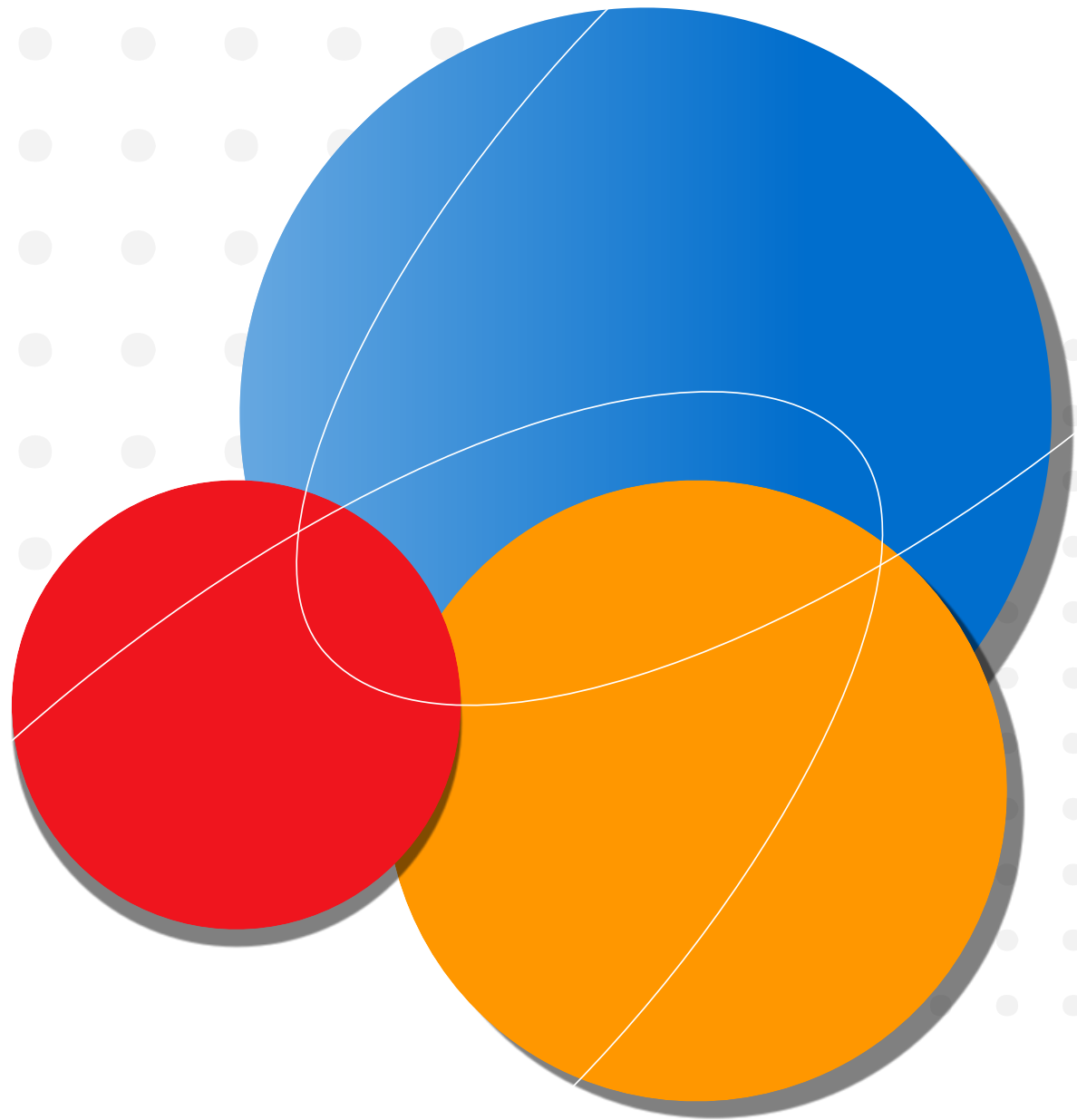



Track 2 Modeling PART



2023 BDA

CONTEST

일시 2023년 6월 4일
팀명 프라모델
후원 CJ제일제당



CONTENTS



01 데이터 분석 및 시각화

- 변수 시각화
- 가설 검증을 통한 데이터 해석

02 데이터 전처리


- 파생 변수 생성
- 칼럼 정규화

03 모델링

- 사용한 방식
- 결과 분석

04 의의 및 한계점

- 모델링이 가지는 의미와 예상 활용 방안
- 한계점 & 보완사항





01

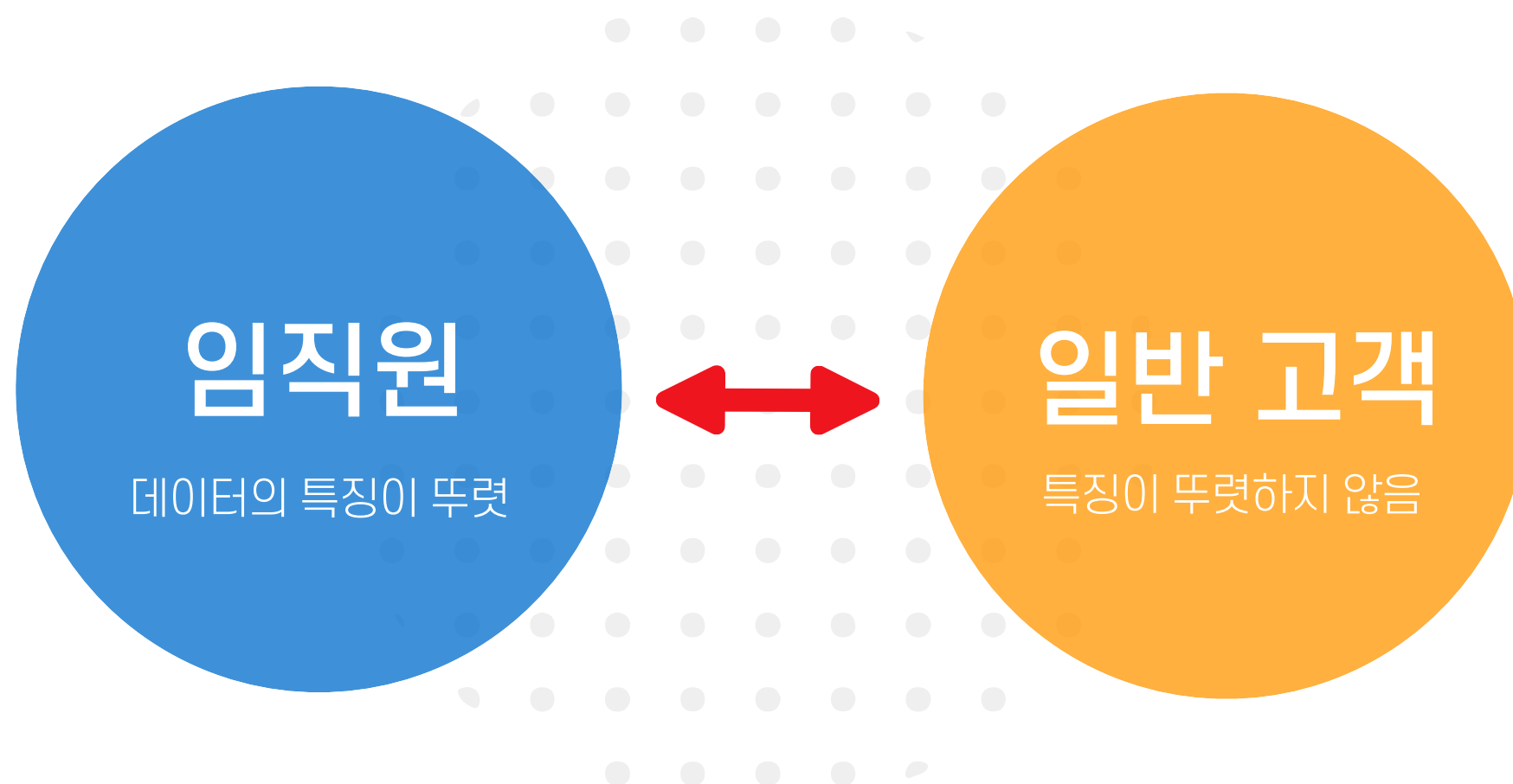
데이터 분석 및 시각화

변수 시각화
가설 검증을 통한 데이터 해석

분석 과제

Track 2 : 모델링 고도화

2023년도 1월 CJ 더마켓 고객 주문 데이터를 활용하여 **프라임 회원 예측** 모델링 진행



- 임직원 - 임직원 데이터셋 사용
- 일반 회원 - 전체 데이터셋 (임직원 + 일반회원) 사용

∴ **성능적 향상**을 얻을 수 있을 것으로 예상

본선 데이터 확인

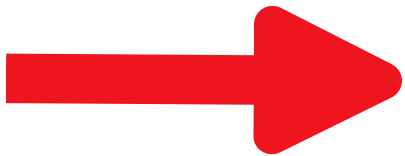
| | scd | product_name | net_order_qty | net_order_amt | gender | age_grp | employee_yn | order_date | prime_yn |
|---|----------------|-------------------|---------------|---------------|--------|---------|-------------|------------|----------|
| 0 | 20230124153976 | 잔칫집 식혜 240ml 30입 | 1 | 9.803170 | F | 2 | Y | 20230124 | N |
| 1 | 20230124155563 | 백설 한입속 비엔나 120g*2 | 1 | 8.256607 | M | 3 | Y | 20230124 | N |
| 2 | 20230125158386 | 비비고 왕교자 1.05kg | 1 | 9.348449 | F | 4 | N | 20230125 | N |
| 3 | 20230126164638 | 고메 바삭쫄깃한 탕수육 900g | 1 | 9.667259 | F | 4 | N | 20230126 | Y |
| 4 | 20230125159705 | 햇반 매일잡곡밥210g | 20 | 9.994653 | M | 4 | N | 20230125 | Y |

