D&A Competition

에바행조 권민지 박준하 신지후 안병민 전지현 정다연

목차

1. 주류 판매와 관련된 요인 6가지 분석, 패턴 파악

2. 주류 판매의 효과적인 전략 구축

3. 여름 신제품 출시에 적용하여 이익 극대화

사용한 데이터

01. customer_data.csv

- 고객 정보 데이터

02. order_data.csv

- 상품 구매 정보 데이터

03. product_data.csv

- 상품 분류 정보 데이터

전처리과정

분석에 필요한 DF 편집

age_group

price

buy_count

buy_date

전처리과정

분석에 필요한 DF 편집

특징 파악할 부분

1. 고객
2. 판매수량
3. 판매가격
4. 일년 추세
5. 시간별 추세
6. 지역

```
# c(고객데이터), o(주문데이터), p(상품데이터) 병합
co=pd.merge(c, o, on='customer_id')
df=pd.merge(co, p, on='product_code')
```

```
# 오프라인 매장만 (온라인 매장 삭제)
df=df.drop(df[df['on_off_div']==2].index)
```

```
# 필요하지 않은 컬럼 삭제
col=['order_id','on_off_div','partner_code','product_code']
df=df.drop(col, axis=1)
```

```
# 주류 데이터만 추출
df=df.loc[df['large_product_cat']=='주류']
```

```
# 인덱스를 삭제하고 재설정
df=df.reset_index(drop=True)
```

	customer_id	gender	age_group	location	market_code	buy_date	buy_hour	buy_amount	buy_count	product	large_product_cat	mid_product_cat
0	M000034966	여성	40대	Z07	A020116	20210919	16	16440.0	2	국산맥주	주류	맥주
1	M000034966	여성	40대	Z07	A043676	20210116	21	10800.0	4	국산맥주	주류	맥주
2	M000136117	여성	30대	Z11	A020092	20210104	17	8220.0	1	국산맥주	주류	맥주
3	M000201112	여성	50대	Z17	A043753	20210725	15	8000.0	4	국산맥주	주류	맥주
4	M000504230	여성	30대	Z05	A030304	20210219	17	9240.0	1	국산맥주	주류	맥주

전처리과정

age_group

```
# 60대 + 70대 -> 70대 이상
my_dict = {'60대': '60대 이상','70대': '60대 이상'}
df['age_group'] = df['age_group'].replace(my_dict)
```

```
df['age_group'].value_counts()
```

```
40대    68838
30대    43666
50대    35058
20대    11251
60대     8606
70대     2415
```

```
df['age_group'].value_counts()
```

```
40대        68838
30대        43666
50대        35058
20대        11251
60대 이상    11021
```

