4
201321	4	1	4
115420	4	2	2
71794	4	2	2
137150	4	2	2
138757	4	2	2
179078	4	2	2
99339	4	2	2
128483	4	2	2
63845	4	2	2
199088	4	3	1.333333333
60546	4	3	1.333333333
151890	4	3	1.333333333
95523	4	3	1.333333333

위 문제와 관련하여 total_days 값이 작은 값을 가진 아웃라이어들을 제거하기로 결정했다.

IQR method (계수 = 0.4)로 설정하여 outlier를 제거한 결과 통계량의 변화는 다음과 같다.

통계량	**outlier** 제거 전	**outlier** 제거 후
표준편차	102.132	79.633231
왜도	0.16806761	0.47189494
IQR	167	127

왜도가 증가한 모습이 보이지만, 표준편차와 IQR값은 감소하였다.

무엇보다, outlier를 제거하기 전 total_days의 range는 (1~19)였는데, outlier 제거 후 range는 (16~315)로 고르게 분포하게 되었다. 즉, **outlier**를 제거하여 데이터의 다양성을 늘릴 수 있었으며, 표본의 수가 더 많은 사용자를 위주로 분석을 진행할 수 있다.

💡 결론: 재구매율이 높은 상위 **5%**에 해당하는 사용자와 하위 **5%** 사용자의 구매 동향을 파악후 비교분석해보자.

1.상품 카테고리별 분석

2. product loyalty 별 분석

3. 시간적 특성에 따른 분석

---

1. 상품 카테고리별 분석

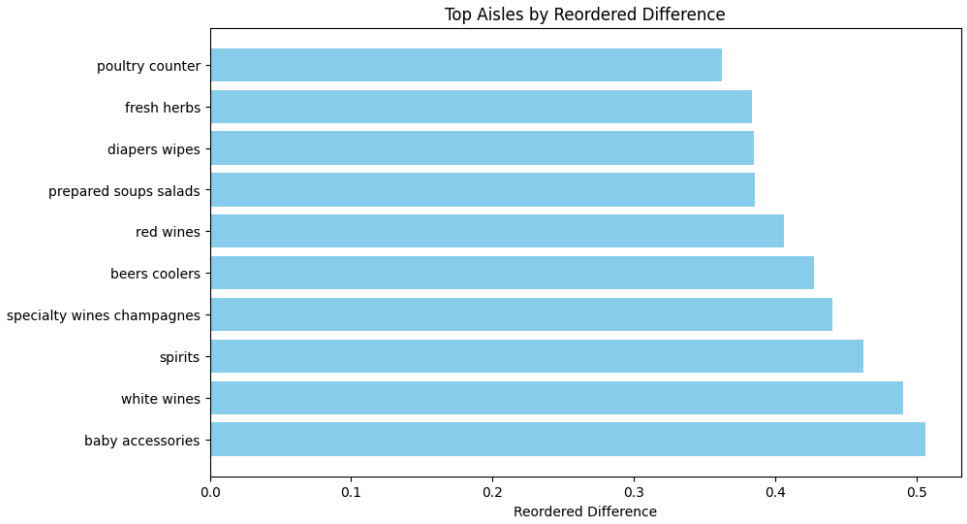
- 각 카테고리에서 재주문율 확인 후, 어떤 카테고리 상품이 더 자주 재주문 되는지 파악한다. 이러한 분석을 통해 향후 유저가 특정 카테고리에서 자주 재주문할때 쿠폰을 지급하여 주문량을 유지하게 하고, 많은 양을 구매하게끔 유도하여 매출을 증가시킬 수 있다고 판단함.

	aisle_id	reordered	aisle
0	84.0	0.851581	milk
1	115.0	0.815373	water seltzer sparkling water
2	24.0	0.797062	fresh fruits
3	32.0	0.794163	packaged produce
4	86.0	0.788359	eggs
5	62.0	0.785342	white wines
6	53.0	0.779012	cream
7	91.0	0.776398	soy lactosefree
8	120.0	0.765695	yogurt
9	124.0	0.754730	spirits

<상위 5%의 재구매율이 높은 상품 카테고리>

- milk와 egg, yogurt와 같은 카테고리가 재주문률이 높았다. 카테고리 특성을 이용한 정기배송 프로모션을 진행하면 매출 증가에 도움이 될 것으로 보인다.

- 더 유의미한 차이를 찾기 위해 두 그룹의 재주문율의 차이를 구해 보았다
    - (상위 5%의 카테고리별 재주문율 - 하위 5%의 카테고리별 재주문)



A) “baby accessories”, “diapers wipes”와 같은 아이와 관련된 상품들이 높은 차이를 보이며,
B) “white wines”, “spirits”, “specialty wines champagnes”, “beers, coolers”, “red wines”와 같은 술 종류의 상품들이 많은 모습을 살펴볼 수 있다.

A)의 정보를 통해 상위 **5%**의 사용자들은 가족 구성원 중 아이가 있을 가능성이 높다고 생각해 볼 수 있다. 물론, 사용자의 가족 구성원에 대한 정보는 없기 때문에 이를 증명하는 것은 어렵다.

위 가설과 B)의 정보, 그리고 고전적인 가설(기저귀를 사는 사람은 맥주도 같이 산다)를 묶어 새로운 가설을 하나 만들어보자. “상위 **5%**의 구매자들이 아이 용품을 구매할 때 술 종류의 상품도 많이 구매할 것이다”

해당 가설에 대해서 확인해 보기 위해 아이용품을 구매한 이력이 있는 사용자의 주문과, 아이용품과 주류를 동시에 포함한 주문의 개수를 계산해 보기로 했다. 우선, “아이 관련 용품”과 “주류”의 기준은 다음과 같이 설정하였다.