- 9개변수, 45875 행으로 이루어진 고객 주문 데이터

독립 변수

주문 번호, 제품명, 주문 수량, 주문 금액,
성별, 연령대, 임직원 유무

* 고객의 한 가지 주문을 상품별로 행을 구분하여 중복값 존재
→ 고유값이 실질적인 1월 주문건수로 총 10653개



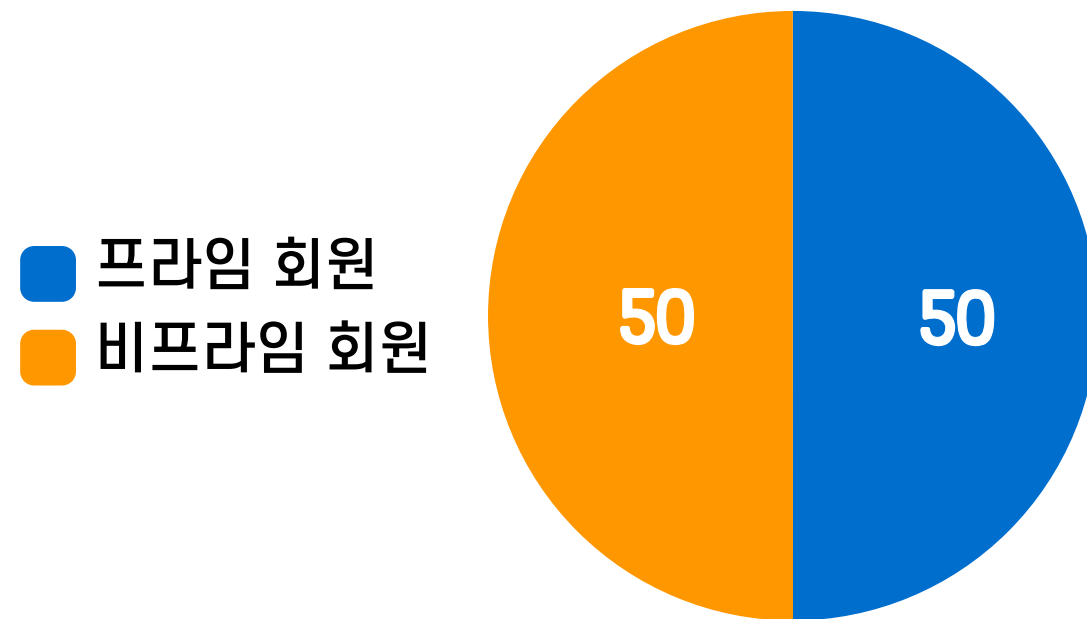
종속 변수

프라임 회원 유무

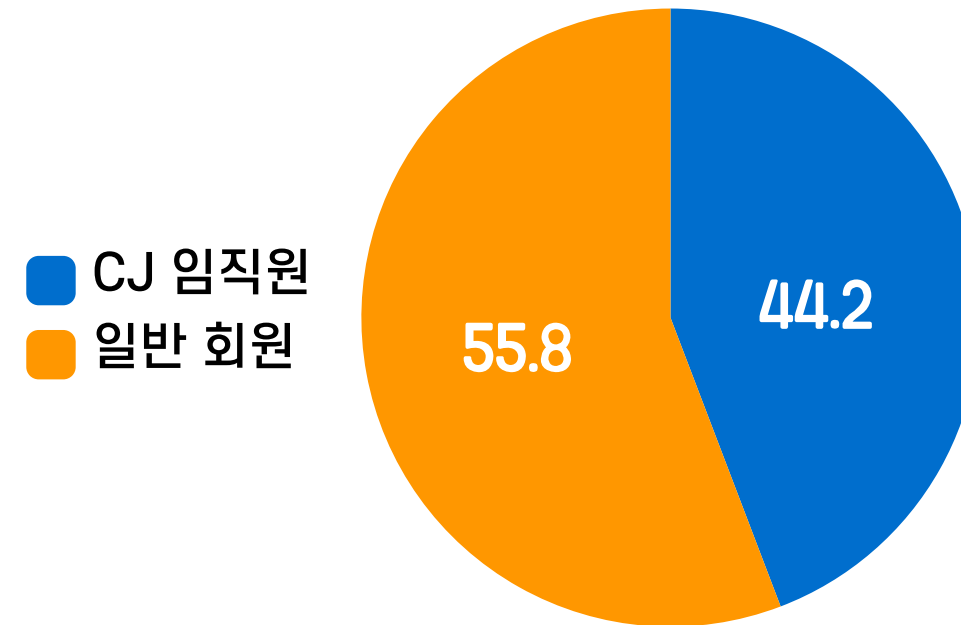
칼럼 분포 확인

프라임 / 임직원

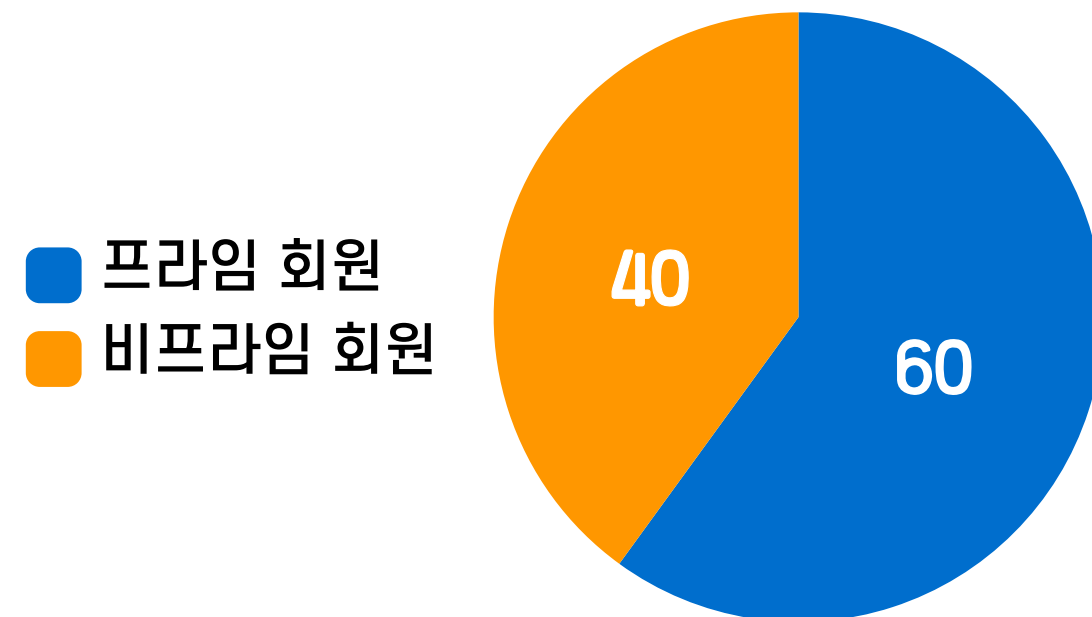
01 프라임 회원 비율



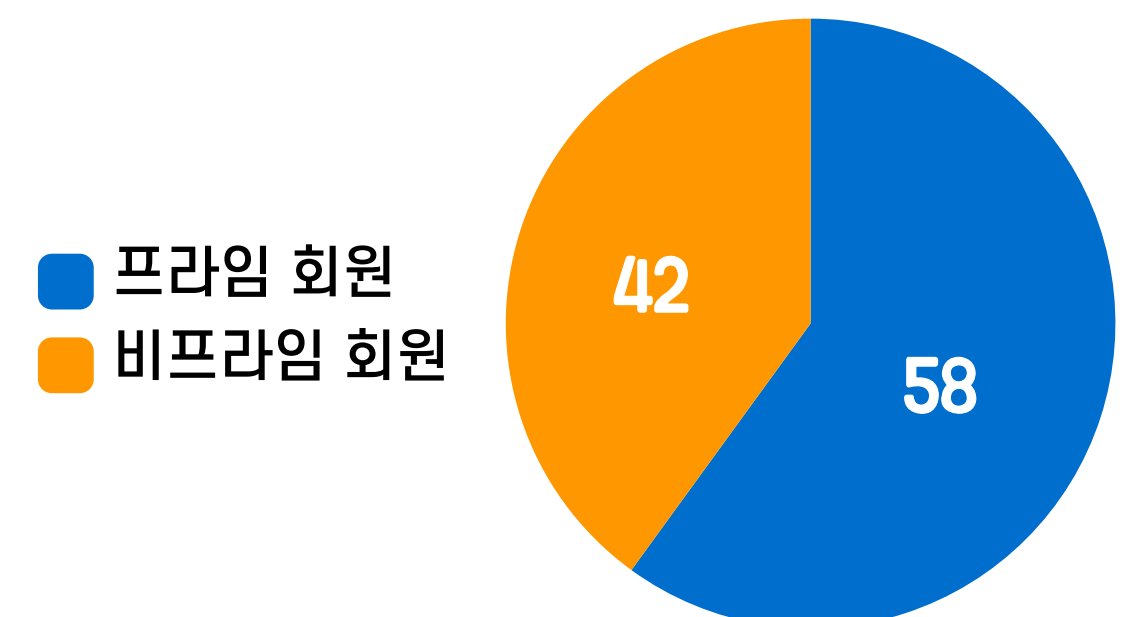
02 CJ 임직원 비율



03 임직원 중 프라임 회원 비율



04 비임직원 중 프라임 회원 비율



가설 검증 1

프라임 회원은 혜택을 받아 일반 회원보다 동일 상품을 저렴하게 구매했을 것이다

● 상품별 주문 수량 & 주문 금액 평균

| 임직원 | 프라임 | 주문 수량 평균 | 주문 금액 평균 |
|--------|-----|----------|----------|
| CJ 임직원 | Y | 1.8321 | 9.3528 |
| | N | 1.7007 | 9.3977 |
| 일반 회원 | Y | 1.7213 | 8.8739 |
| | N | 1.5808 | 9.0347 |

- 주문 수량 평균
임직원 / 비임직원 모두
프라임 회원 주문 수량이 높음
- 주문 금액 평균
임직원 / 비임직원 모두
프라임 회원 주문 금액이 저렴함

가설 검증 1

프라임 회원은 혜택을 받아 일반 회원보다 동일 상품을 저렴하게 구매했을 것이다

● 상품명과 주문 수량별로 묶어 주문 금액 평균 구하기

● 프라임 회원

| product_name | net_order_qty | net_order_amt_prime |
|---|---------------|---------------------|
| (냉동) 비비고 테이블 특 선물세트 (특양지곰탕 700gx2개+특설렁탕700gx1개) | 1 | 10.220281 |
| (냉동) 비비고 테이블 특 선물세트 (특양지곰탕 700gx2개+특설렁탕700gx1개) | 2 | 10.914124 |
| (냉동) 비비고 테이블 특설렁탕 700g | 1 | 8.910623 |
| (냉동) 비비고 테이블 특설렁탕 700g | 2 | 9.708680 |
| (냉동) 비비고 테이블 특설렁탕 700g | 3 | 9.879576 |
| ... | ... | ... |
| 헬씨누리 침향환 환심 10환 | 1 | 9.514658 |
| 헬씨누리 침향환 환심 10환X6입(1BOX)_행사 | 1 | 11.043706 |
| 훈제대란 20구 | 1 | 9.126706 |
| 훈제대란 20구 | 2 | 9.649498 |
| 훈제대란 20구 | 3 | 10.410456 |

