



Walmart 고객 세분화 전략

2026년 2월



- 분석 목표 - 목적, 배경, 필요성
- 데이터 설명 : 출처, 특성 및 구조
- 데이터 분석 개요 : 상세코드나 전처리 과정을 자세히 적지 않는다.
 - 전처리 과정 (결측치, 이상치 제거 등) 설명
 - 분석 결과에 영향을 미친 중요한 데이터 저처리 과정을 자세히 기재한다.
 - 데이터 분석 프로세스 및 아키텍처 관점의 도면 작성
- EDA : 주요 탐색 내용 요약 & 시각화 자료 포함
- 주요 결과
 - 분석 결과 요약
 - 인사이트 및 제안 사항
- 결론 : 전체적인 결론 및 향후 방향
- 질문과 답변

1. 분석목표



본 분석의 목적은

월마트 고객 구매 데이터를 기반으로 고객을 행동 중심으로 세그멘테이션하고,
매출 성장을 위해 가장 효과적으로 공략해야 할 마케팅 타겟을 도출하는 것이다.

기존의 인구통계 기반 타겟팅은 고객의 실제 구매 행동을 충분히 설명하지 못하며,
특히 상품 가격대와 소비 여력 간의 관계를 정량적으로 반영하지 못하는 한계가 있다.

이에 본 분석은 **구매 행동(Purchase)**을 중심으로 **고객의 소비 성향을 재정의** 하고,
이를 **마케팅 실행 관점(Defend / Grow / Expand)**으로 연결하는 데 초점을 두었다.



2. 데이터 출처, 특성 및 구조

본 데이터는 Kaggle에 공개된 **Walmart 고객 구매 데이터**로,
개별 구매 트랜잭션 단위로 고객 정보와 상품 정보, 구매 금액을 포함하고 있다.

주요 데이터 구조

- **고객 정보**: 성별(Gender), 연령대(Age), 결혼 여부(Marital_Status), 직업 코드(Occupation)
- **상품 정보**: Product_Category
- **거래 정보**: Purchase (구매 금액)

특히 **Occupation**, **Product_Category**는 실제 의미를 갖는 값이 아닌 **익명화된 범주형 인덱스**로,
해당 값 자체에는 서열이나 크기 의미가 존재하지 않는다.

10개 Column과 55068 row(index)로 이루어진 **DataFrame**이고, 결측치가 없다.



2. 데이터 출처, 특성 및 구조

변수 목록과 타입 확인

- User_ID: 식별자
 - Product_ID: 식별자
 - Gender: 범주형 (Nominal)
 - Age: 범주형 (Ordinal)
 - Marital_Status: 범주형 (이진: 0 또는 1)
 - City_Category: 범주형 (Nominal)
 - Stay_In_Current_City_Years: 범주형 (Ordinal)
 - Occupation: 범주형 (Nominal)
 - Product_Category: 범주형 (Nominal)
-
- Purchase: 수치형 (연속형)