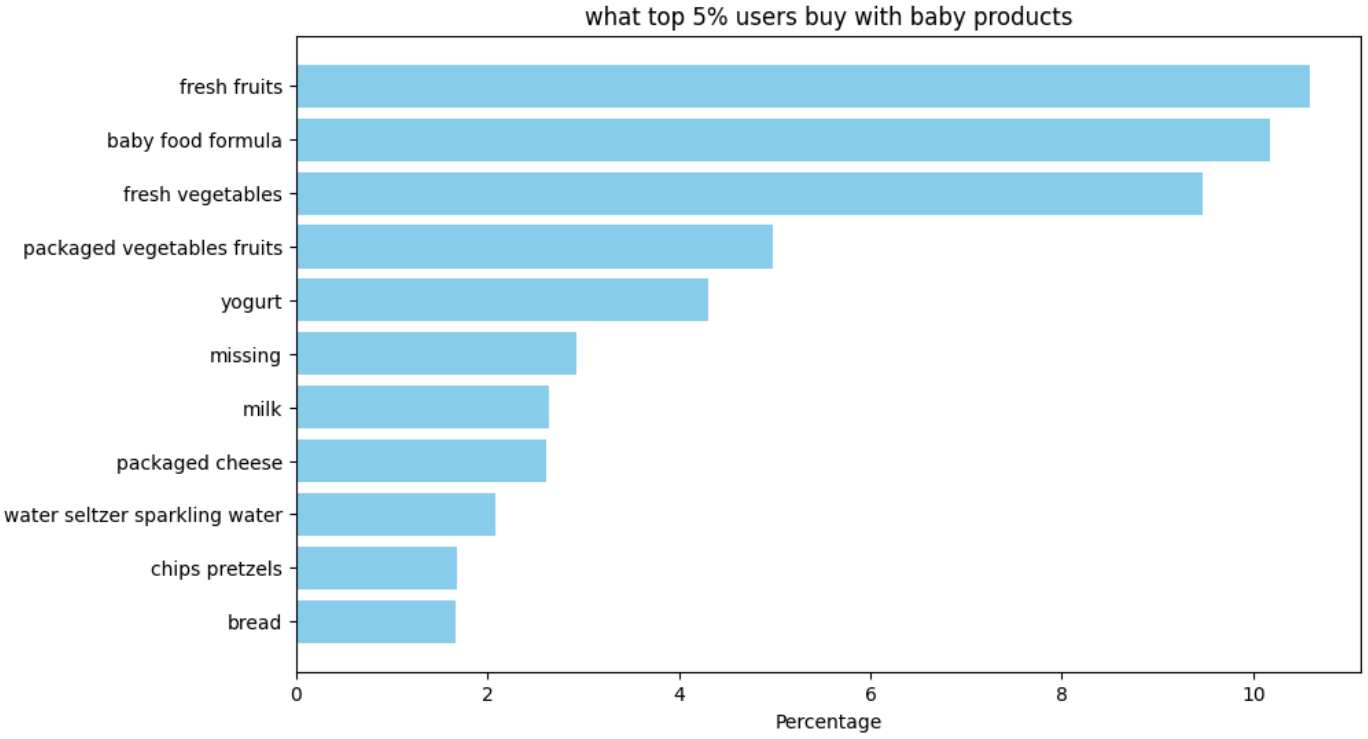
category	aisle	aisle_id
아이 관련 용품	56	diapers wipes
	82	baby accessories
	92	baby food formula
	100	infant needs
	102	baby bath body care
주류	27	beers coolers
	28	red wines
	62	white wines
	124	spirits
	134	specialty wines champagnes

➢ 아이 관련 용품”이 등장한 주문의 개수는 총 17,685개 였다.

➢ “아이 관련 용품”과 “주류”가 동시에 등장한 주문의 개수는 총 335개 였다.

➢ 결과적으로, 1.9%의 주문만 가설을 만족한다. 결국 상위 **5%**의 구매자들은 아이 용품을 구매할 때 술 종류의 상품도 같이 구매할 것이다라는 가설은 옳지 않다**.**

다음으로 상위 5%의 구매자들이 아이 용품을 구매할 때 같이 구매하는 상품의 종류를 찾아보았다.



➢ instacart에서 사용자들이 많이 구매하는 상품의 점유율과 크게 다르지 않은 양상이다. 즉, 해당 정보로 **“**아이 관련 용품**”**을 구매하는 사용자들에 대한 유의미한 **insight**를 이끌어내기는 어려워 보인다 (상위 10개의 점유율에 대한 분석뿐만 아니라, 더 낮은 점유율에서도 비슷한 양상을 보였다).

2. **product loyalty** 별 분석

	product_id	reorder_count	product_name
10135	13176	31703	Bag of Organic Bananas
19217	24852	26868	Banana
16319	21137	17796	Organic Strawberries
36515	47209	15521	Organic Hass Avocado
16920	21903	14296	Organic Baby Spinach
21633	27966	11756	Organic Raspberries
21535	27845	10010	Organic Whole Milk
36941	47766	9933	Organic Avocado
36830	47626	7881	Large Lemon
20264	26209	7460	Limes
12964	16797	7280	Strawberries
17731	22935	7279	Organic Yellow Onion
34793	45007	7189	Organic Zucchini
19303	24964	6207	Organic Garlic

<상위 5%_ 가장 많이 재주문되는 제품 >

	product_id	reorder_count	product_name
14104	24852	4359	Banana
7466	13176	2734	Bag of Organic Bananas
12470	21903	2071	Organic Baby Spinach
12034	21137	1437	Organic Strawberries
27064	47766	1404	Organic Avocado
125	196	1216	Soda
26986	47626	1203	Large Lemon
9561	16797	1119	Strawberries
26757	47209	1116	Organic Hass Avocado
14883	26209	953	Limes
6990	12341	892	Hass Avocados
25305	44632	834	Sparkling Water Grapefruit
11203	19660	833	Spring Water

<하위 5%_가장 많이 재주문되는 제품 >

➢ 재구매를 많이 하는 상품 중 상위에 있는 품목들을 보면 대부분 과일이나 채소 등 department == produce 인 것들이다.

➢ 하위 5%에서 재구매를 많이 하는 상품 중 상위에 있는 품목들 또한 과일이나 채소 등으로 상위 5%와 별 다른 차이가 없음을 알 수 있다.

- 사용자의 첫 번째 주문과 그 이후 주문 간의 시간 간격 계산하여 오랜기간 주문을 유지하는 고객의 충성도를 알 수 있다.

	user_id	days_since_prior_order
6116	139506	314.0
7831	179260	314.0
3990	90510	314.0
8991	205185	314.0
1604	36481	314.0
6411	145974	314.0
7368	168176	314.0
1748	39021	314.0
5297	121231	314.0
8759	200316	314.0

<상위 5% 유저들의 주문 유지 기간>

	user_id	days_since_prior_order
674	13469	314.0
5542	113676	313.0
7685	157775	313.0
2457	51104	313.0
3354	68773	313.0
2519	52524	312.0
4320	89001	311.0
5896	120894	311.0
8579	177188	310.0
458	8963	309.0
1240	25091	309.0
5481	112556	308.0
7363	150645	307.0
1736	35379	307.0
7036	143980	307.0
1179	23770	307.0
9214	189719	306.0
8486	174976	306.0
4700	96694	306.0
4189	86659	306.0
5868	120208	305.0

<하위 5% 유저들의 주문 유지 기간>

➢ 상위 **5%**와 하위 **5%**의 충성도가 별 차이가 없는 것을 발견했다.