● 일반 회원

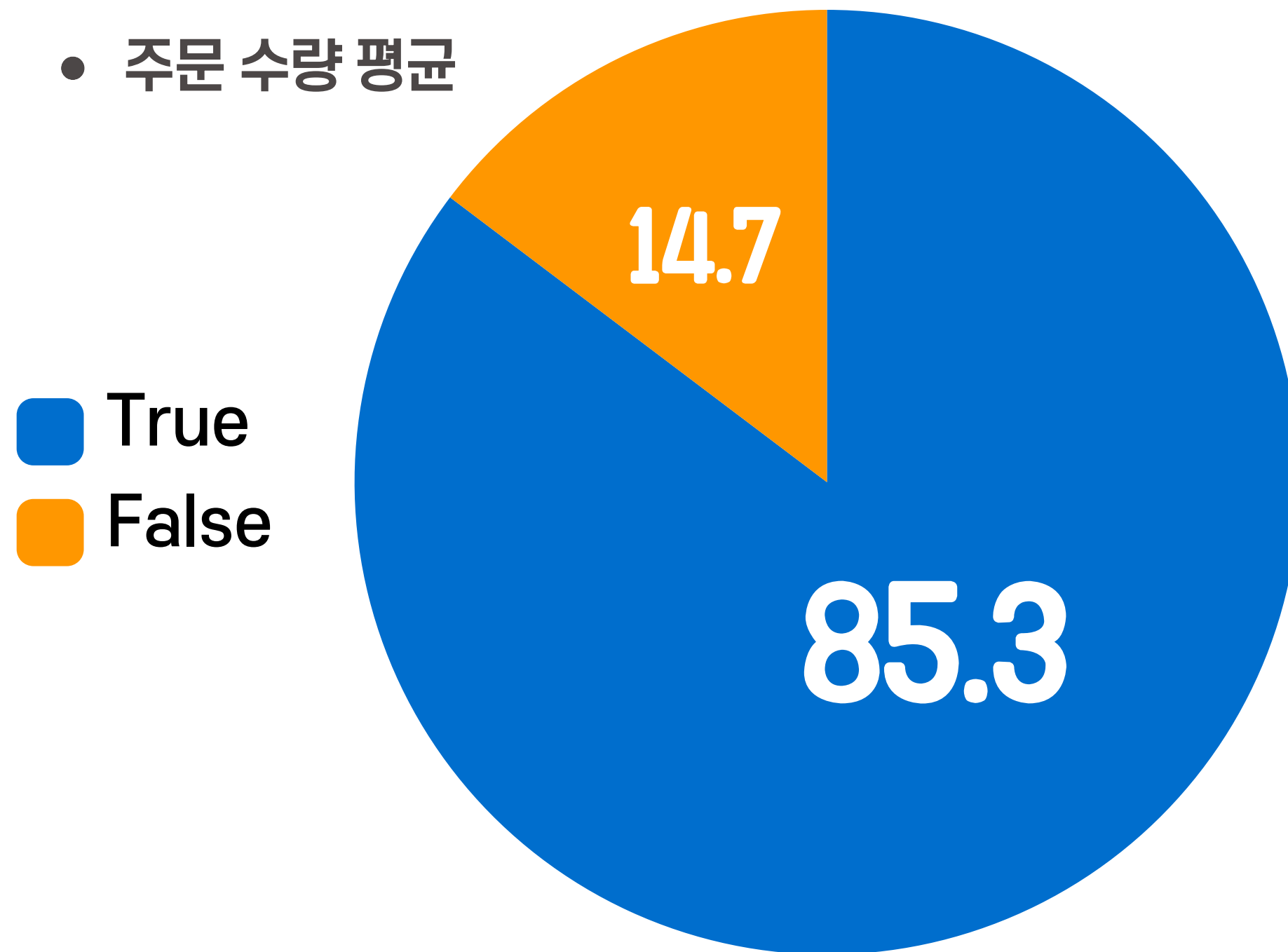
| product_name | net_order_qty | net_order_amt_nonprime |
|--|---------------|------------------------|
| (냉동) 비비고 테이블 특 선물세트 (특양지곰탕700gx2개+특설렁탕700gx1개) | 1 | 10.270577 |
| (냉동) 비비고 테이블 특설렁탕 700g | 1 | 8.980387 |
| (냉동) 비비고 테이블 특설렁탕 700g | 2 | 9.713655 |
| (냉동) 비비고 테이블 특설렁탕 700g | 5 | 10.896758 |
| (냉동) 비비고 테이블 특설렁탕 700gx2개 | 1 | 9.875140 |
| ... | ... | ... |
| 행복한콩 폭신폭신티 두부볼 750g | 3 | 10.263118 |
| 헬씨누리 침향환 환심 10환 | 1 | 9.562475 |
| 헬씨누리 침향환 환심 10환X6입(1BOX)_패밀리데이 | 1 | 11.091499 |
| 훈제대란 20구 | 1 | 9.255307 |
| 훈제대란 20구 | 2 | 10.052812 |

가설 검증 1

프라임 회원은 혜택을 받아 일반 회원보다 동일 상품을 저렴하게 구매했을 것이다

● 상품명과 주문 수량별로 묶어 주문 금액 평균 구하기

● 주문 수량 평균



가설 검증

동일 상품에 대해 프라임 회원이 일반 회원보다 **저렴하게 구매**하는 경향이 있다.