전처리과정

price

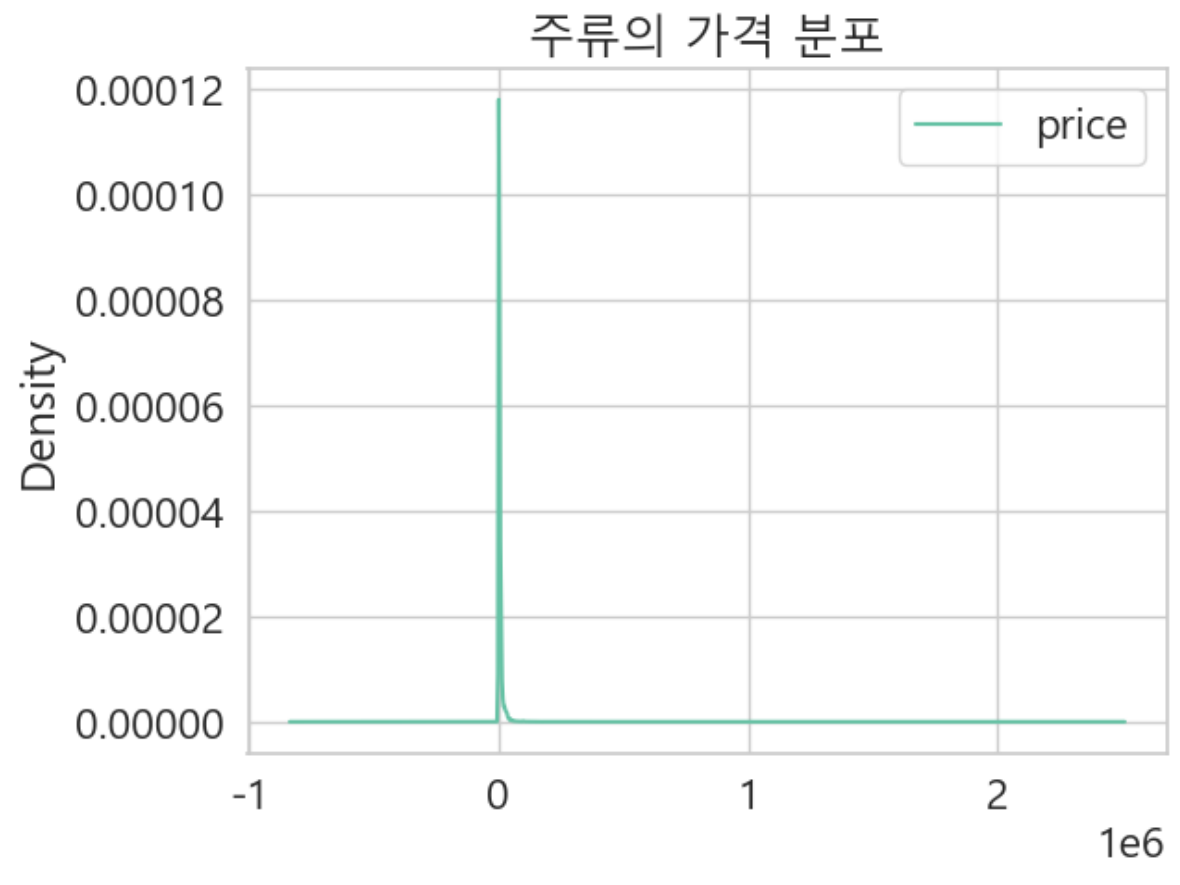
```
# 상품 하나당 가격(price) 컬럼 추가
price=df['buy_amount']/df['buy_count']
df.insert(8,'price',price)
```

```
df.loc[:,df.columns[6:9]].head()
```

	buy_amount	buy_count	price
156312	16440.0	2	8220.0
156313	10800.0	4	2700.0
156314	8220.0	1	8220.0
156315	8000.0	4	2000.0
156316	9240.0	1	9240.0

전처리과정

price




```
df['price'].describe()
```

```
count    1.698340e+05
mean     6.277998e+03
std      1.877288e+04
min      1.000000e+01
25%      1.570000e+03
50%      2.500000e+03
75%      7.080000e+03
max      1.672000e+06
Name: price, dtype: float64
```

```
# 주류 중분류별로 따로 떼내기
beer=df[df['mid_product_cat']=='맥주']
soju=df[df['mid_product_cat']=='소주']
trad=df[df['mid_product_cat']=='전통주']
wine=df[df['mid_product_cat']=='와인']
oset=df[df['mid_product_cat']=='주류세트']
fore=df[df['mid_product_cat']=='양주']
```