이를통해 상위 **5%**의 유저라고 해서 충성도가 높은건 아닌것을 알 수 있었다.

- 주문을 자주 하는 고객들이 재주문을 더 많이 할 수 있을까? 분석해보았다.

	user_id	total_orders
152339	152340	100
185640	185641	100
185523	185524	100
81677	81678	100
70921	70922	100
136869	136870	100
81703	81704	100
119931	119932	100
119834	119835	100
91034	91035	100
24530	24531	100
166135	166136	100
76593	76594	100
24561	24562	100
12640	12641	100

<유저별 총 주문 횟수가 높은 순>

	user_id	order_id	reordered
1278913	122556	1361422	1.0
2708591	167138	2882287	1.0
2708598	29602	2882295	1.0
2353422	9882	2504561	1.0
483906	108257	515193	1.0
1284257	121778	1367092	1.0
2353423	48412	2504562	1.0
1871084	4166	1991489	1.0
1871091	111651	1991497	1.0
2708592	157382	2882288	1.0
1284253	156389	1367088	1.0
483902	9503	515188	1.0
2708589	196970	2882285	1.0
1284251	195118	1367086	1.0

<재주문율이 높은 순>

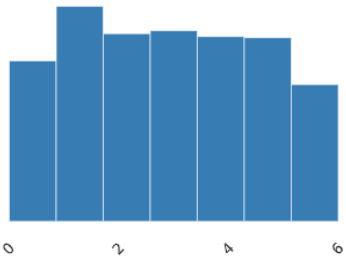
➢ 두 표의 상위에 있는 유저 아이디가 모두 다른것을 보니 주문을 자주한다고 재주문을 많이 하는것이 아니겠구나 생각하려 하였으나, 데이터가 많아 상관관계를 구해봐야겠다고 판단함

➢ 고객별 총 주문 횟수와 고객별 재주문율의 상관관계를 구해보니 0.0023379017344325828이 나왔다. 값이 1이면 완벽한 양의 상관, -1이면 완벽한 음의 상관, 0이면 상관이 없다는 것을 나타내는데 수치를보아 양의 상관관계임을 알 수 있다. (상관관계 별로 의미없는 수치임)

3. 시간적 특성

<상위 5% 사용자의 요일별 주문 수 분석>

Distinct	7	Minimum	0
Distinct (%)	< 0.1%	Maximum	6
Missing	0	Zeros	35762
Missing (%)	0.0%	Zeros (%)	12.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.8915998	Memory size	2.1 MiB



More details

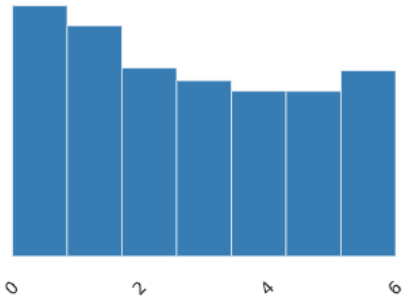
Statistics | Histogram | Common values | Extreme values

Value	Count	Frequency (%)
1	47908	17.1%
3	42503	15.1%
2	41816	14.9%
4	41189	14.7%
5	40951	14.6%
0	35762	12.7%
6	30470	10.9%

- 0은 일요일, 6은 토요일을 나타낸다
- 가설 **:** 요일에 따라 주문 수의 차이는 크게 없을 것이다
- “100% / 7 = 14.29%”과 양의 방향으로 가장 큰 차이는 월요일(+2.81%p), 음의 방향으로 가장 큰 차이는 토요일(-3.39%p)로 확인되며, 대부분의 구간에서는 큰 차이가 없다
- 한 주를 주중/주말로 나누어 볼때, 평등하게 분배된다면 주중이 71.45%, 주말이 28.55% 정도로 분배되어야 한다 (기준점).
- 하지만 상위 5%의 경우 주중이 76.4%, 주말이 23.6%로 4.95%p의 차이를 보인다. 이를 통해 상위 **5%**는 주말보다는 주중에 물품을 구매하는 경향성이 있다고 판단할 수 있다.

<하위 5% 사용자의 요일별 주문 수 분석>

Distinct	7	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	6
Missing	0	**Zeros**	11262
Missing (%)	0.0%	**Zeros (%)**	18.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2.7436086	**Memory size**	478.1 KiB



