



월마트 데이터 분석

2026.02-10

KPMG Future Academy 8기

박순선

목차



1. 분석 목표
2. 데이터 설명
3. 데이터 분석 개요
4. EDA
5. 주요 결과
6. 결론
7. 질문과 답변

1. 분석 목표 - 목적, 배경, 필요성



목적: 월마트의 고객 데이터를 통해 핵심 고객군을 추출하여 핵심 타겟을 위한 전략 수립

목표: 남녀별, 연령대별 소비패턴이 다를 것으로 예상되어 이들을 기준으로
의미있는 시각화 자료를 도출한다.

배경: 월마트를 방문하는 고객들 중 핵심 고객을 알지 못하는 상태

필요성: 매출의 많은 부분을 차지하는 핵심 고객군을 알아내어 그들을 만족시키기 위한 제품군부
제품배열, 마케팅 전략까지 함께 재점검할 필요있음

2. 데이터 설명



데이터 출처 : Kaggle에서 매우 유명한 "Black Friday Sales" 데이터셋

- 2013년의 블랙 프라이데이(Black Friday) 기간 동안 발생한 거래 내역
- Analytics Vidhya의 해커톤을 위해 공개되었다가 Kaggle에 공유되었음
- 소비자 구매 패턴을 분석하기 위해 만들어졌으며 사용자ID와 제품ID는 비식별화 처리됨
- 약 13년전의 데이터 이므로 오늘날의 소비패턴과는 차이가 있다.

해당 경로 : <https://www.kaggle.com/code/akkyjaicar/walmart-case-study-clt-ci>

특성 및 구조 : 데이터 프레임에서 수치형은 구매액 1개, 나머지는 모두 범주형이다.

데이터 프레임 상에서 수치로 표현되어 있지만, 그것은 비식별화 처리하거나 분류코드로 표현되었다.

2. 데이터 설명: 구매액만 수치형, 나머지는 범주형이다.

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	7969

User_ID: 비식별화 처리, 숫자가 아닌 범주형이다.

Product_ID: 비식별화, 범주형

Gender: 범주형

Age: 나이 숫자를 범위로 구성함 0~17 같은 방식. 범주형

Occupation: 직업번호로 범주형이다. 데이터 분석가들에 의하면 미국의 표준 직업 분류와 일치할 것으로 예상함

예를 들어 10: 전문직, 7: 경영관리직, 4: 학생, 15: 교육

City_Category: 인구 통계 데이터와 결합했을 때 A가 가장 높은 구매 총액과 인구 밀도를 보였으므로 'Tier 1(대도시)'으로 정의함.

(A: New York/LA급, B: 중간 규모, C: 소규모)