범주형 데이터의 기초 통계량 (**Count, Ratio, Cum_Ratio**)
확인 및 히스토그램 분포 시각화 작업 진행



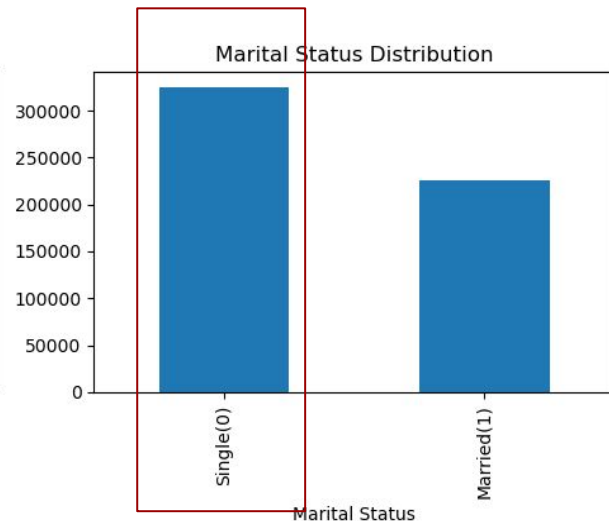
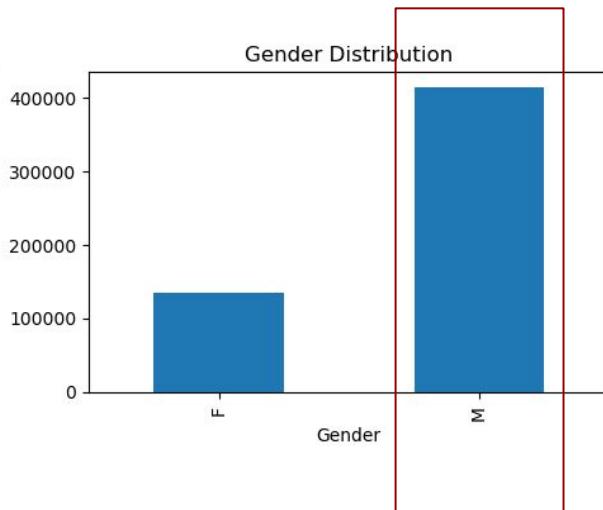
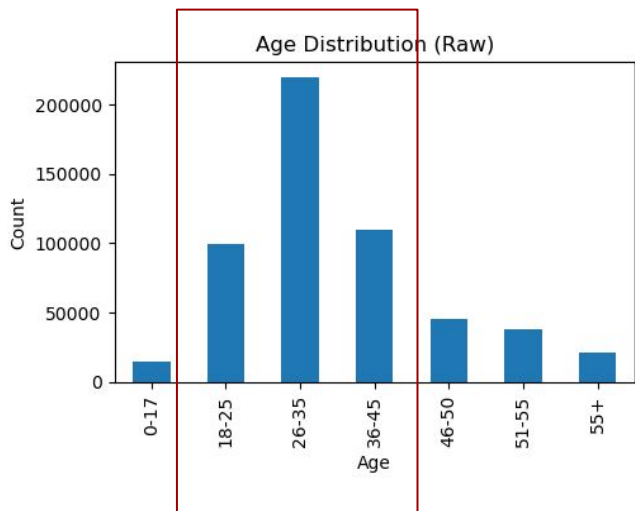
수치형 데이터는 이상치 확인을 위해 **Boxplot** 시각화 작업
진행

2. 데이터 출처, 특성 및 구조



Demography 특성:

남(75.3%), 미혼(59.0%), 연령 26-35(39.9%) > 36-45(19.9%) > 18-25(18.1%)로
청장년층의 남자 미혼 고객이 높은 분포를 보이고 있음