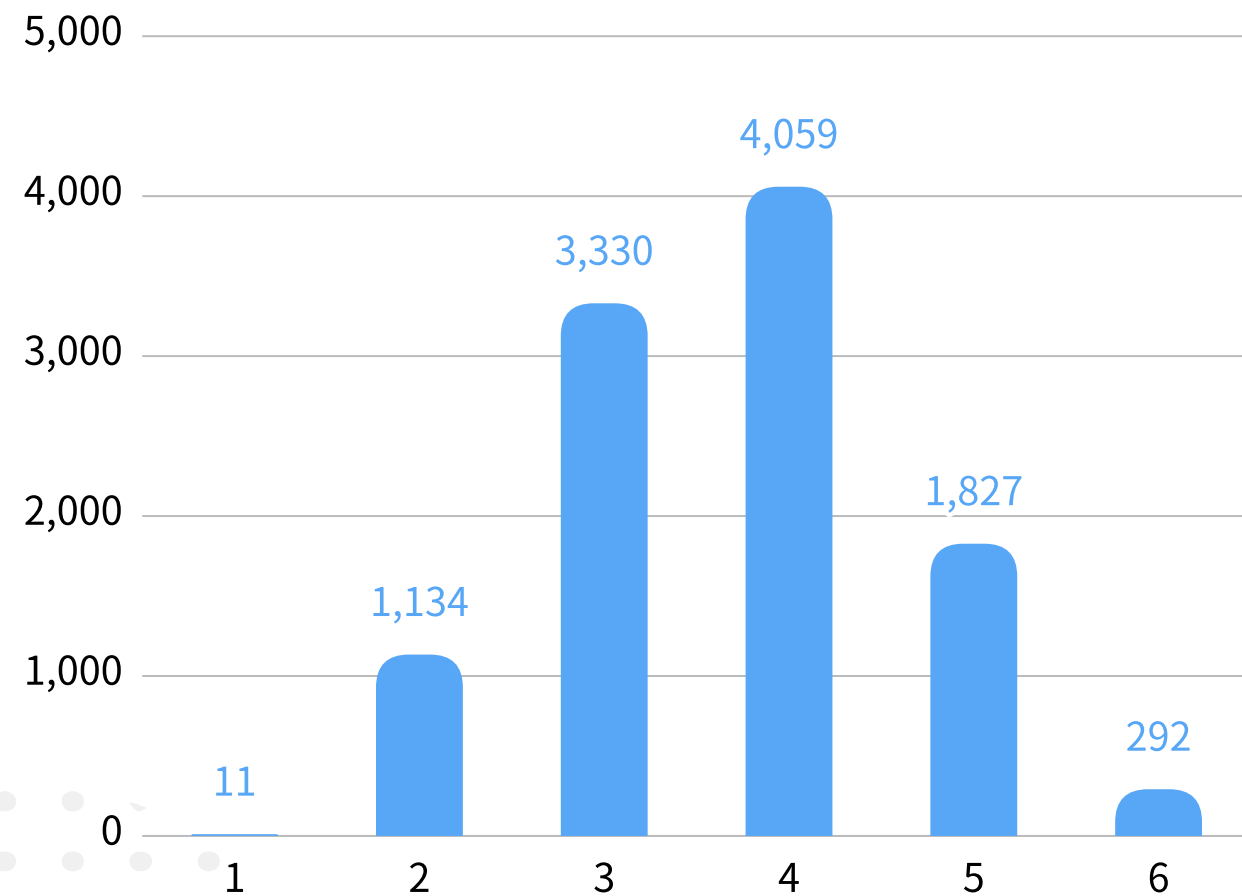
* 1월 프라임회원 혜택 7% 무제한 할인 존재

∴ **'is_price_lower_than_avg'** 변수로 추가 예정

가설 검증 2

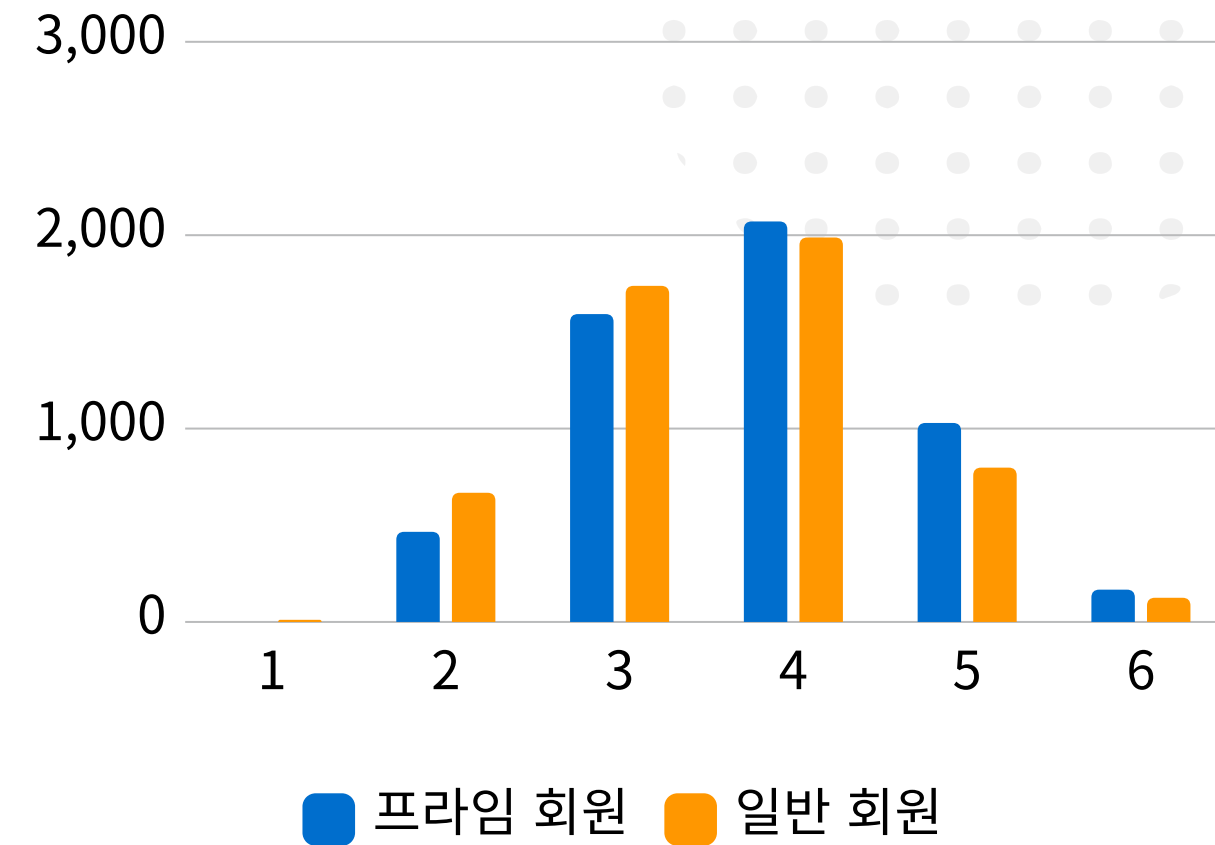
30 ~ 40대 회원 중 프라임 회원이 많을 것이다

연령층별 분포

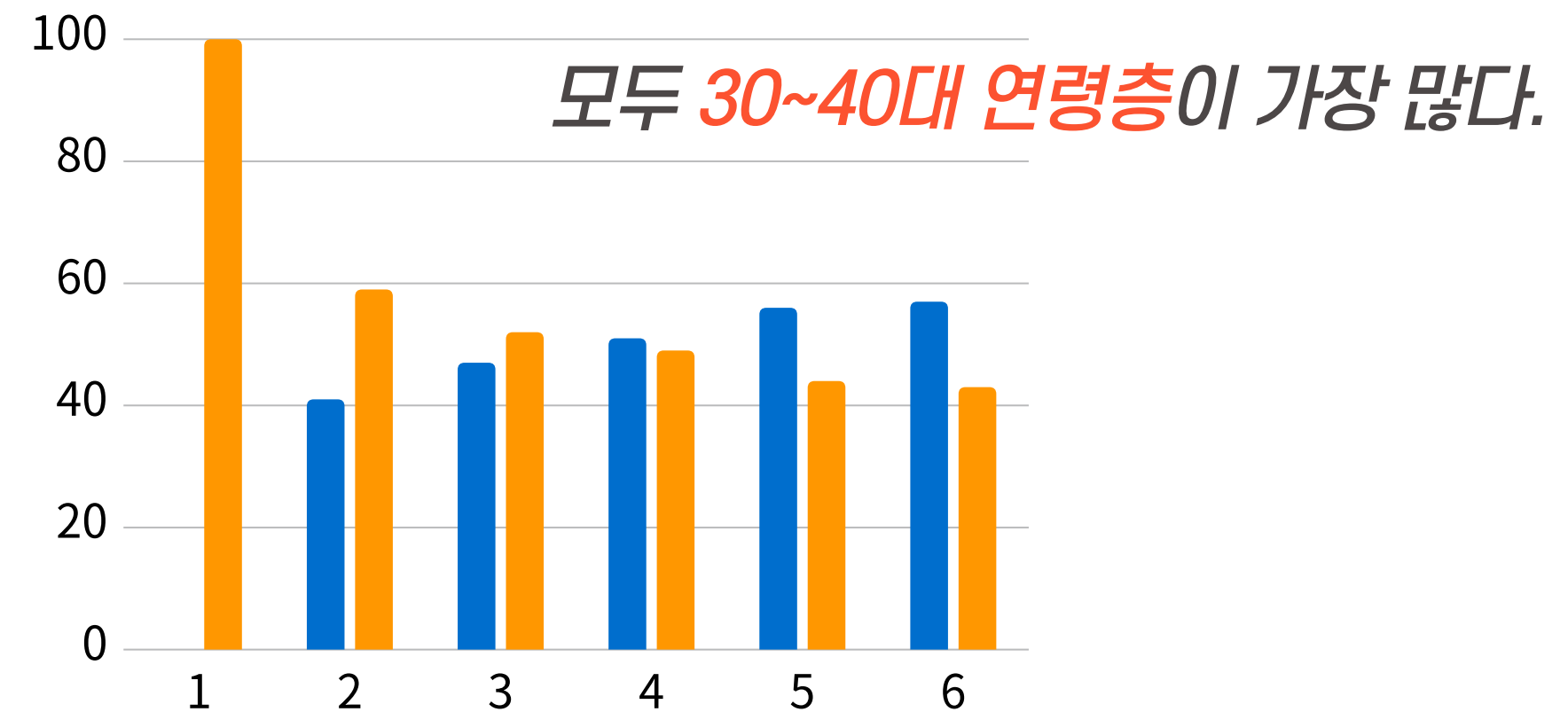


수입이 안정적이며 신선식품물을 애용하는 30~40대층
수입이 적은 10대와 온라인 활용률이 낮은 60대 이상층

연령층별 프라임 회원 수



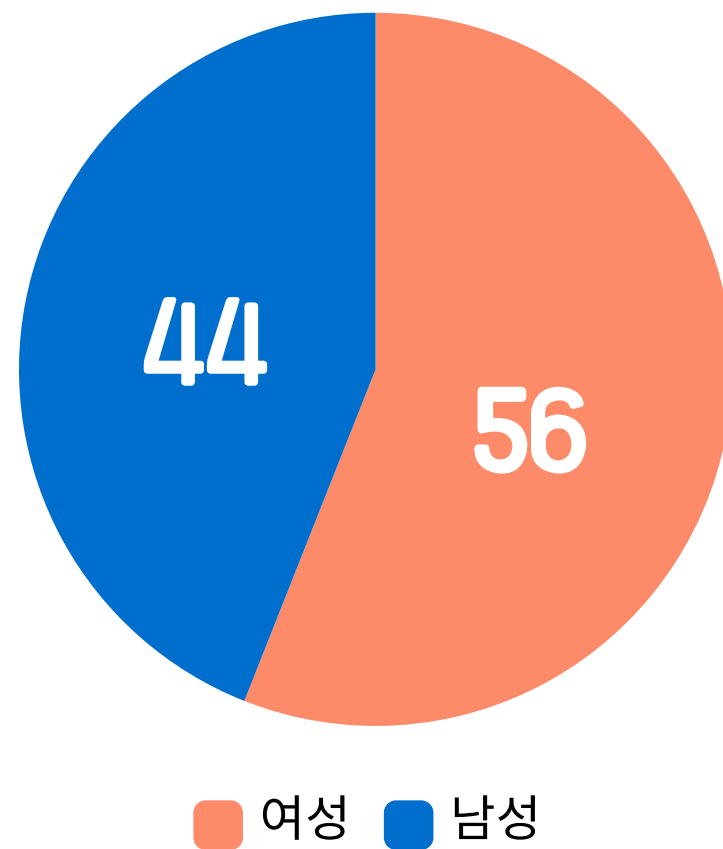
연령층별 프라임 회원 비율



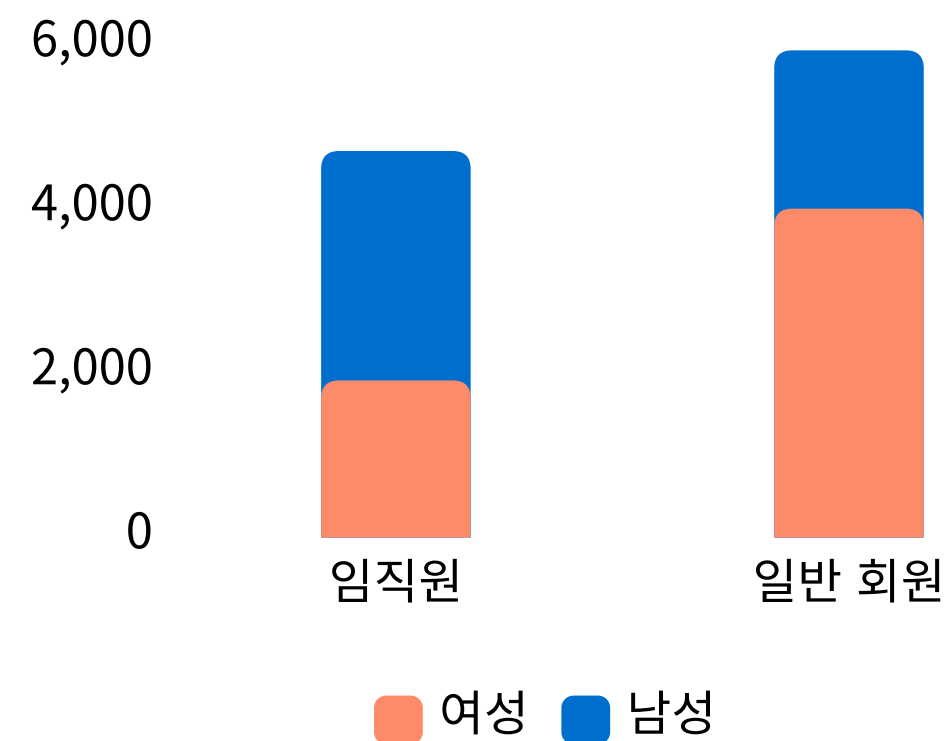
가설 검증 3

여성 회원 중 프라임 회원이 많을 것이다

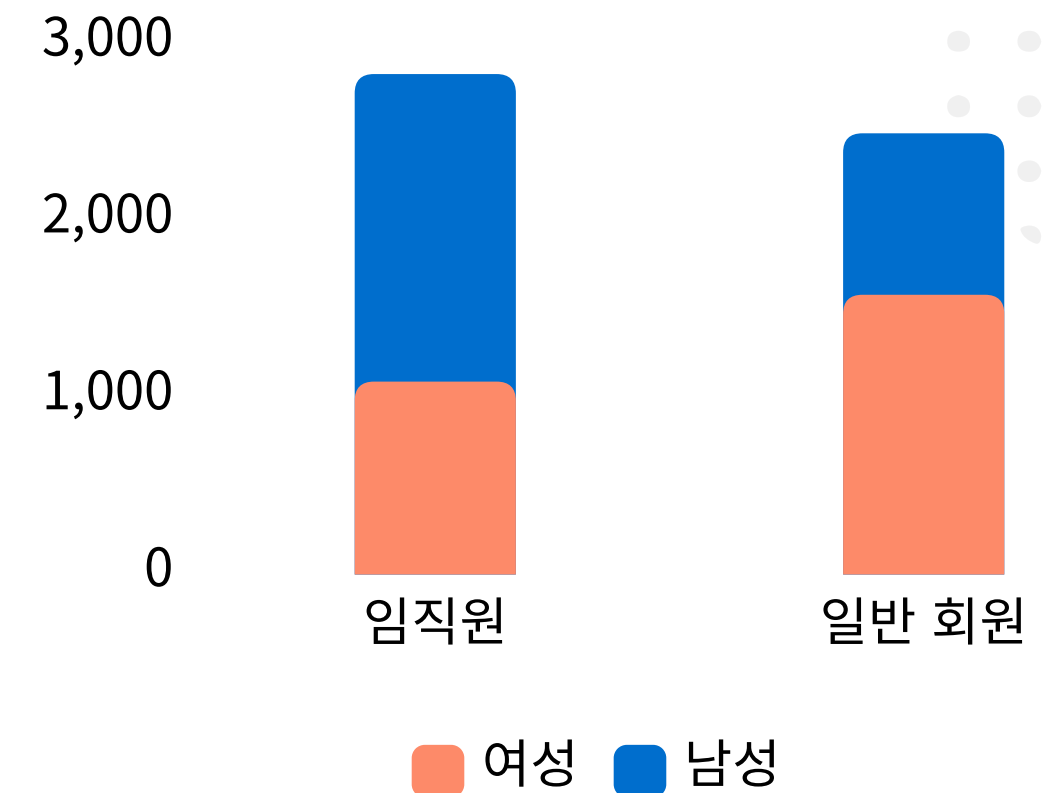
• 성별 분포



• 임직원 여부에 따른 성별 분포



• 프라임 회원 성별 분포



가설 검증

여성과 남성의 프라임 수는 크게 차이 없으나,
임직원 중에는 남성이, 일반 회원은 여성 프라임 회원이 많다.