Stay_In_Current_City_Years: 2, 4+ 등으로 범주화됨

Marital_Status: 0은 미혼, 1은 기혼

Product_Category: 1은 가전제품, 5,8은 의류 및 패션잡화, 12는 생활용품/소형가전, 3은 장난감/취미로 추정됨


특정 카테고리의 평균 가격이 \$15,000~\$20,000 사이인 경우, 블랙 프라이데이가 당시 가장 인기 있었던

고가 가전제품(TV, 노트북 등)으로 명명

Purchase: 구매액, 달러기준

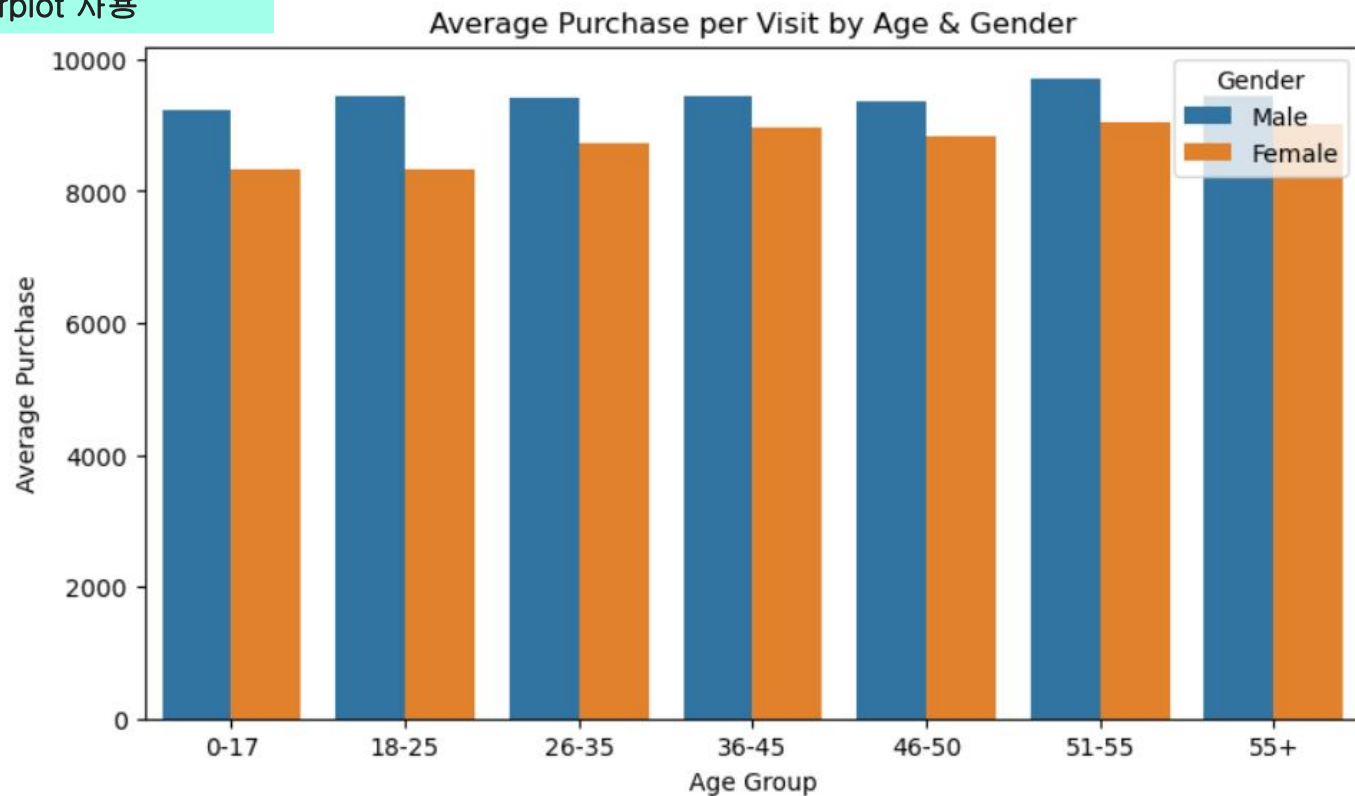
(위 내용은 추정치이며 제미나이를 통해 검색된 결과임. 입증형식은 Kaggle 커뮤니티에서 분석가들에 의해 추정된 내용)

3. 데이터 분석 개요 (전처리 과정, 결과가 나오기 까지의 과정 설명)

- 
- 1) 범주형의 속성값 변경작업하여 새 데이터 프레임에 저장: 숫자로 표시되어 있지만 범주형으로 혼동할 수 있어서 일부 컬럼의 항목값을 미혼, 기혼으로 변경
 - 2) 성별, 미혼, 연령대를 기준으로 방문수, 총구매액, 평균 구매액 계산: 평균 구매액은 그룹별 구매액을 더한후 방문수로 나눔
 - 3) 남성과 여성의 데이터 프레임 구분: 남성과 여성을 구분하여 구매 패턴을 보고자 함
 - 4) 각 성별로 sum: 각 성별로 연령대를 구분하여 방문수, 구매액, 평균 구매액 계산
 - 5) 연령대별 남녀의 평균 구매액 비교
 - 6) 연령대별 남녀의 총 구매액 비교
 - 7) 남녀 구매액 누적 비교