전처리과정

price

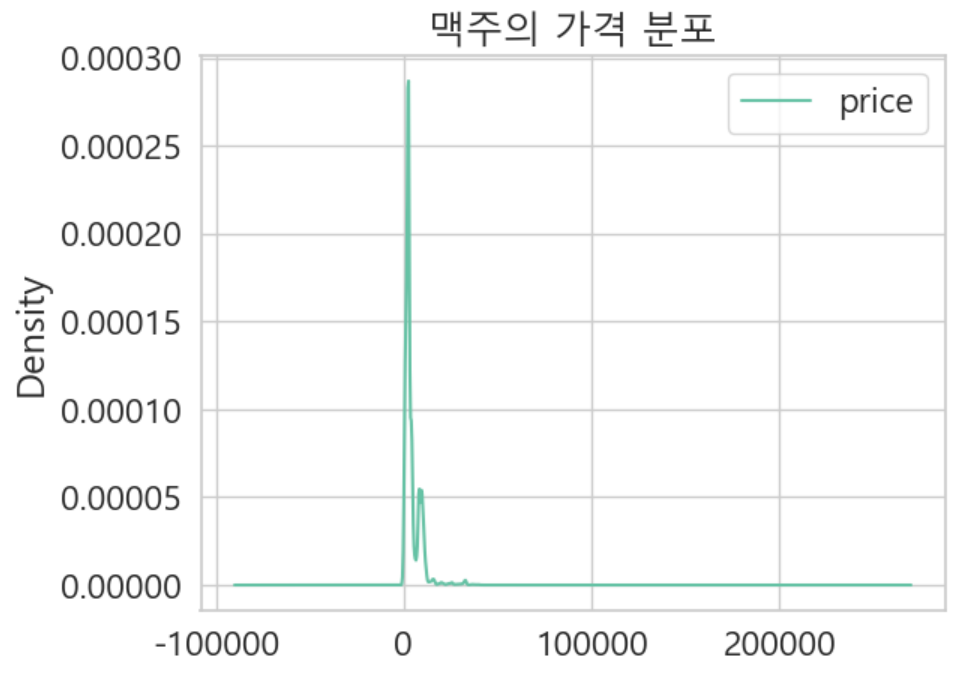
```
# 극단치 처리 함수

def deal_with_extremes(kind):

    # 상위 10% -> 상위 10% 속하는 값들의 중앙값
    q9=kind['price'].quantile(q=0.9, interpolation='nearest')
    m9=kind.loc[kind['price']>=q9]['price'].median()
    kind['price'].loc[kind['price']>=q9]=m9

    # 하위 10% -> 하위 10% 속하는 값들의 중앙값
    q1=kind['price'].quantile(q=0.1, interpolation='nearest')
    m1=kind.loc[kind['price']<=q1]['price'].median()
    kind['price'].loc[kind['price']<=q1]=m1
```

```
deal_with_extremes(beer)
deal_with_extremes(soju)
deal_with_extremes(trad)
deal_with_extremes(wine)
deal_with_extremes(oset)
deal_with_extremes(fore)
```

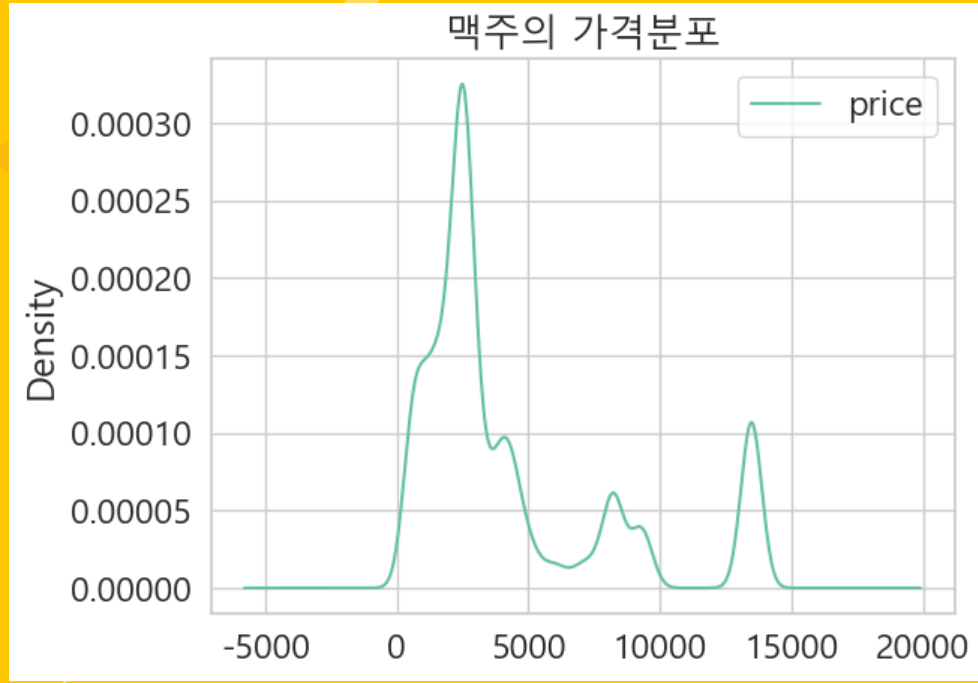
전처리과정

price

〈 극단치 처리 전 〉



〈 극단치 처리 후 〉



가격 범위가 10원~18만원에서 약 600원~13000원으로 줄어들어 분포를 **더 쉽고 정확하게** 파악 가능

전처리과정

buy_count

```
# 7개 이상을 7로 표기
df['buy_count'] = df['buy_count'].clip(upper=7)
```

```
df['buy_count'].value_counts().tail(15)
```

```
27    4
40    3
72    2
60    2
26    2
28    1
63    1
34    1
33    1
37    1
36    1
96    1
29    1
31    1
35    1
Name: buy_count,
```

```
df['buy_count'].value_counts()
```

```
1      108674
2       35084
4        9904
3        8848
6        2415
5        2413
8         819
10        392
12        316
7         310
20        139
```

```
df['buy_count'].value_counts()
```

```
1      108674
2       35084
4        9904
3        8848
7        2496
6        2415
5        2413
Name: buy_count, dtype: int64
```