가설 검증 4

프라임 회원은 프라임이 아닌 회원보다 총 주문 품목 개수가 많을 것이다

● 프라임 회원과 일반 회원의 상품별 주문 수량 비교

| 프라임 | 수량 1개 | 수량 2개 이상 |
|--------|--------------|-------------|
| 프라임 회원 | 17,502 (52%) | 7,152 (58%) |
| 일반 회원 | 16,078 (48%) | 5,143 (42%) |

둘 다 프라임 회원의 비율이 높으나

2개 이상일 때 프라임 회원과 일반 회원 간의 차이가 큼

● 주문 건당 총 주문 상품 개수

| 프라임 | 평균 개수 | 중위값 |
|--------|-------|-----|
| 프라임 회원 | 8개 | 5개 |
| 일반 회원 | 6개 | 4개 |

평균과 중위값 모두 **프라임 회원**이 더 많음

가설 검증 4

프라임 회원은 프라임이 아닌 회원보다 총 주문 품목 개수가 많을 것이다

● 총 주문 수량 별 프라임/ 일반 회원 구성 비율

● 프라임 회원

| | net_order_qty | cnt | prime_yn | prop | cummulative_prop |
|----|---------------|------|----------|-----------|------------------|
| 0 | 1 | 1062 | Y | 19.943662 | 19.943662 |
| 1 | 2 | 649 | Y | 12.187793 | 32.131455 |
| 2 | 3 | 459 | Y | 8.619718 | 40.751174 |
| 3 | 4 | 372 | Y | 6.985915 | 47.737089 |
| 4 | 5 | 370 | Y | 6.948357 | 54.685446 |
| 5 | 6 | 294 | Y | 5.521127 | 60.206573 |
| 6 | 7 | 230 | Y | 4.319249 | 64.525822 |
| 7 | 8 | 207 | Y | 3.887324 | 68.413146 |
| 8 | 9 | 157 | Y | 2.948357 | 71.361502 |
| 9 | 10 | 201 | Y | 3.774648 | 75.136150 |
| 10 | 11 | 147 | Y | 2.760563 | 77.896714 |

● 일반 회원

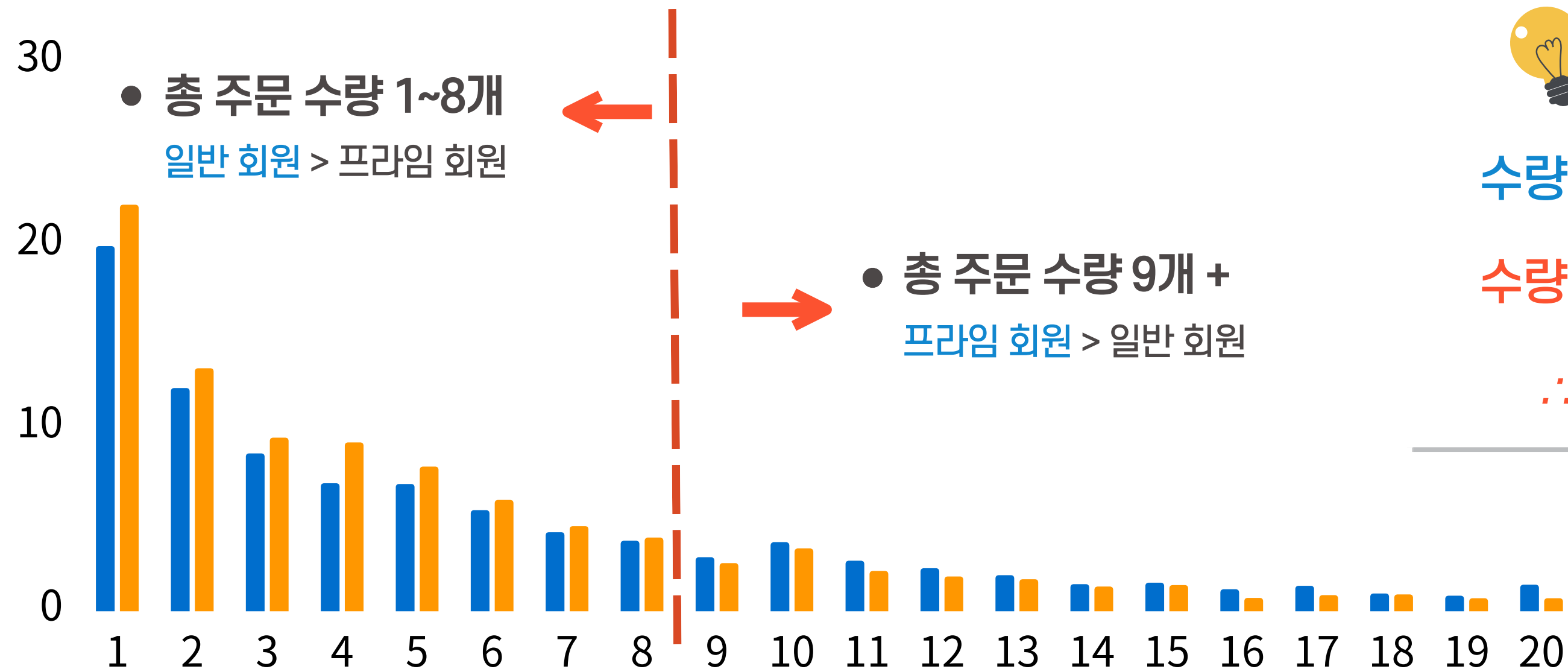
| | net_order_qty | cnt | prime_yn | prop | cummulative_prop |
|----|---------------|------|----------|-----------|------------------|
| 0 | 1 | 1183 | N | 22.203453 | 22.203453 |
| 1 | 2 | 707 | N | 13.269520 | 35.472973 |
| 2 | 3 | 505 | N | 9.478228 | 44.951201 |
| 3 | 4 | 491 | N | 9.215465 | 54.166667 |
| 4 | 5 | 421 | N | 7.901652 | 62.068318 |
| 5 | 6 | 324 | N | 6.081081 | 68.149399 |
| 6 | 7 | 248 | N | 4.654655 | 72.804054 |
| 7 | 8 | 214 | N | 4.016517 | 76.820571 |
| 8 | 9 | 140 | N | 2.627628 | 79.448198 |
| 9 | 10 | 183 | N | 3.434685 | 82.882883 |
| 10 | 11 | 117 | N | 2.195946 | 85.078829 |

가설 검증 4

프라임 회원은 프라임이 아닌 회원보다 총 주문 품목 개수가 많을 것이다

● 총 주문 수량 별 프라임/ 일반 회원 비율 분포

■ 프라임 회원 ■ 일반 회원



가설 검증

수량 ▼ - 비프라임 회원 비율 ↑

수량 ▲ - 프라임 회원 비율 ↑

∴ 'tot_qty_9' 변수 추가

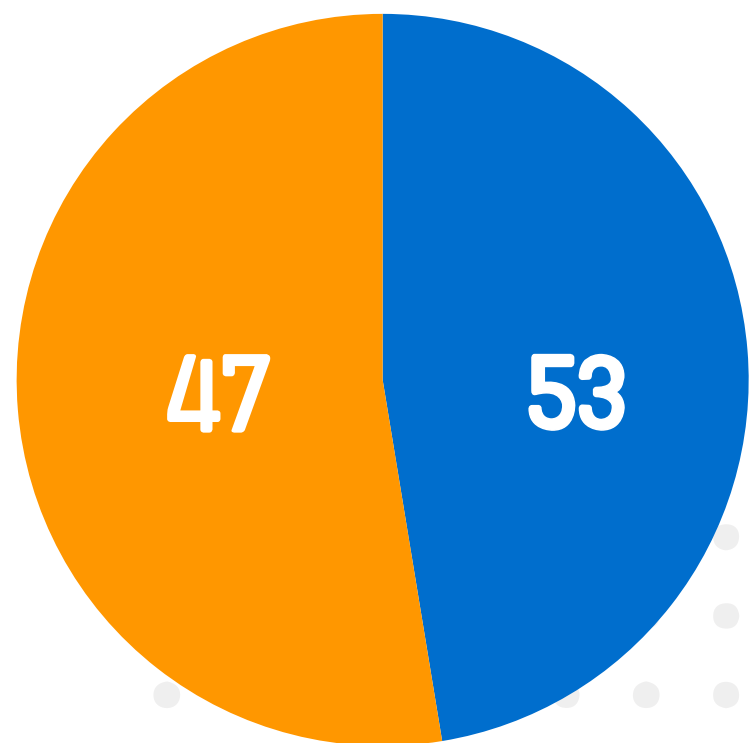
가설 검증 5

구매 품목 중 인기 상품의 비중이 높은 회원은 프라임 회원일 확률이 높을 것이다

인기 상품 추정

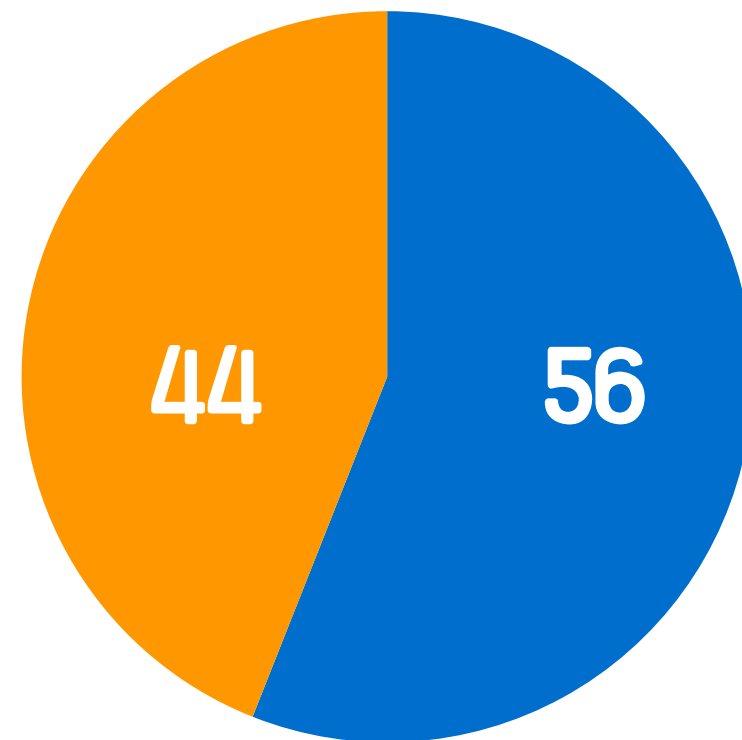
상품명 개수를 count해서 상위 100개 상품을 1월의 인기상품으로 정의

인기 상품 구매 고객



■ 프라임 회원
 ■ 일반 회원

미구매 고객



■ 프라임 회원
 ■ 일반 회원



가설 검증

∴ 'pop_product' 변수 추가

가설과 반대로,

일반 회원이 인기 상품을 구매하는 경향이 있다.

| 임직원 | 인기 상품 구매 | 프라임 회원 비율 |
|-------|----------|-----------|
| CJ 직원 | O | 59% 41% |
| | X | 63% 37% |
| 일반 회원 | O | 54% 43% |
| | X | 49% 51% |