4. EDA 요약

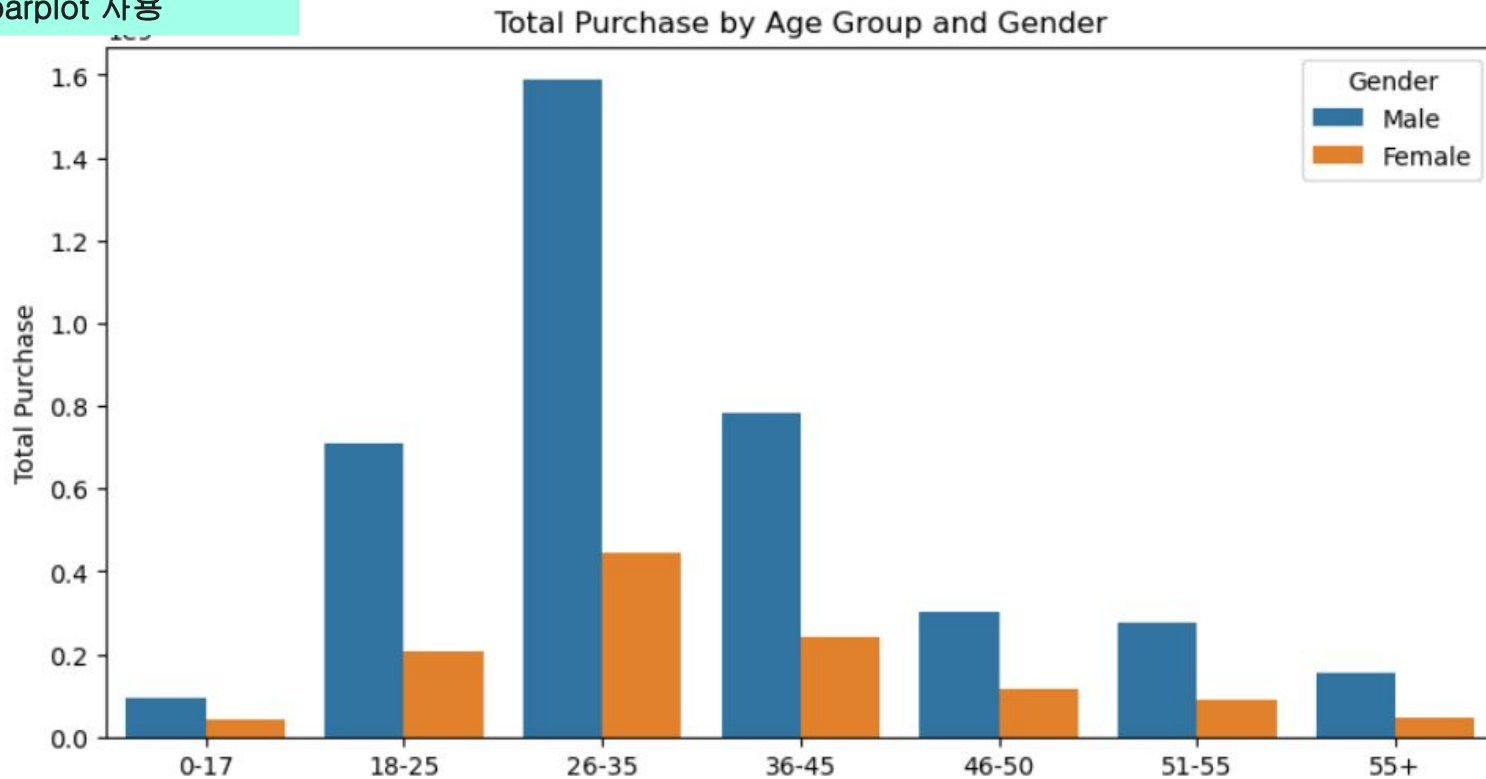
seaborn barplot 사용



연령대별 그룹을 남녀로 구분하여 평균 구매액을 보여주는 컬럼차트
(전 연령대에서 남성의 구매액이 더 크다)