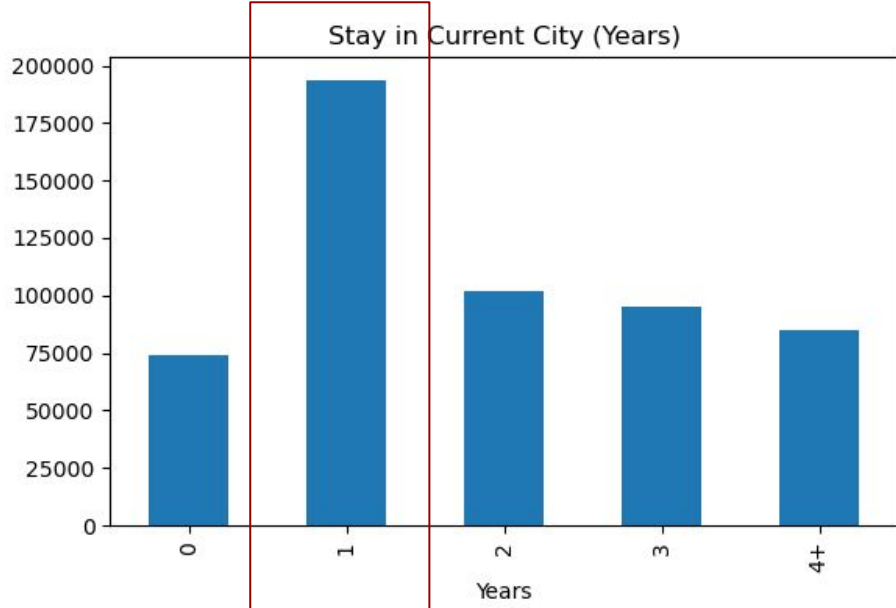
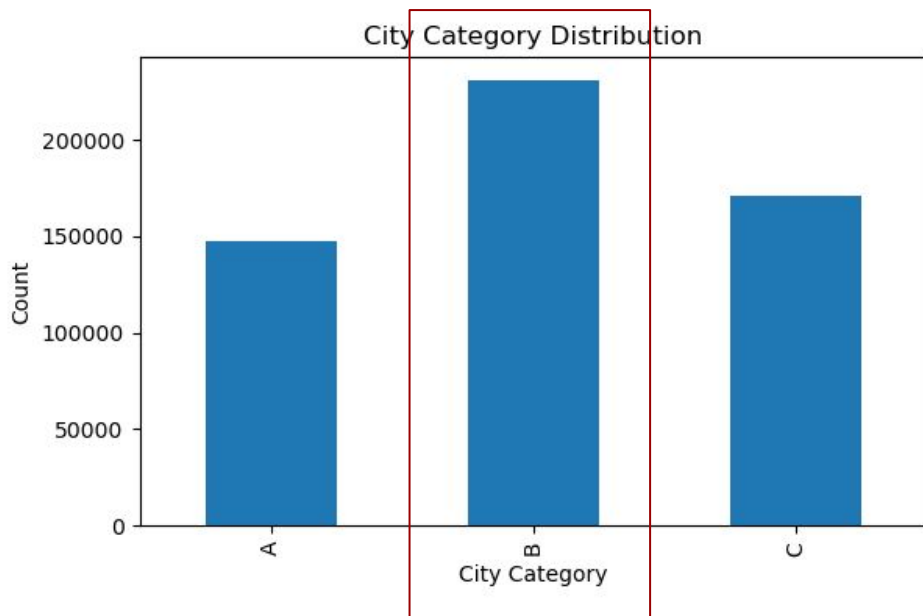


2. 데이터 출처, 특성 및 구조



거주지 특성:

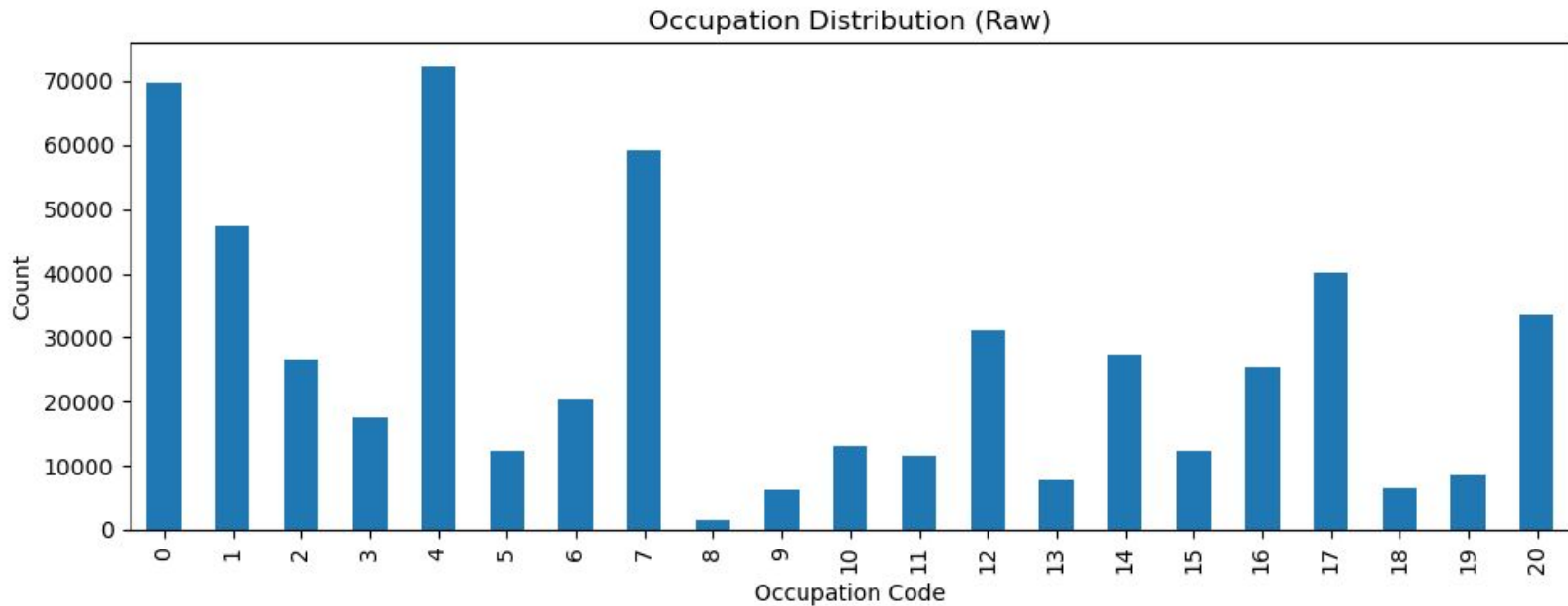
City Category Distribution(고객 생활권)은 B(중간규모도시) 42.0% > C(소도시) 31.1% > A(대도시) 26.9% 순이며, 현재 도시 거주 기간은 1년이 35.2%로 가장 높게 나타남