가설 검증 6 & 변수 정리

프라임 회원의 최종 구매 금액이 일반 회원보다 높을 것이다

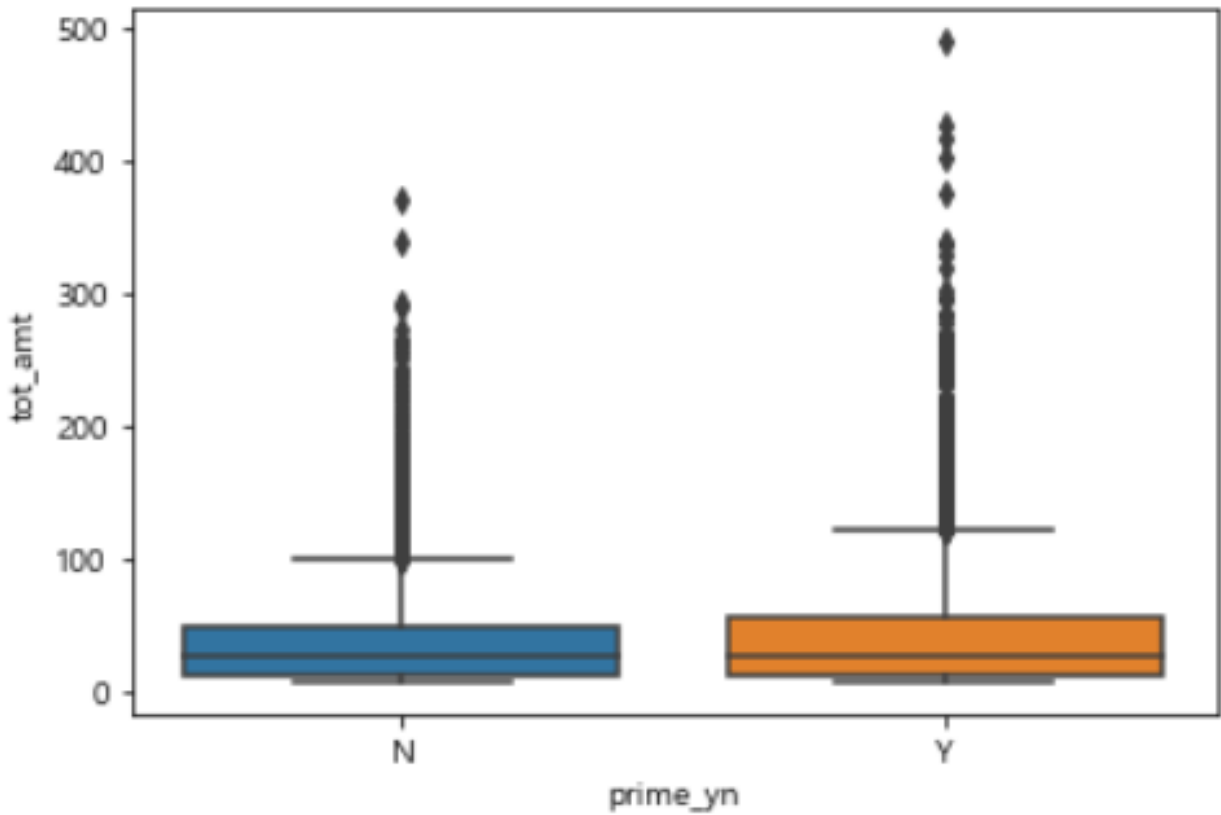
- 최종 구매 금액 도출

동일 주문 번호끼리 더하여 구매 금액의 총합 계산 *'tot_amt'* 변수 추가

| 프라임 | 최소 구매 금액 | 최대 구매 금액 |
|--------|----------|----------|
| 프라임 회원 | 7.11 | 488.5 |
| 일반 회원 | 6.55 | 369.4 |

차이가 큼

프라임 회원 구매 금액이 더 높음



✓ 가설 검증 과정에서 생성된 4가지 변수

1. *is_price_lower_than_avg*

3. *pop_product*
2. *tot_qty_9*

4. *tot_amt*



02

데이터 전처리

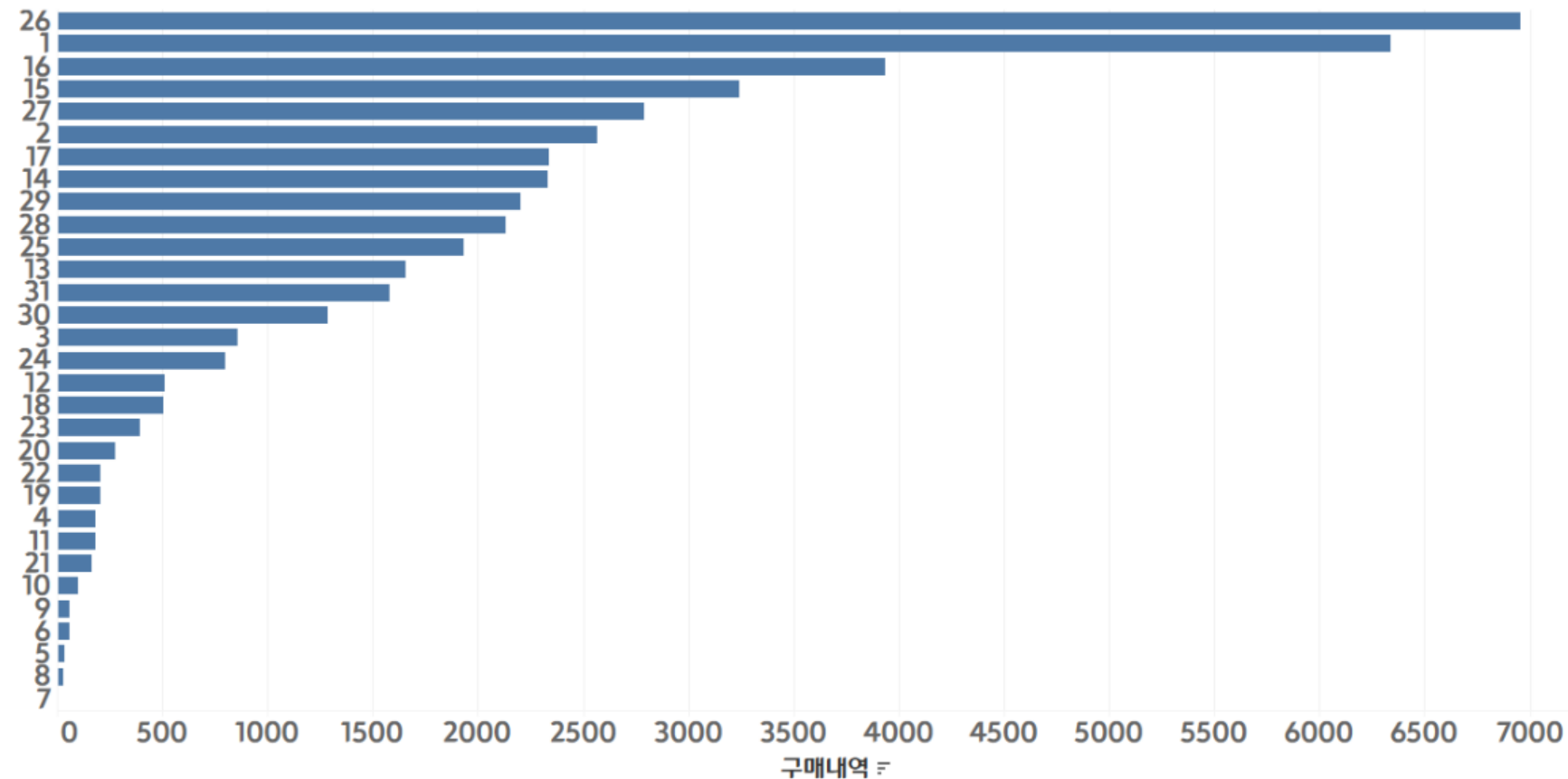
추가 파생 변수 생성
변수 별 전처리

날짜 파생 변수 생성

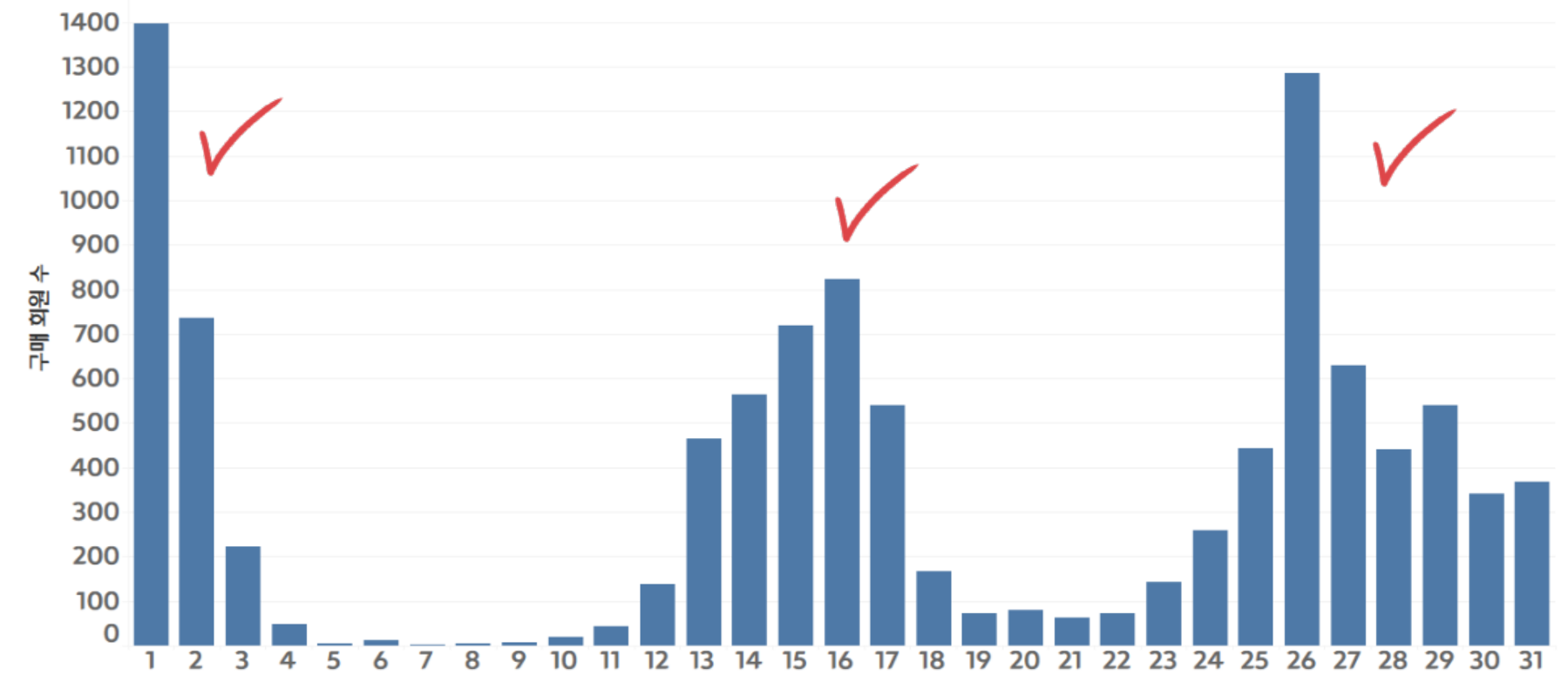
날짜 관련 추가 분석 진행

- 기존 'order_date' 변수 분해
 - (1) $(20230101+4)\%31 = 1$ 을 활용하여 'date' 변수 생성
 - (2) $\text{date} \% 7$ 을 통해 0~6까지 분류한 'date_of_week' 변수 생성
 - 1 : 일 , 2 : 월 , 3 : 화 , 4 : 수 , 5 : 목 , 6 : 금 , 0 : 토
- * 기존 order_date 변수 삭제