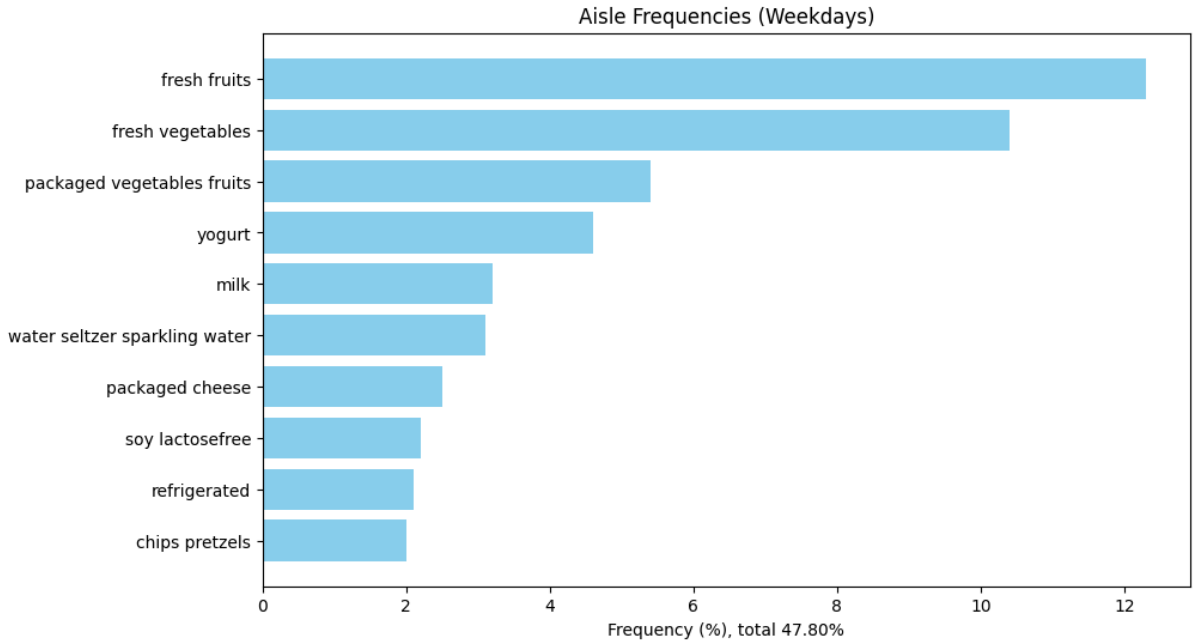
More details

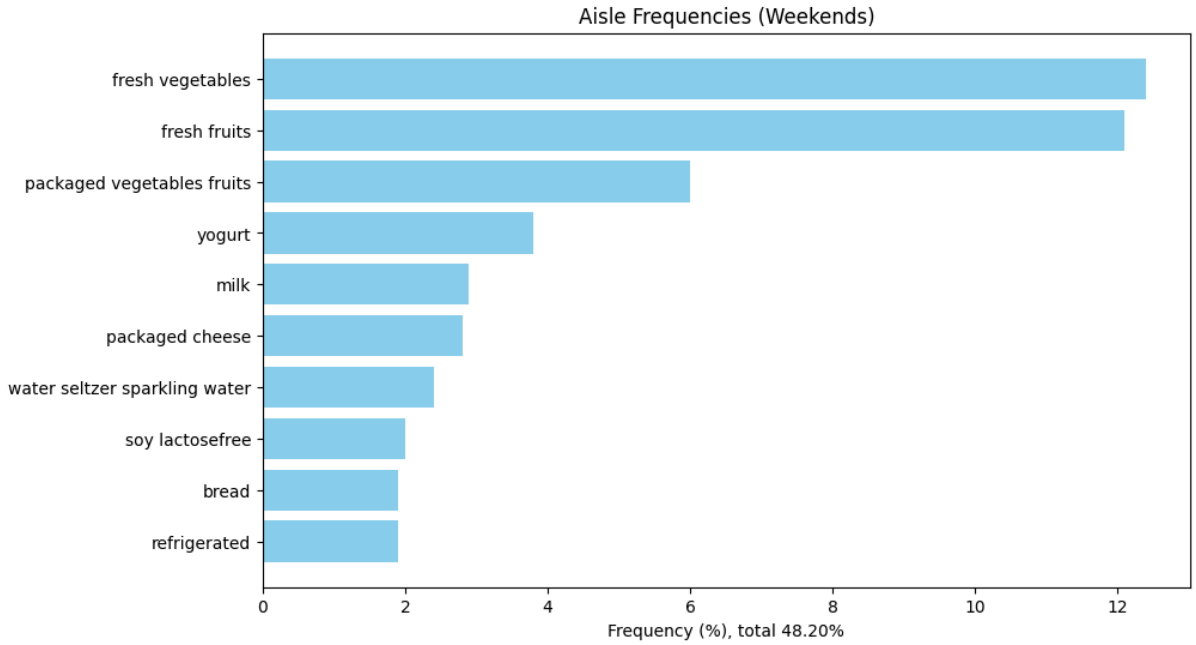
Statistics | Histogram | Common values | Extreme values

Value	Count	Frequency (%)
0	11262	18.4%
1	10366	16.9%
2	8466	13.8%
6	8346	13.6%
3	7893	12.9%
4	7425	12.1%
5	7418	12.1%

- “100 / 7 = 14.29%”과 양의 방향으로 가장 큰 차이는 일요일(+4.11%p), 음의 방향으로 가장 큰 차이는 금요일(-2.19%p)로 확인되며, 대부분의 구간에서는 큰 차이가 없다. 상위 5%와 비교해 볼 때 양의 방향으로 차이가 좀 있는 편이다.
- 하위 5%의 경우 주중이 68%, 주말이 32%로 3.45%p의 차이를 보인다. 하위 5%의 경우가 더 평등하다고 판단할수도 있다. 하지만 더욱 중요한 부분은 상위 5%의 분석과는 다른 양상을 가졌다는 점이다. 즉, 하위 **5%**의 경우는 주중보다는 주말에 물품을 구매하는 경향성이 강하다고 볼 수 있다.

<상위 5% 사용자가 주중/주말에 구매하는 상품의 종류 분석>





- 주중과 주말 주문 건수의 차이가 큰 관계로 점유율이 높은 10종류의 카테고리를 분석하기로 결정했다.
- 주중과 비교하여 주말에는 점유율 1, 2위의 순서가 바뀌며 bread가 추가되며 chips pretzels가 순위권에서 사라진다. 하지만 이는 작은 차이이며, 주중과 주말의 전반적인 카테고리 점유율이 비슷한 양상을 가진다(즉, 상품의 점유율은 요일에 dependent하지 않다)는 점에 집중하는 것이 좋을 것 같다.

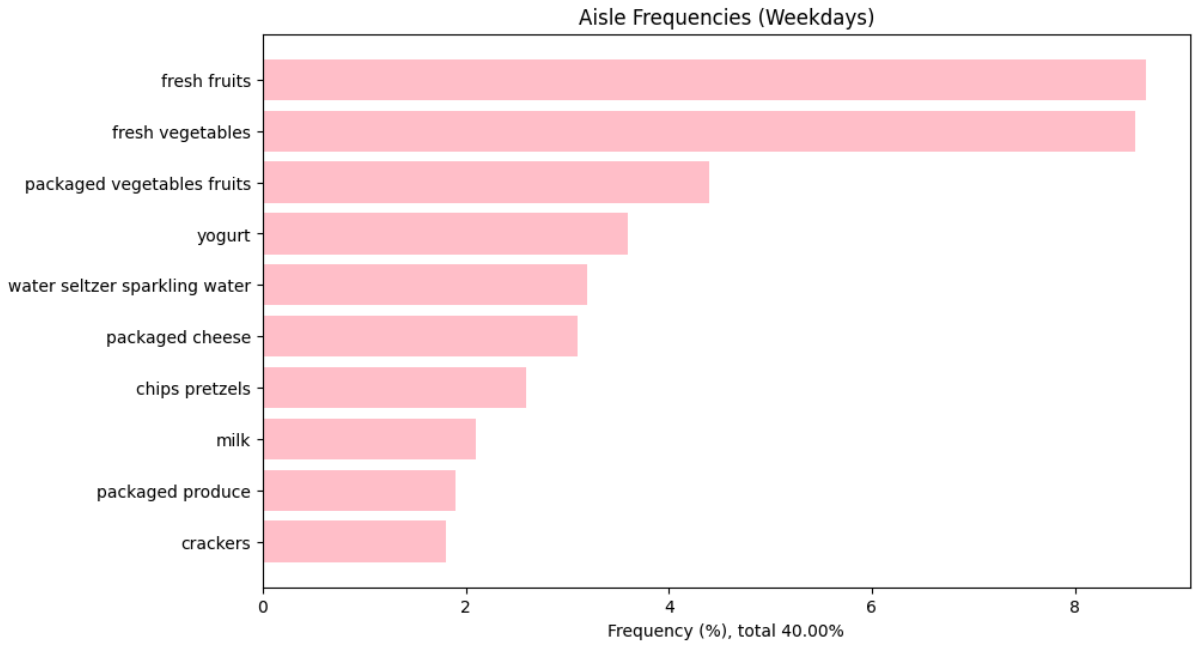
➢ 또한, 여기서 상위 5%의 사용자들이 instacart에 대해 어떻게 인식하는지 확인해 볼 수 있다. 그 전에 instacart의 메인페이지에 적혀있는 문구를 확인해보자.