2. 데이터 출처, 특성 및 구조



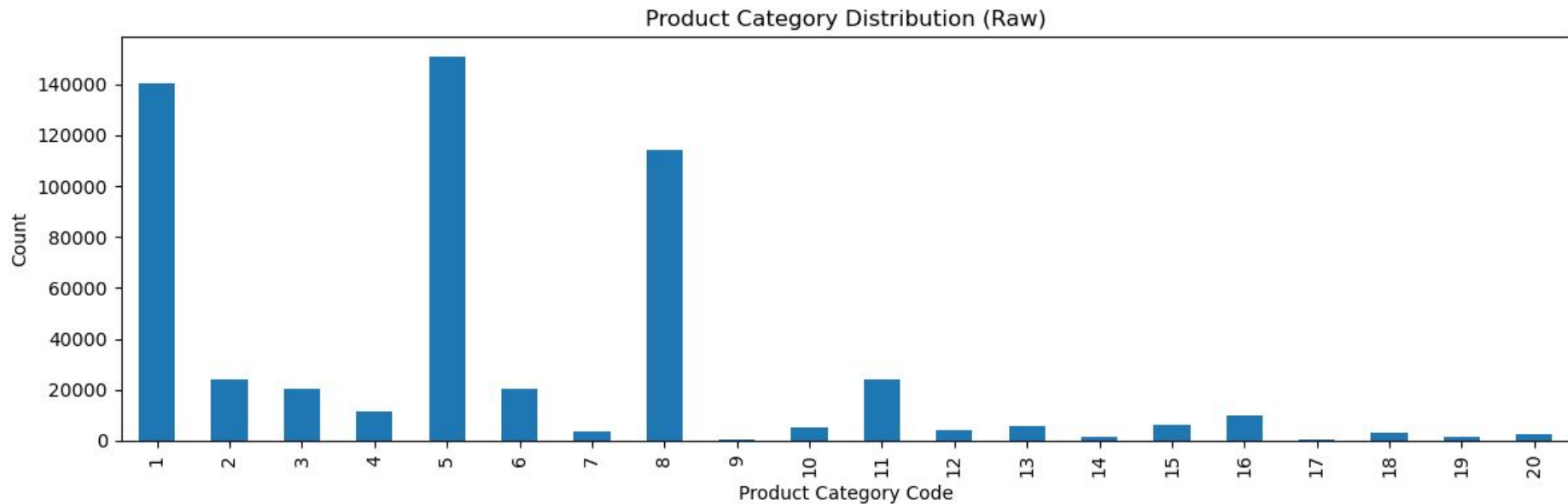
익명화된 범주형 인덱스 1. (Occupation)



2. 데이터 출처, 특성 및 구조



익명화된 범주형 인덱스 2. (Product_Category)