<일자 별 구매 빈도>



<일자 별 구매 회원 수>

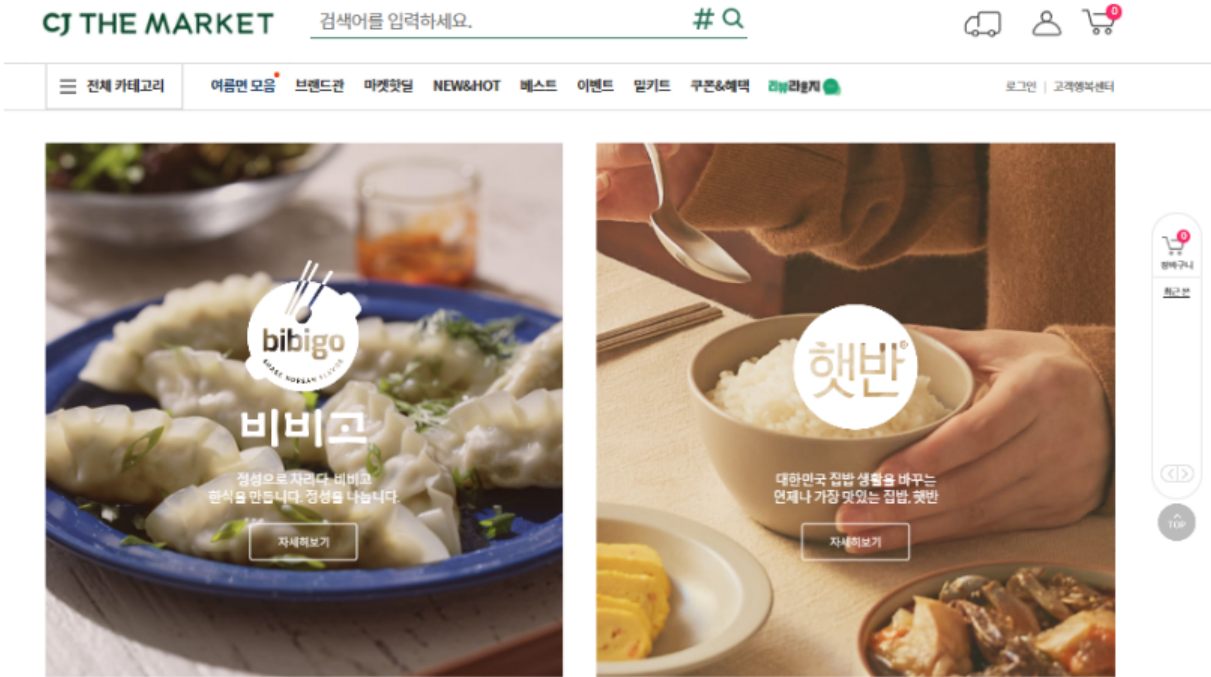


- 1월 간 3번의 상승 하락 주기 확인
 - 1/1 ~ 1/2, 1/13 ~ 1/17, 1/25 ~ 1/31
 - ∴ 구매량이 압도적으로 많은 날짜 파생 변수 생성
- 1월에 존재하는 공휴일 변수 생성
 - 설날 혹은 새해 첫날 (1/1)

상품명 파생 변수 생성

브랜드 및 카테고리 변수 생성

Task 1.
홈페이지에서
브랜드 크롤링



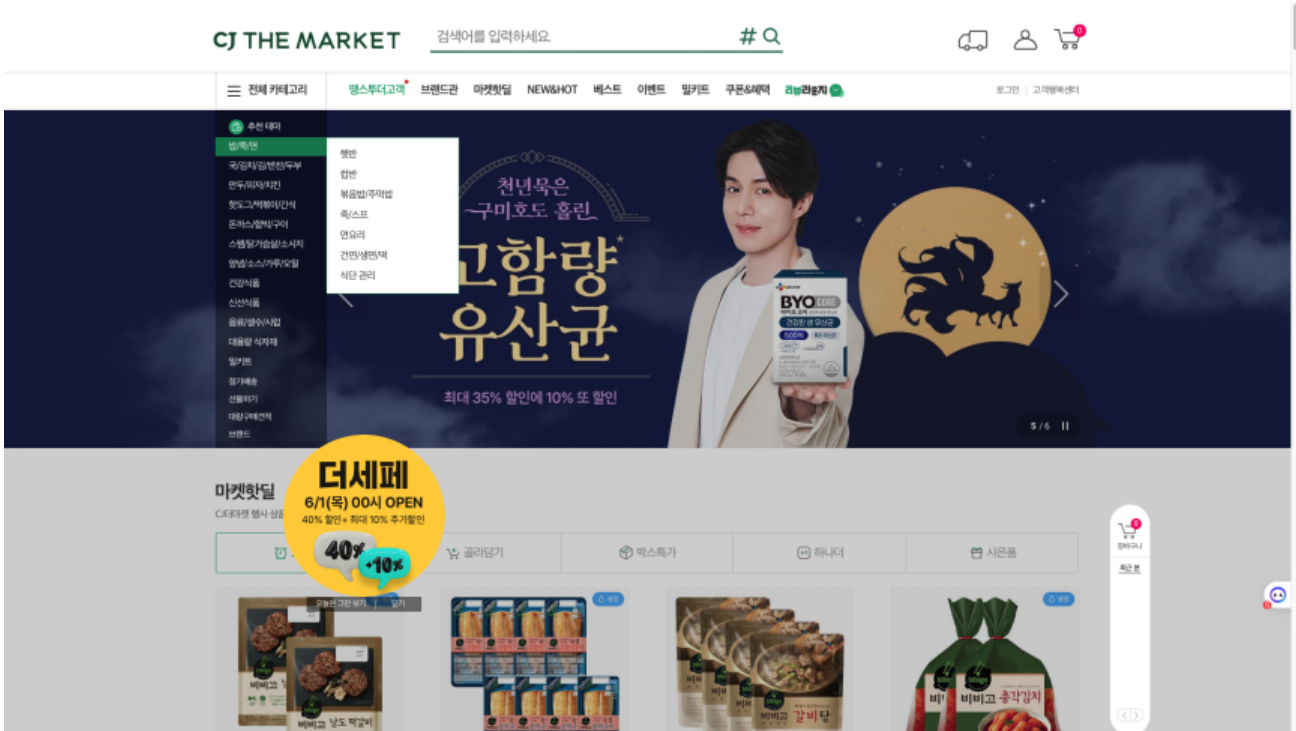
크롤링한 브랜드를 이용해
브랜드별 구매 빈도 수가 특정값 이상이면 '인기 브랜드' 로 정의

∴ 인기 브랜드 구매 여부 관련 더미 변수 생성

크롤링한 카테고리 별 category dictionary 제작

밥 | 찬 | 점심 | 스낵 | 돈 | 햄 | 소스 | 건강 | 신선 | 음료

∴ 카테고리별 상품 주문 수량 나타내는 변수 생성



Task 2.
상품 분류
카테고리 크롤링

상품명 파생 변수 생성

이벤트 관련 변수 생성

Step 01

정규표현식 활용 대괄호로 묶여있는
이벤트 정보 추출

- 임직원 / 일반 회원 분리하여 개별 적용 이벤트 확인

Step 03

할인 / 한정 행사 파생 변수 추가

- 'sale' - 증정 | 할인 | 특가 | ONLY | 침착맨
'limit' - 한정

Step 05

임직원 / 일반 회원 대상 이벤트 리스트
각각 분리된 데이터셋에 파생 변수로 추가

Step 02

임직원 / 일반 회원 분리
거래수 100건 이상 행사 변수 생성

Step 04

임직원 대상 행사 별도 분리하여
파생 변수 생성

- 'em_e' - 패밀리데이 | 행사 | 오프

Event

행사 내용 추출

기타 변수 생성 및 전처리

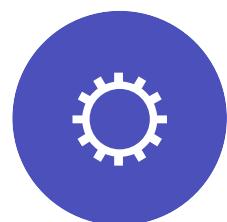
추가 변수 생성



- 상품명에서 'box' 검출하여 **'box'** 파생 변수 생성



- 여러 개를 묶어 한 번에 묶어 판매하는 상품을 검출하는 **'multi'** 변수
→ 상품명 n (개, 인분, 번들, ea, 입) 포함



- 최종 구매 상품 종류 개수를 구분하는 **'product_cnt'** 변수
→ 한 가지 상품이 여러 개 포함 될 수 있기 때문



- 한 가지 상품에 대한 구매 수량이 37개 이상인 주문 건을 검출하는 **'bulk'** 변수
→ 37 이상 데이터들은 2023 설선물 세트와 유관

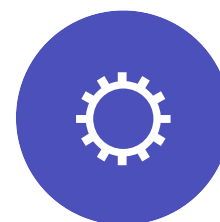
기타 전처리



- 형식이 다른 **상품명 표기 형식 통일화**
→ 상품명과 수량/ 금액으로 비교하여 통일



- **boolean형** 변수 주문번호로 그룹화하여 **sum** 적용
→ 이벤트 상품 / 박스 상품 등 구매 건수 도출



- 모델링을 위해 **연령대 / 요일** 변수 **더미 변수**로 변형
→ True / False 혹은 Y / N 나타내는 데이터 1,0 형식 통일



- scd 별 / 행별 파생변수 종합된 데이터 프레임 생성하여 **train, test셋 분리**
→ 임직원 유무에 따라 데이터 셋 분리

03

모델링

모델링에 사용한 방식
분석 결과 해석

Modeling

사용 라이브러리 소개



- 머신 러닝 워크플로우를 자동화하는 오픈 소스 라이브러리

: 분류, 회귀, 클러스터링 등 다양한 Task에서 사용하는 모델들을 동일한 환경에서 한 줄의 코드로 실행

| | Description | Value |
|----|-----------------------------|------------|
| 0 | Session id | 6473 |
| 1 | Target | prime_yn |
| 2 | Target type | Binary |
| 3 | Original data shape | (4713, 53) |
| 4 | Transformed data shape | (4713, 53) |
| 5 | Transformed train set shape | (3299, 53) |
| 6 | Transformed test set shape | (1414, 53) |
| 7 | Ordinal features | 19 |
| 8 | Numeric features | 33 |
| 9 | Categorical features | 19 |
| 10 | Preprocess | True |