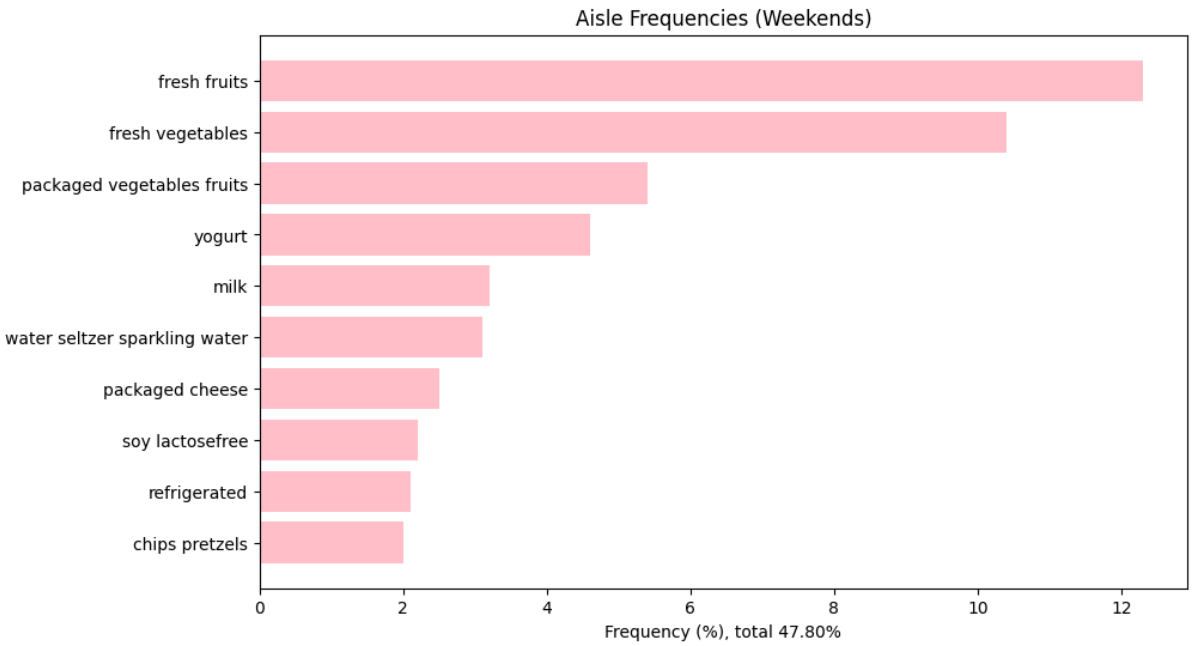
> ***Order groceries for delivery or pickup today***
> ***Whatever you want from local stores, brought right to your door.***

➢ 상점과 계약하여, 사용자들이 식료품을 배달하거나 pick up 할 수 있게해주는 서비스이다. 메인 페이지에서 사용하는 이미지는 신선한 채소 및 과일류로, instacart의 "건강한, 신선한" 이미지를 만들어준다.

➢ 상위 5%의 상품 종류도 이러한 이미지와 같은 선상에 있다. 즉, 상위 **5%**의 사람들은 **instacart**에서 건강하고 신선한 상품을 기대한다는 뜻이다.

<하위 **5%** 사용자가 주중**/**주말에 구매하는 상품의 종류 분석**>**





➢ 주중과 주말 주문 건수의 차이가 큰 관계로 점유율이 높은 10종류의 카테고리를 분석하기로 결정

➢ 두가지 관점에서 분석을 진행해보자.

1) 상위 5%의 상품 종류와 하위 5%의 주말 상품 종류의 비교

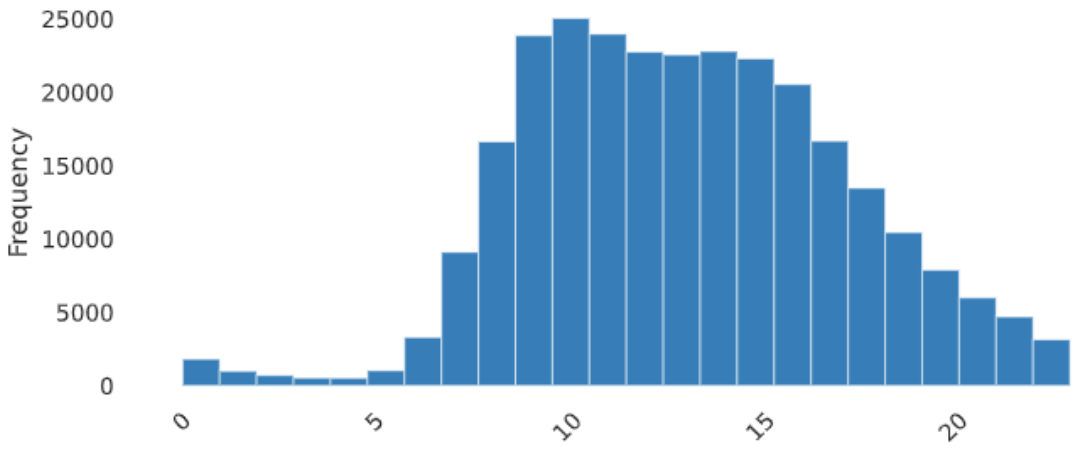
- 앞에서도 언급했듯이, 상위 5%의 그것은 주중 / 주말의 유의미한 차이는 없다.
- 하위 5%의 주말 상품 종류를 살펴보면, 이 또한 상위 5%의 그것과 유의미한 차이가 없다. 앞으로는 "상위 **5%**의 주중 **/** 주말 상품 종류"와 **"**하위 **5%**의 주말 상품 종류"는 같다고 생각하자**.**

2) 하위 5%의 주중과 상위 5%의 상품 점유율 비교

- 전반적인 점유율의 양상은 비슷하지만, **fresh fruits**와 **fresh vegetables**의 점유율이 크게 떨어졌다**.** 두 상품 모두 (상위 5%와 비교하여) 4%p에 가까운 하락폭을 보여준다. 다른 상품 종류의 변화폭이 1%p 내외인 것을 고려하면, 이는 큰 변동이다.
- 위에서 설명한 변동과 관련하여 상위 10개의 상품 종류의 총 점유율이 **8%p** 정도 떨어진 모습도 확인 할 수 있다. 즉, 하위 5%는 주중에 (상위 5%와 비교하여) 다양성이 높은 주문을 한다는 뜻이다.