2. 데이터 출처, 특성 및 구조

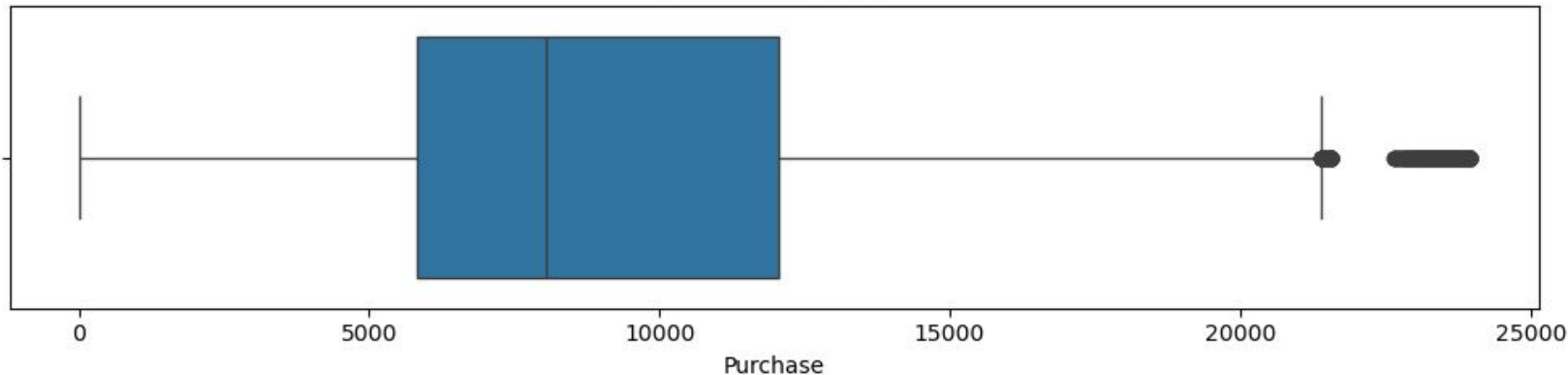


Purchase:

수치형 데이터 박스플롯 시각화

- 중앙값: 8,000~9,000
- 1Q~3Q 범위: 5,500~12,000 (금액 분산이 큼)
- 하단 수염: 0 근처 -> 음수 없음(데이터 이상 없음)
- 상단 수염: 21,000~22,000 (정상범위 내에서도 고가 구매 꽤 많은 것으로 보임)
- 이상치: 22,000~24,000 부근에 점이 많음 -> 이상치 여부 확인 필요

Purchase Boxplot



3. 데이터 분석 개요



본 분석은 다음의 흐름으로 진행되었다.

1. 데이터 구조 및 분포 파악
2. 구매 금액(Purchase)의 분포 특성 분석
3. 구매 행동 기반 파생 변수 생성
4. 고객 세그멘테이션 설계
5. 세그먼트별 매출 기여도 및 성장 가능성 평가
6. 마케팅 관점 타겟 선정 및 세부 프로파일링

상세 코드 구현보다는 의사결정에 영향을 주는 데이터 처리와 구조적 판단에 집중하였다.

4. 전처리 과정



결측치

- 핵심 분석 변수(Purchase, Gender, Age, Occupation 등)에 치명적인 결측치는 존재하지 않아, 별도의 제거 없이 분석을 진행하였다.

이상치

- Purchase 분포 분석 결과, 극단적으로 높은 값들이 존재했으나 이는 오류 데이터가 아닌 **고가 상품 구매라는 실제 고객 행동**으로 판단하였다.
- 따라서 이상치를 제거하지 않고, **분포 왜곡을 완화하는 방식**으로 접근하였다.