Preprocess
중점



Baseline Code

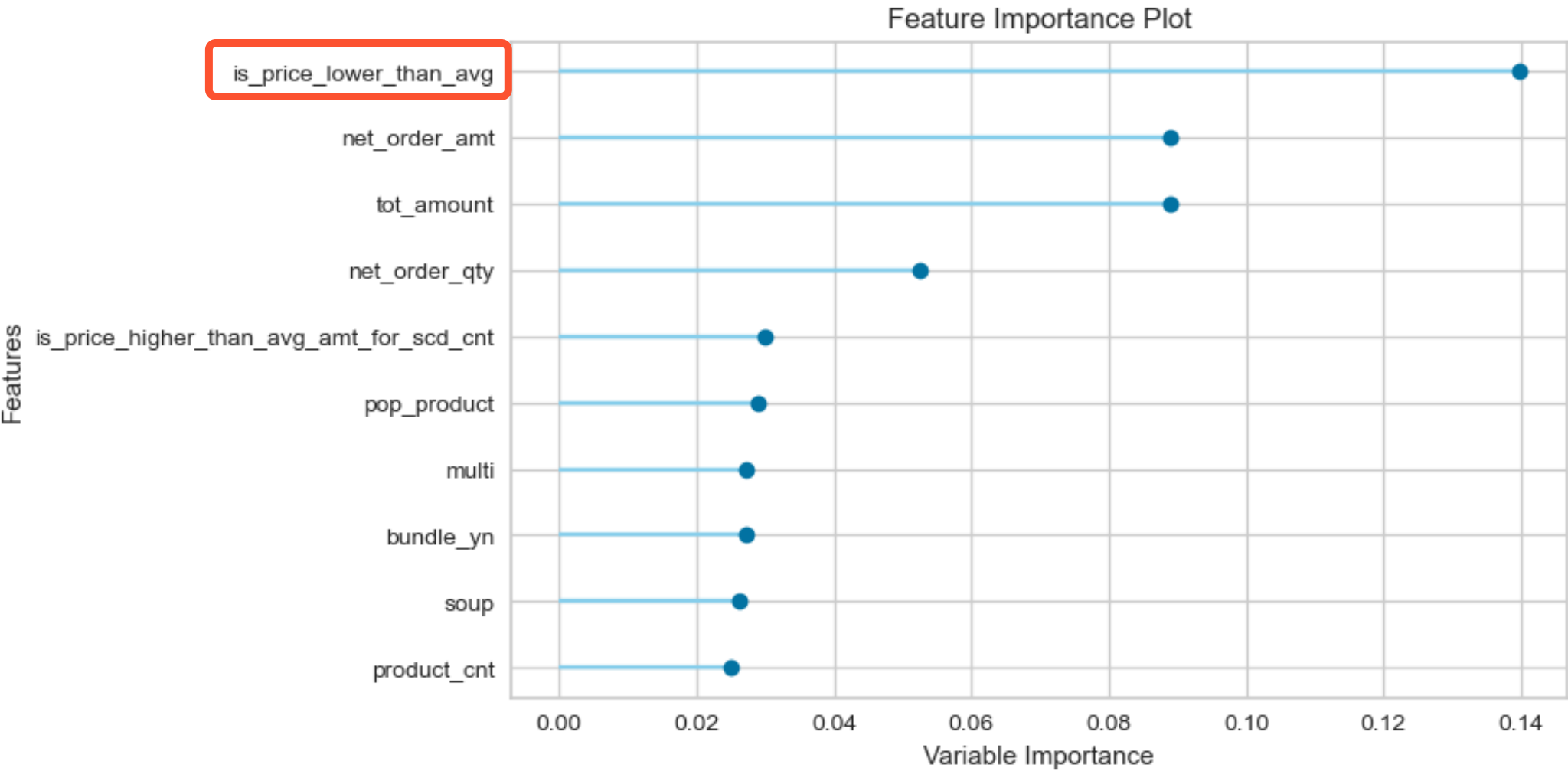
- Random Forest

분류에 널리 사용되는 의사결정 나무의 과적합 한계를 극복하기 위한 앙상블 모델

- Baseline Model 성능 확인

| | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|--|----------|--------|--------|--------|--------|--------|--------|
| click to scroll output; double click to hide | | | | | | | |
| 0 | 0.7452 | 0.8035 | 0.6954 | 0.6954 | 0.6954 | 0.4764 | 0.4764 |
| 1 | 0.7236 | 0.7883 | 0.6114 | 0.6948 | 0.6505 | 0.4234 | 0.4257 |
| 2 | 0.6947 | 0.7792 | 0.5543 | 0.6644 | 0.6044 | 0.3591 | 0.3630 |
| 3 | 0.7260 | 0.8028 | 0.5943 | 0.7075 | 0.6460 | 0.4252 | 0.4295 |
| 4 | 0.7452 | 0.8171 | 0.6171 | 0.7347 | 0.6708 | 0.4655 | 0.4702 |
| 5 | 0.7139 | 0.7761 | 0.5657 | 0.6972 | 0.6246 | 0.3976 | 0.4032 |
| 6 | 0.7620 | 0.8402 | 0.6229 | 0.7676 | 0.6877 | 0.4988 | 0.5059 |
| 7 | 0.7236 | 0.8081 | 0.5771 | 0.7113 | 0.6372 | 0.4178 | 0.4238 |
| 8 | 0.7349 | 0.7861 | 0.6437 | 0.7000 | 0.6707 | 0.4495 | 0.4506 |
| 9 | 0.7084 | 0.7916 | 0.5747 | 0.6803 | 0.6231 | 0.3881 | 0.3917 |
| Mean | 0.7278 | 0.7993 | 0.6057 | 0.7053 | 0.6510 | 0.4301 | 0.4340 |
| Std | 0.0188 | 0.0184 | 0.0401 | 0.0272 | 0.0282 | 0.0407 | 0.0406 |

- 랜덤 포레스트 모델에서의 특징 중요도



임직원 데이터 모델링 분석 결과

- F1 Score 기준 상위 5개 모델

| Model | Accuracy | AUC | Recall | F1 |
|------------------------------|----------|--------|--------|--------|
| XGBoost | 0.7760 | 0.8524 | 0.8178 | 0.8144 |
| Gradient Boosting Classifier | 0.7754 | 0.8572 | 0.7905 | 0.8086 |
| Light GBM | 0.7693 | 0.8544 | 0.8087 | 0.8081 |
| Random Forest | 0.7596 | 0.8413 | 0.8072 | 0.8013 |
| Logistic Regression | 0.7545 | 0.8222 | 0.7638 | 0.7887 |

일반 직원 데이터 모델링 분석 결과

• F1 Score 기준 상위 5개 모델

| Model | Accuracy | AUC | Recall | F1 |
|------------------------------|----------|--------|--------|--------|
| Light GBM | 0.7513 | 0.8362 | 0.6780 | 0.6931 |
| XGBoost | 0.7422 | 0.8308 | 0.6634 | 0.6832 |
| Gradient Boosting Classifier | 0.7467 | 0.8305 | 0.6486 | 0.6826 |
| LDA | 0.7275 | 0.7817 | 0.6686 | 0.6731 |
| Ridge Classifier | 0.7268 | 0.0000 | 0.6657 | 0.6715 |

상위 5개 모델 블렌딩 모델링 분석 결과

• 임직원 데이터 예측

| | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|------|----------|--------|--------|--------|--------|--------|--------|
| Fold | | | | | | | |
| 0 | 0.7848 | 0.8642 | 0.7904 | 0.8414 | 0.8151 | 0.5585 | 0.5600 |
| 1 | 0.7697 | 0.8480 | 0.8283 | 0.7961 | 0.8119 | 0.5153 | 0.5160 |
| 2 | 0.7955 | 0.8724 | 0.8081 | 0.8443 | 0.8258 | 0.5784 | 0.5792 |
| 3 | 0.7788 | 0.8663 | 0.8161 | 0.8161 | 0.8161 | 0.5386 | 0.5386 |
| 4 | 0.7602 | 0.8462 | 0.7828 | 0.8115 | 0.7969 | 0.5046 | 0.5050 |
| Mean | 0.7778 | 0.8594 | 0.8051 | 0.8219 | 0.8132 | 0.5391 | 0.5398 |
| Std | 0.0121 | 0.0104 | 0.0166 | 0.0184 | 0.0094 | 0.0271 | 0.0273 |

• 일반 회원 데이터 예측

| | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|------|----------|--------|--------|--------|--------|--------|--------|
| Fold | | | | | | | |
| 0 | 0.7596 | 0.0000 | 0.6800 | 0.7301 | 0.7041 | 0.5021 | 0.5030 |
| 1 | 0.7728 | 0.0000 | 0.7114 | 0.7389 | 0.7249 | 0.5316 | 0.5318 |
| 2 | 0.7344 | 0.0000 | 0.6447 | 0.6988 | 0.6706 | 0.4487 | 0.4497 |
| 3 | 0.7653 | 0.0000 | 0.6934 | 0.7333 | 0.7128 | 0.5147 | 0.5153 |
| 4 | 0.7341 | 0.0000 | 0.6304 | 0.7051 | 0.6657 | 0.4460 | 0.4480 |
| Mean | 0.7532 | 0.0000 | 0.6720 | 0.7212 | 0.6956 | 0.4886 | 0.4896 |
| Std | 0.0161 | 0.0000 | 0.0302 | 0.0161 | 0.0234 | 0.0350 | 0.0345 |



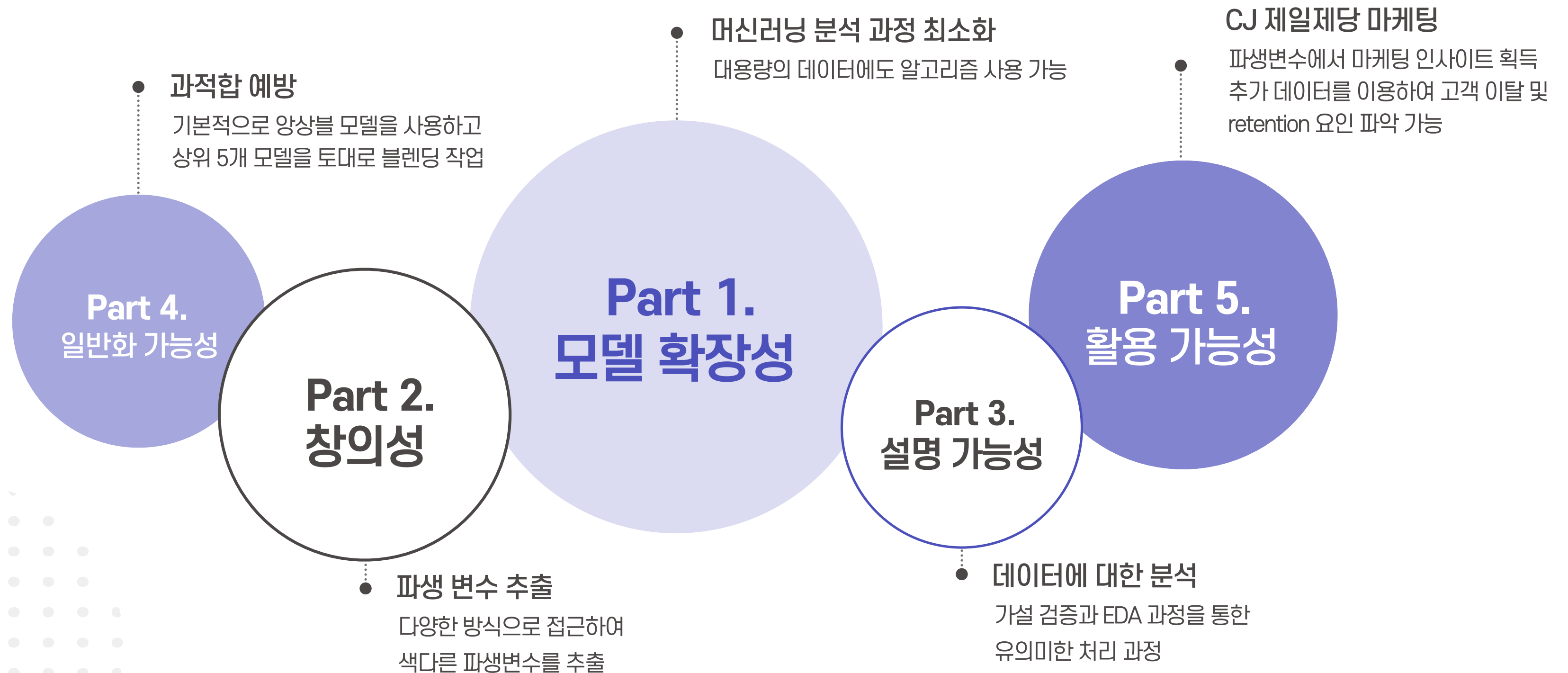
04

의의 및 한계점

모델링이 가지는 의미와 예상 활용 방안
한계점 & 보완사항

프라모델 조의 강점

모델링이 가지는 의미와 예상 활용 방안



한계점과 보완사항



- 일반 회원 데이터에서는 프라임 예측을 위한 뚜렷한 특징을 찾지 못 했음
→ 임직원에 비해 낮은 예측 결과로 이어짐
- 초기에 계획한 전처리 사항 중 시간 문제로 분석하지 못한 내용이 있음
→ 무게, 배송 기한 등에 대해 처리하지 못함

- 고객 식별 정보, 프라임 회원 지속 기간 등의 데이터가 추가 되거나 장기간의 데이터를 사용한다면 보완할 수 있을 것으로 예상됨
- 추가적인 시간이 주어진다면 분석해볼만 한 가치가 있을 것으로 예상됨