➢ 결론적으로, 하위 5%의 경우 **instacart**를 사용하는 목적성이 **"**건강하고 신선한**"** 상품을 구매하기 보다는**, "**장보기"에 더 가깝다고 분석할 수 있다.

<상위 **5%** 사용자의 시간별 **order** 수 분석**>**



Quantile statistics

Minimum	0
5-th percentile	7
Q1	10
median	13
Q3	16
95-th percentile	20
Maximum	23
Range	23
Interquartile range (IQR)	6

Descriptive statistics

Standard deviation	4.2438908
Coefficient of variation (CV)	0.32297143
Kurtosis	-0.070872441
Mean	13.140143
Median Absolute Deviation (MAD)	3
Skewness	0.016580755
Sum	3687111
Variance	18.01061
Monotonicity	Not monotonic

➢ 오전 9시부터 5시까지 주문량이 증가하는 것을 알 수 있다.

➢ 해당 분석을 통해 다음과 같은 프로모션을 진행하여 매출 향상에 도움을 줄 수 있다고 판단

1. 인기 시간대 강화 프로모션 : 가장 주문이 많이 발생하는 시간대에 할인이나 묶음상품을 제공하여 구매 유도
2. 고객 특화 프로모션 : 특정시간대에 주문한 고객에게 추가 혜택(ex. 포인트 2배!)을 제공하여 고객 충성도 향상 및 유지, 더불어 해당 시간대에 주문을 유도할 수 있음
3. 시간대별 할인 이벤트 : 주문량이 낮은 20~22시까지의 시간동안 무료 배송등의 혜택을 부여하여 구매유도 , 또는 특정 시간동안에만 유효한 한정된 기간의 이벤트 개최하여 고객의 호기심 자극하여 구매를 유도할 수 있음.

➢ 조금 더 구체적인 분석을 위해 시간대에 따른 판매량의 차이에 대해 분석해보자. 이를 위해서 우선 0~23의 정수로 표현된 시간대를 instacart라는 서비스의 특성에 맞게 나눠줘야 한다. instacart의 주문 방식에 대해서 알아본 결과 다음과 같았다.

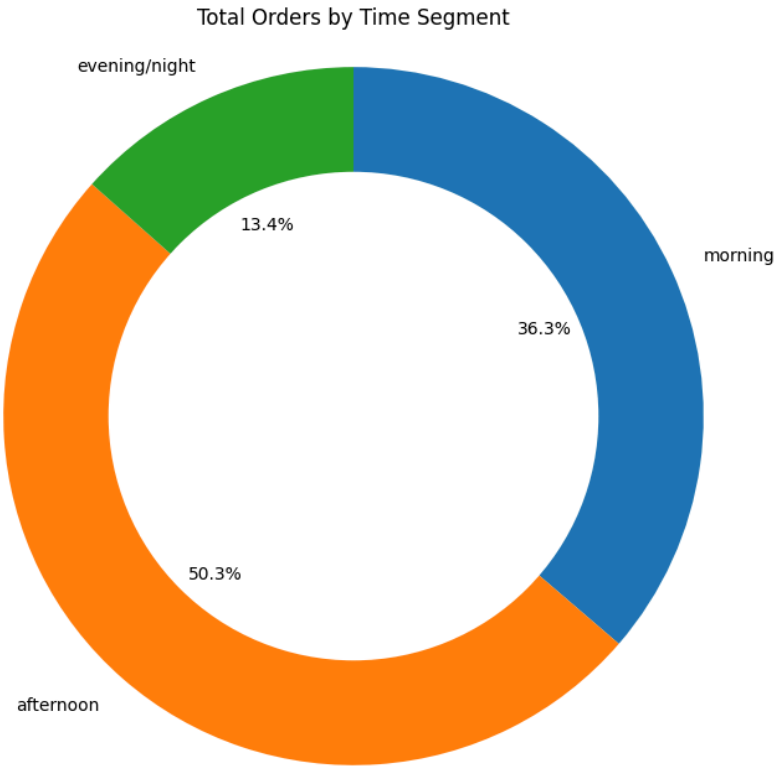
A) 2~5 시간 이내에 배송

B) 특정 날짜의 특정 시간을 예약하여 배송

C) 사용자가 직접 가게에서 픽업

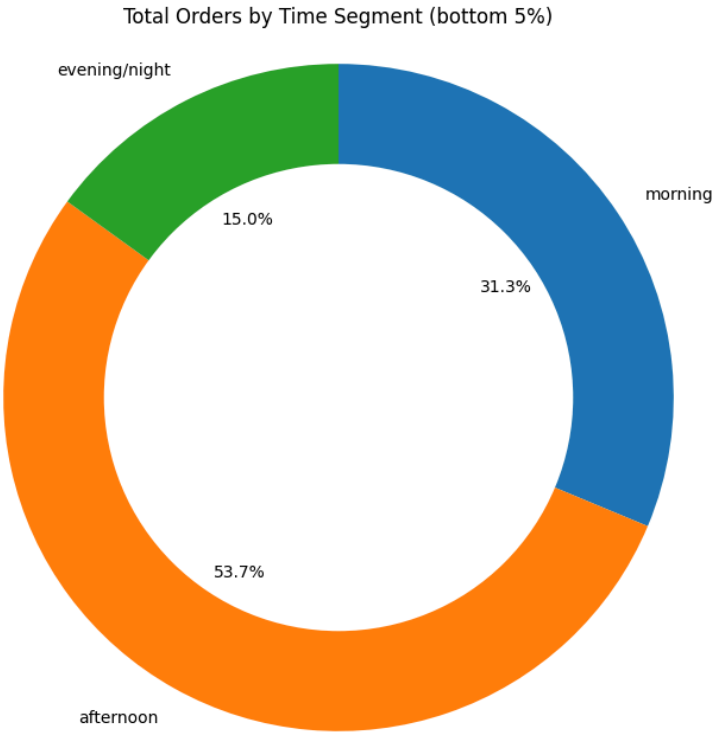
➢ 주문 방식은 위와 같이 분류되어 있지만, 데이터셋에서는 하나의 **order_id**가 어떤 주문 방식을 사용하는지 확인할 수 없기 때문에 서비스의 특성에 맞게 **time segment**를 설정할 수 없었다. 그래서, time segment의 경우 일반적인 상식에 부합하는 선에서 설정하였으며, 이는 아래와 같다.