5. 분석 결과에 영향을 미친 주요 데이터 전처리 과정



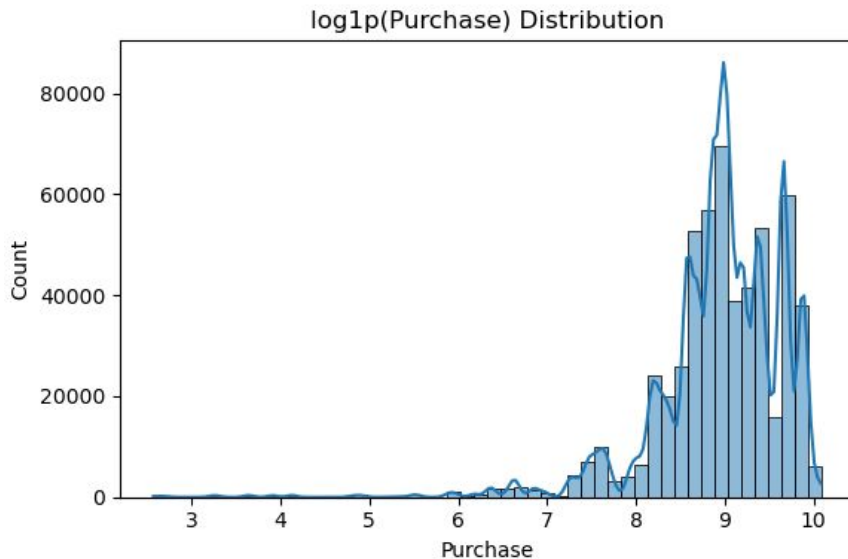
1) Purchase의 로그 변환 ($\log_{10}(\text{Purchase})$)

Purchase는 강한 양의 왜도를 가진 분포였으며, 소수의 고가 구매가 평균과 모델링에 과도한 영향을 주는 구조였다.

이에 구매 금액을 제거하거나 자르지 않고, $\log_{10}(\text{Purchase})$ 변환을 통해:

- 극단값의 영향 완화
- 구매 행동 간 상대적 차이 유지
- 안정적인 비교 기준 확보

를 달성하였다.



5. 분석 결과에 영향을 미친 주요 데이터 전처리 과정



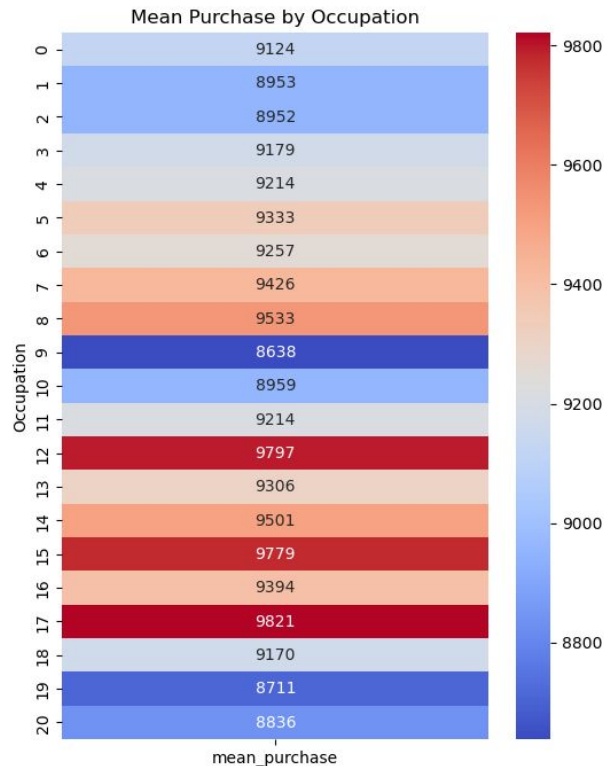
2) Occupation 재정의 (소비 행동 기반)

기존 Occupation 인덱스는 의미 없는 코드이므로,
`log1p(Purchase)` 평균을 기준으로:

- 저소비 (Occ_Low)
- 중소비 (Occ_Mid)
- 고소비 (Occ_High)
- 희소 그룹 (Occ_Other)

으로 재구성하였다.