4. EDA 요약

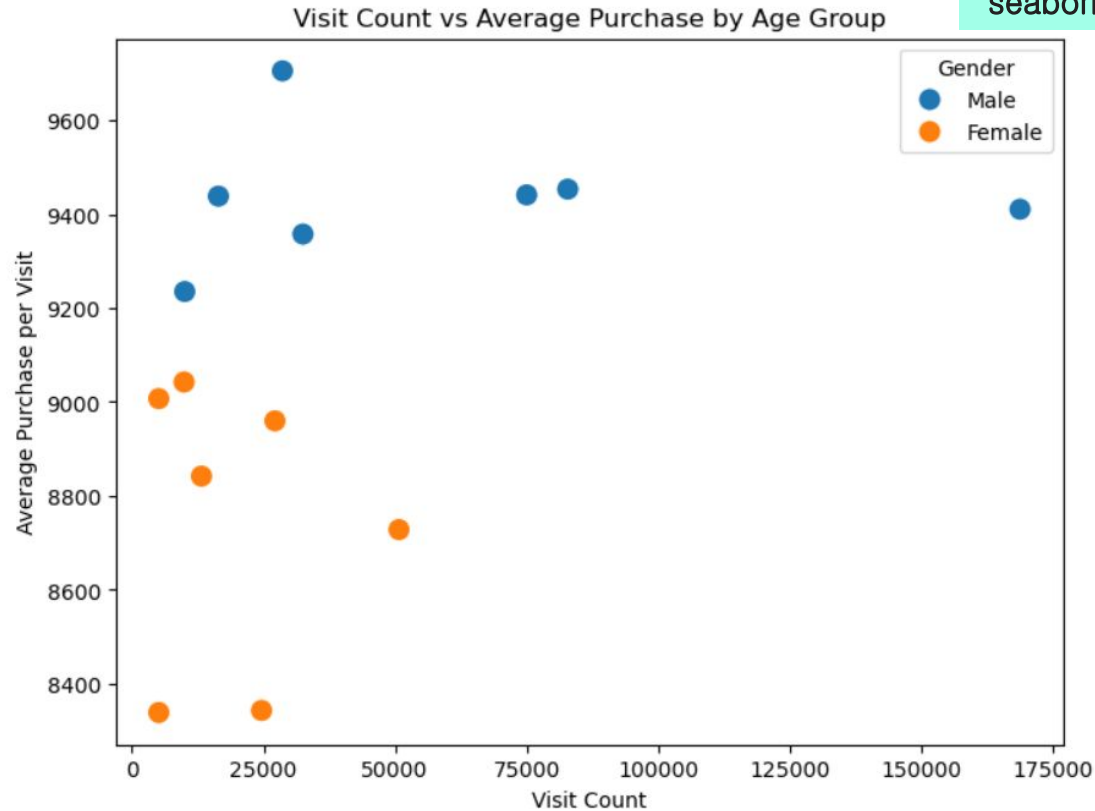
seaborn barplot 사용



연령대별 그룹을 남녀로 구분하여 총 구매액을 비교한 컬럼차트
(남녀 모두 26~35세 사이 구매액이 크다.)

4. EDA 요약

seaborn scatterplot 사용



남성과 여성의 방문수별 평균 구매액 산점도

여성은 비교적 방문횟수 적고, 구매금액도 적다.

남성은 방문수가 적지만 구매금액 크다.

우상단에 배치된 군은 남성으로 자주 방문하면서 돈을 많이 쓰는 군


5. 주요결과



- 1) 월마트에서 주로 남성이 돈을 더 많이 쓴다.
- 2) 남녀 구분을 막론하고 26~35세 사이의 연령대가 가장 많이 돈을 쓴다.
- 3) 남성은 여성보다 방문할때마다 돈을 더 많이 쓴다.

6. 결론 및 인사이트

(이 데이터는 블랙 프라이데이이고 이 시기에 가전제품에 대한 관심도가 높기 때문에 생긴 결과이므로
평상시의 구매 데이터와는 차이가 있을 것으로 예상됨)



1) 남성을 타겟으로 하는 제품군을 늘릴 필요가 있다.

제품군의 경우 방문시 구매가가 높은 가전제품에 대한 높은 수요예상

2) 제품군에서 젊은층을 공략하기

남녀 모두 젊은 층이 가장 많이 돈을 쓰는 만큼 젊은이들을 위한 제품군을 늘려야 한다.

3) 여성을 끌어들이기 위한 이벤트 행사 등을 보강할 필요

전 연령대에서 여성 구매액이 적은 만큼 매력적인 프로모션을 통해 유입증가를 모색해야 함.



Q&A