time segment	time
morning	06~11
afternoon	12~18
evening/night	19~05



<상위 5% 사용자의 time segment별 주문 점유율>

➢ 상위 5% 그룹의 경우 afternoon 시간대가 50%의 점유율을 가지고 있다



<하위 5% 사용자의 time segment별 주문 점유율>

➢ 상위 5%와 비교한 하위 5% 그룹의 점유율 차이는 다음과 같다

time segment	delta
morning	+5%p
afternoon	-3.4%p
evening/night	-1.6%p

➢ 감소한 afternoon과 evening/night 그룹의 수치에 비해 morning 그룹의 증가량이 큰 모습을 확인할 수 있다. 따라서, 하위 **5%**에게 마케팅을 한다고 가정한다면**, morning segment**에 해당하는 시간대에 진행하는 것이 유리할 것이다**.**

<상위 **5%**의 **time segment**별 상품 카테고리의 점유**>**

aisle_id	count	Frequency	aisle
24	112998	12.70%	fresh fruits
83	91829	10.30%	fresh vegetables
123	48496	5.50%	packaged vegetables fruits
120	43880	4.90%	yogurt
84	30256	3.40%	milk
115	27365	3.10%	water seltzer sparkling water
21	23275	2.60%	packaged cheese
91	19860	2.20%	soy lactosefree
31	18033	2.00%	refrigerated
107	17485	2.00%	chips pretzels

➢ 해당 내용은 **<**상위 **5%** 사용자가 주중**/**주말에 구매하는 상품의 종류 분석**>**과 큰 차이가 없었다. 해당 내용을 대표할 수 있는 morning 그룹의 표만 첨부한다.

<하위 **5%**의 **time segment**별 상품 카테고리의 점유**>**

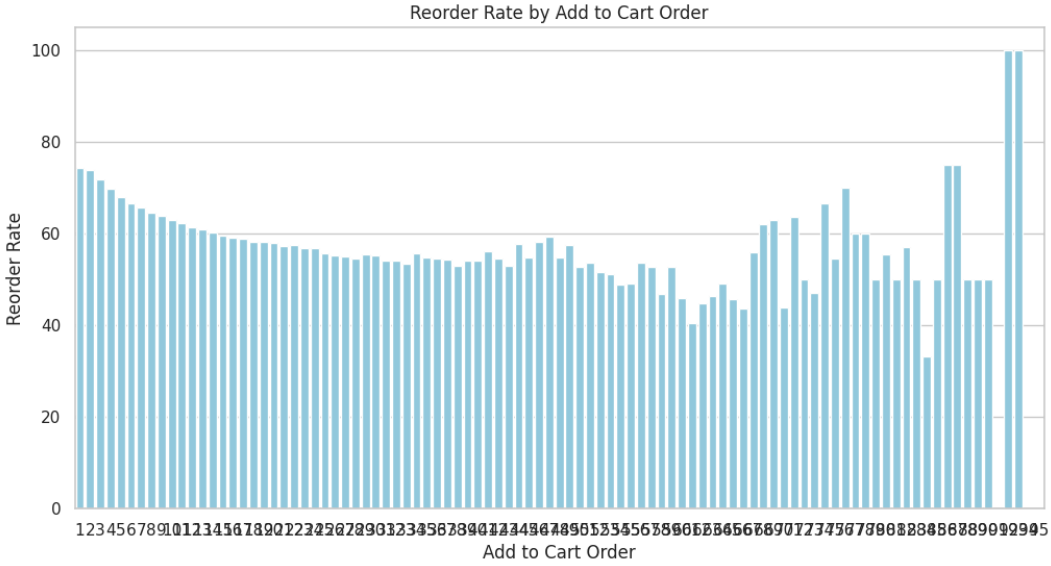
aisle_id	count	Frequency	aisle
24	14528	8.90%	fresh fruits
83	14312	8.70%	fresh vegetables
123	7316	4.50%	packaged vegetables fruits
120	5992	3.70%	yogurt
115	5404	3.30%	water seltzer sparkling water
21	4917	3.00%	packaged cheese
107	4020	2.50%	chips pretzels
32	3517	2.10%	packaged produce
84	3474	2.10%	milk
78	2945	1.80%	crackers

➢ 해당 내용 또한 **<**하위 **5%** 사용자가 주중**/**주말에 구매하는 상품의 종류 분석**>**과 큰 차이가 없었다. 해당 내용을 대표할 수 있는 morning 그룹의 표만 첨부한다.

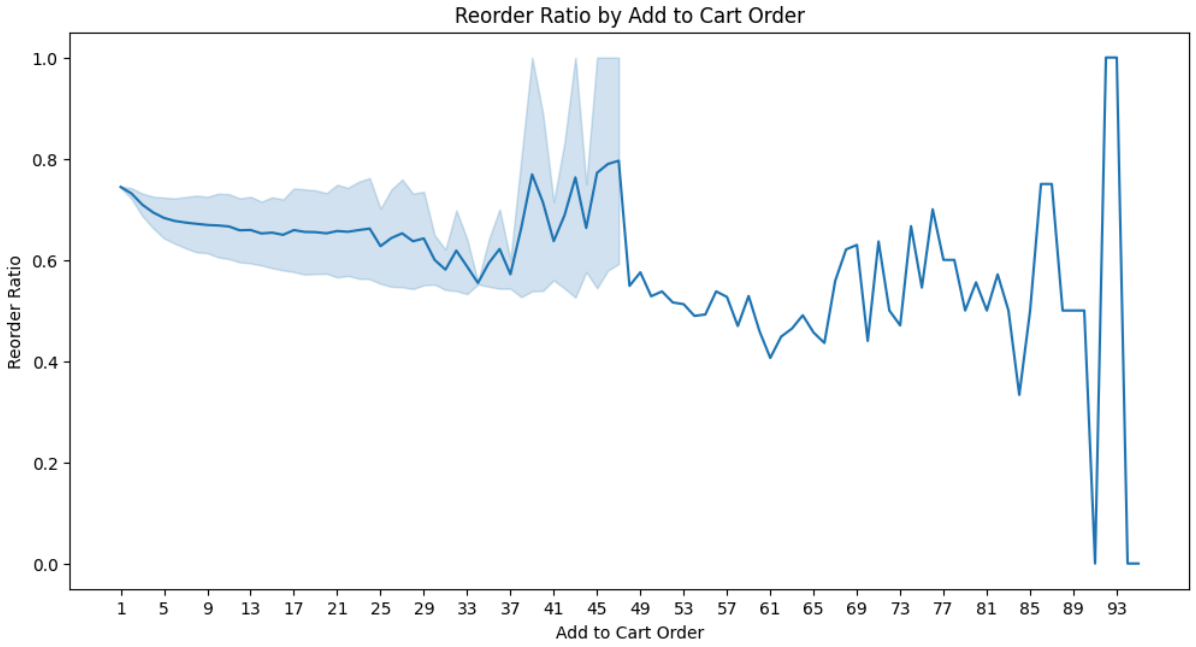
➢ 결국, 두 그룹의 time segment별 상품 카테고리의 점유율을 사용한 분석을 하면 **<**주중**/**주말에 구매하는 상품의 종류 분석**>**과 같은 결과가 나온다.