→ 직업 코드 → 소비 성향 세그먼트로 의미를 변환



5. 분석 결과에 영향을 미친 주요 데이터 전처리 과정



3) Product_Category → Price_Segment 재설계

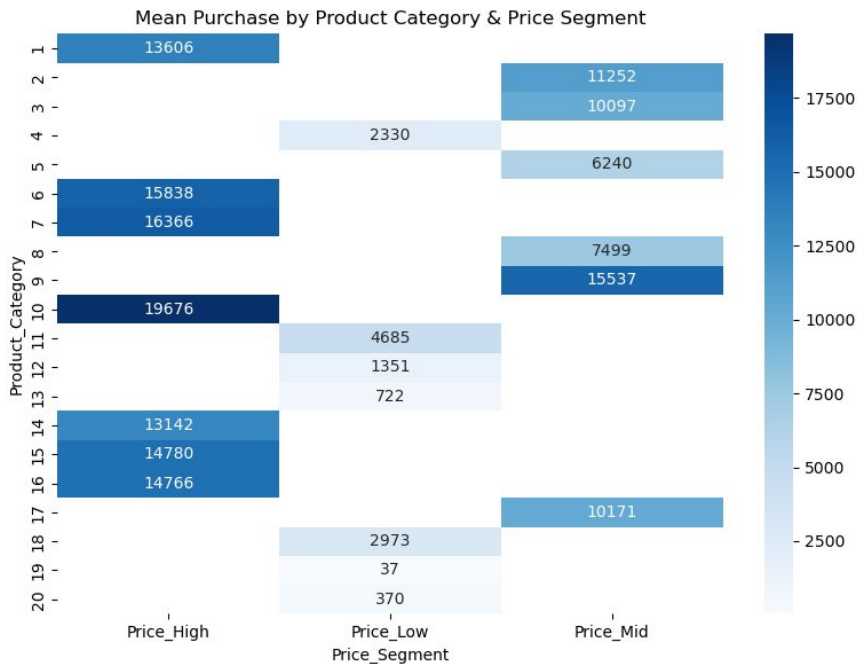
Product_Category는 병합 시에도 인덱스 수가 과도하게 많아,
차원 축소 및 해석 측면에서 비효율적이었다.

이에 Product_Category를 직접 사용하지 않고,
카테고리별 구매 금액 중앙값을 기준으로:

- Price_Low
- Price_Mid
- Price_High

의 **가격대 세그먼트(Price_Segment)**로 재정의하였다.

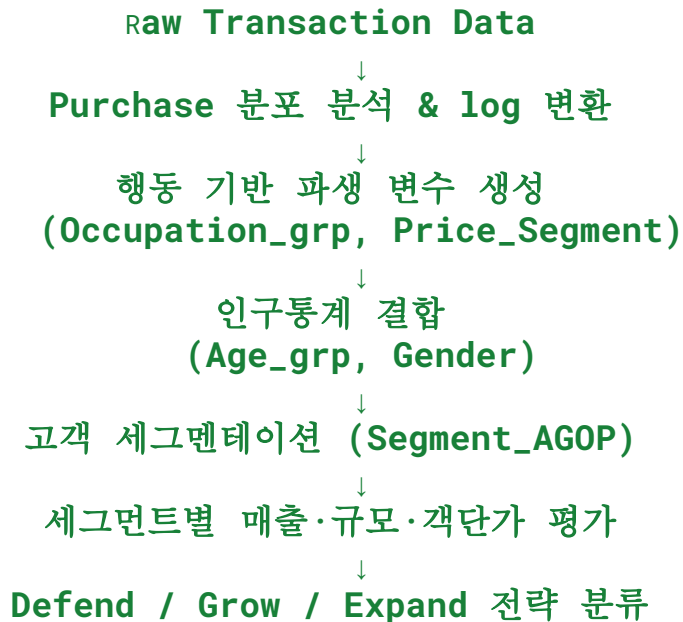
→ 상품 코드 → 가격 수용도 지표로 전환



6. 데이터 분석 프로세스 및 아키텍처 관점 설명



분석 프로세스는 다음 구조를 따른다.



이 구조는 데이터 처리 → 분석 → 마케팅 실행으로 자연스럽게 이어지도록 설계되었다.

7. EDA



주요 탐색 결과

- Purchase는 다봉 분포를 가지며, 가격대별 구매 행동이 뚜렷하게 구분됨
- 로그 변환 후 분포의 안정성이 크게 개선됨
- Price_Segment별 구매 금액은 명확한 계층 구조를 보임
- 특정 연령·성별·소비 성향 조합에서 고가 상품 구매 집중 현상 확인

EDA는 이후 세그멘테이션 설계의 근거로 활용되었다.