<상위 **5%** 사용자의 재주문과 카트 추가 순서의 분포 분석**>**

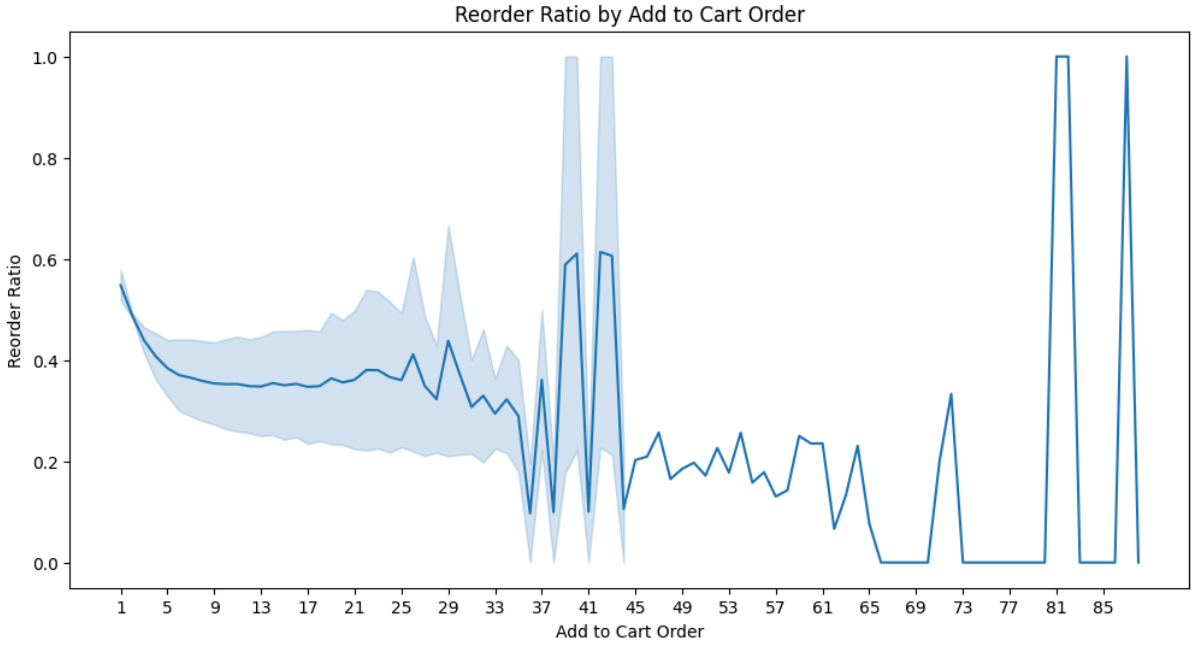
- 재주문이 일어날 때 카트에 제품을 추가한 순서의 분포를 확인하여 특정 순서에서 재주문이 더 자주 발생하는지 분석해보자.



- 상위 5% 사용자는 마지막 순서에서 재주문이 많이 발생하는 것을 알 수 있다.
- 이를 통해 상위 5% 사용자는 마지막 순서에서 재주문을 많이 하는것을 알 수 있다. 하지만 시각화 가독성이 떨어짐으로 x축 segment를 나눠보았다.



<상위 5% 'add_to_cart_order'에 따른 재주문 비율>



<하위 5% 'add_to_cart_order'에 따른 재주문 비율 >

두 그래프로 보았을 때 상위5%나 하위 5%의 그래프 양상은 비슷한걸로 보인다.

다만 상위 5%의 사용자들은 하위 5%사용자보다 약 1~ 73 순서에서 더 높은 재주문율을 보임을 알 수 있다. 이를 통해 하위 5% 사용자에게 재주문을 유도하는 프로모션이나 할인쿠폰을 적용하여 매출상승에 도움을 줄 수 있을 것으로 판단된다.

📌기대효과 및 결론

📃 이탈 고객 분석 : 이탈 고객은 장바구니 사용률이 높고 한번에 대량 구매를 하는 경향이 있어 구매주기가 긴 것으로 확인되었다. 이러한 특징을 보았을때, 배송비에 부담을 느껴 대량구매를 한다고 결론 지었다. 문제 해결 방안으로 정기적인 배송 서비스를 할 시에 배송비 할인을 해주는 이벤트 등 을 고민해볼 수 있다.

📃 재구매를 많이 하는 사용자의 특징 분석

- milk와 egg, yogurt와 같은 카테고리가 재주문률이 높았다. 이러한 점을 고려하여 카테고리 특성을 이용하여 정기배송같은 프로모션을 진행하면 매출 증가에 도움이 될 것으로 보인다.
- 하위 5%에게 마케팅을 한다고 가정한다면, morning segment에 해당하는 시간대에 진행하는 것이 유리할 것이다
- 상위 5% 사용자를 대상으로 마케팅할때는 주말보다는 주중에 하는것이 매출상승에 도움이 된다.
- 오전 9시부터 오후 5시까지 주문량이 증가하므로 인기 시간대 강화 프로모션, 고객 특화 프로모션,시간대별 할인 이벤트 등을 기획하여 매출 증가를 기대할 수 있다.
- 위와 같은 항목들은 **A/B Test**와 같은 실험을 통해 효용성을 시험해 볼 수 있다.
- 하위 5%의 사용자는 instacart를 사용하는 목적성이 “건강하고 신선한” 상품을 구매하기 보다는, “장보기”에 더 가깝다고 분석
    - 해당 분석 결과는 상위 사용자를 겨냥하여 “건강하고 신선한” 상품들의 홍보에 집중할 것인지(즉, whole food와 같은 건강한 이미지를 계속 유지할 것인지), 아니면 하위 사용자를 고려하여 대중성이 높은 상품들의 홍보에 집중할 것인지를 판단해야 할 것 같다.