8. 고객 Segmentation

고객 세그먼트

활용변수

- 인구통계 : Age_Grp, Gender
- 경제력/행동: Occupation_grp
- 상품전략: Price_Segment
- 타겟지표: Purchase

순서

세그먼트 도출 >> 세그먼트별 매출 기여도 도출 >> 스코어링 >>
타겟 후보 결정 >> 타겟 프로파일링

8. 고객 Segmentation



세그먼트 도출

	bucket	Segment_AGOP	customers	transactions	revenue	revenue_share	avg_purchase	target_score
48	Defend	26-35 M Occ_High Price_High	533	20242	286821908	0.056286	14169.642723	0.745101
72	Defend	36-45 M Occ_High Price_High	294	10646	154307579	0.030281	14494.418467	0.492667
96	Defend	46+ M Occ_High Price_High	299	8691	127101817	0.024942	14624.533080	0.465017
27	Defend	18-25 M Occ_High Price_High	148	5487	78243337	0.015354	14259.766175	0.337756
39	Defend	26-35 F Occ_High Price_High	93	2487	35189600	0.006906	14149.416968	0.262285
55	Expand	26-35 M Occ_Mid Price_Low	679	7573	24470690	0.004802	3231.307276	0.340274
49	Expand	26-35 M Occ_High Price_Low	519	4612	14773881	0.002899	3203.356678	0.259786
34	Expand	18-25 M Occ_Mid Price_Low	454	4577	14389016	0.002824	3143.765785	0.230520
103	Expand	46+ M Occ_Mid Price_Low	437	4330	13228676	0.002596	3055.121478	0.220550
79	Expand	36-45 M Occ_Mid Price_Low	375	4149	13373469	0.002624	3223.299349	0.196557
56	Grow	26-35 M Occ_Mid Price_Mid	698	43925	314016392	0.061622	7148.921844	0.743957
35	Grow	18-25 M Occ_Mid Price_Mid	487	26919	196099930	0.038483	7284.814815	0.517786
104	Grow	46+ M Occ_Mid Price_Mid	460	20104	152676346	0.029961	7594.326801	0.460124
80	Grow	36-45 M Occ_Mid Price_Mid	384	20921	154301182	0.030280	7375.420965	0.426119
47	Grow	26-35 F Occ_Mid Price_Mid	306	16860	124142191	0.024362	7363.119276	0.357247

9. 주요 결과



1) 최우선 핵심 타겟(Defend)

- 26-35세 남성 · 고소비 직업군 · 고가 상품 구매 고객
 - 높은 객단가와 큰 총매출(Gross Sales)을 동시에 보유
 - 이탈 방지 및 프리미엄 강화 필요
-

2) 성장 핵심 타겟(Grow)

- 중소비 성향 · 중가 상품 중심 고객
 - 규모와 전환 가능성이 모두 큼
 - 업셀링 전략 최적 대상
-

3) 시급 개선 타겟(Expand)

- 소비 여력은 있으나 저가 상품에 고착된 대규모 고객군
- 가격 장벽이 아닌 가치 인식 부족이 주요 원인



9. 분석 결과 요약

고객의 실제 구매 행동을 기준으로 한 세그멘테이션은
기존 인구통계 기반 접근보다 훨씬 높은 설명력을 가짐

가격대 세그먼트는 상품 카테고리보다
매출과 직결되는 핵심 변수로 작용

고객은 단일 타겟이 아니라
유지·성장·확장 전략이 다른 집단으로 구분되어야 함

10. 인사이트 및 제안 사항



Defend 타겟

- 할인보다 경험·프리미엄 강화
- VIP 관리 전략 필요

Grow 타겟

- 번들·업셀 중심 캠페인
- 중가 → 고가 전환 유도

Expand 타겟

- 엔트리 프리미엄
- 체험·보증 기반 메시지 설계
- KPI는 매출이 아닌 가격대 이동을

11. 결론



본 분석은 고객을 ‘누구인가’가 아니라 ‘어떻게 소비하는가’로
재정의 함으로써,
월마트의 매출 성장을 위한 실질적인 타겟 전략을 도출하였다.

향후에는:

- 세그먼트별 캠페인 실험(A/B 테스트)
- 전환율 기반 성과 측정
- 장기 고객 가치(LTV) 분석

으로 확장 가능하다.



12. 질문과 답변

- 왜 Product_Category를 제거했는가?
- Grow와 Expand의 차이는 무엇인가?
- 할인 전략은 어디에 써야 하는가?

→ 본 분석 결과는 위 질문들에 대한 **데이터 기반 근거**를 제공한다.