전처리과정

buy_date

```
df['buy_date'] = pd.to_datetime(df['buy_date'].astype(str), format='%Y-%m-%d')
```

```
# 년 추출 -> year 컬럼 추가
year = df['buy_date'].dt.year
df.insert(5,'year',year)
```

```
# 월 추출 -> month 컬럼 추가
month = df['buy_date'].dt.month
df.insert(5,'month',month)
```

```
# 일 추출 -> day 컬럼
day = df['buy_date'].dt.day
df.insert(5,'day',day)
```

```
df.loc[:,df.columns[4:8]].head()
```

	buy_date	day	month	year
156312	2021-09-19	19	9	2021
156313	2021-01-16	16	1	2021
156314	2021-01-04	4	1	2021
156315	2021-07-25	25	7	2021
156316	2021-02-19	19	2	2021

전처리과정

전처리과정 적용

```
# 모든 주류 중분류에 전처리 과정 적용

beer=df[df['mid_product_cat']=='맥주']
soju=df[df['mid_product_cat']=='소주']
trad=df[df['mid_product_cat']=='전통주']
wine=df[df['mid_product_cat']=='와인']
oset=df[df['mid_product_cat']=='주류세트']
fore=df[df['mid_product_cat']=='양주']
```

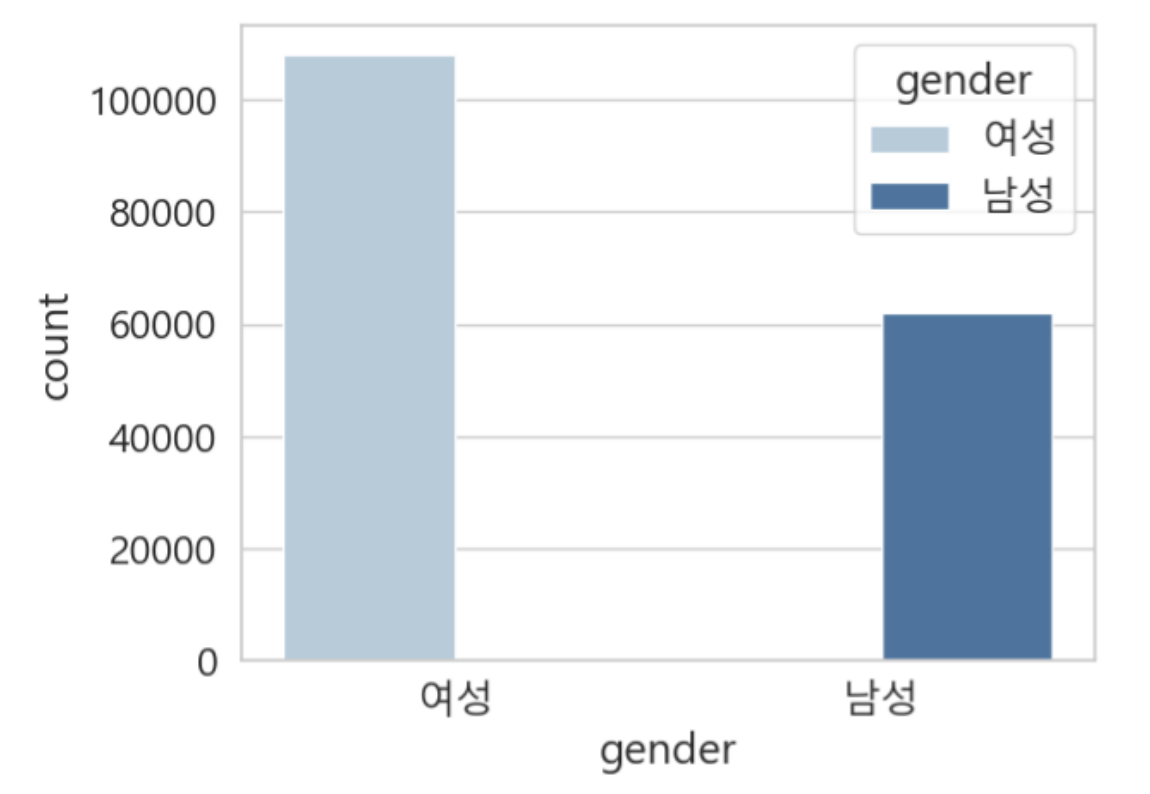
분석과 인사이트

고객 수량 가격

시기 시간 지역

분석과 인사이트

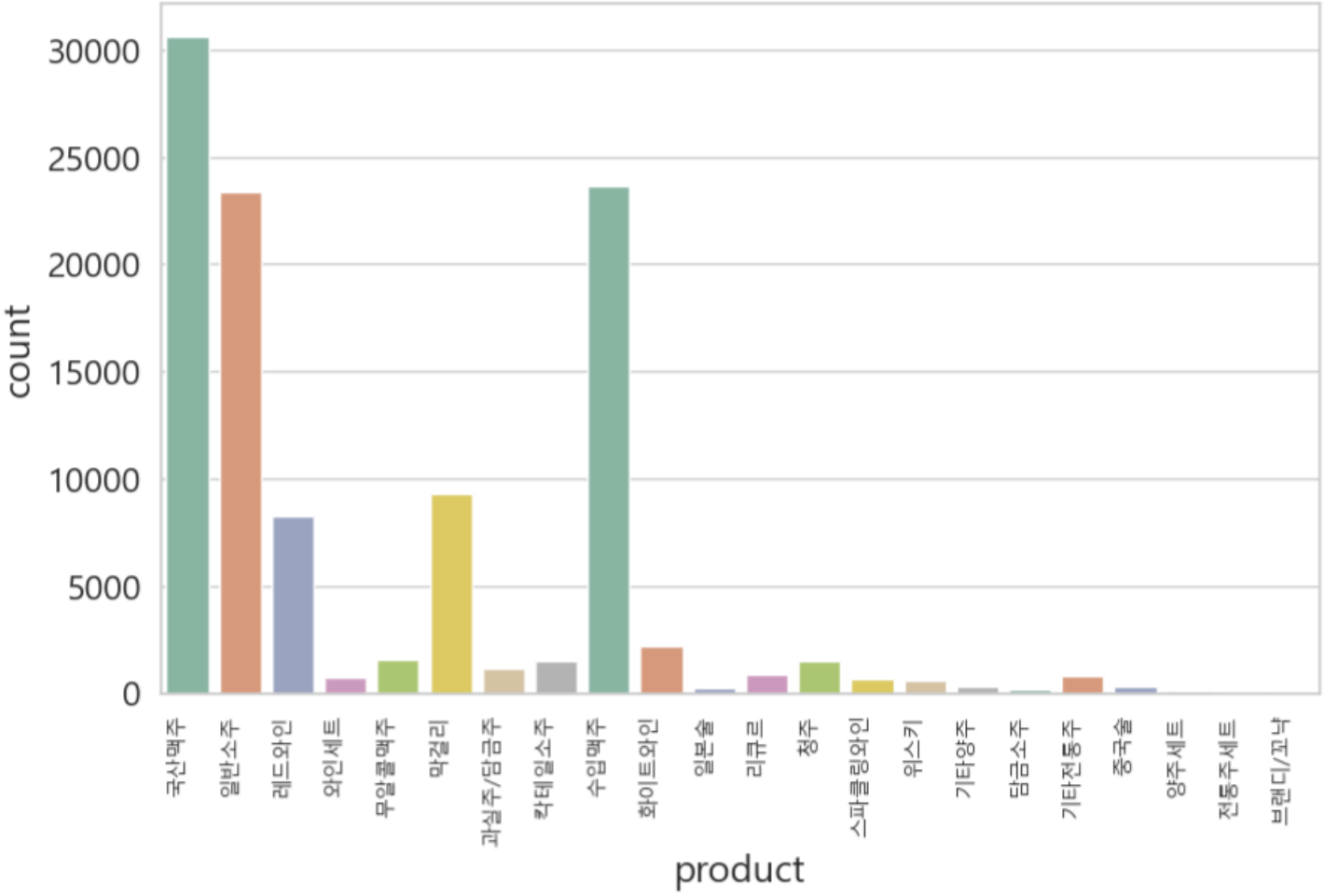
고객



여성의 주류 구매량 > 남성의 주류 구매량

분석과 인사이트

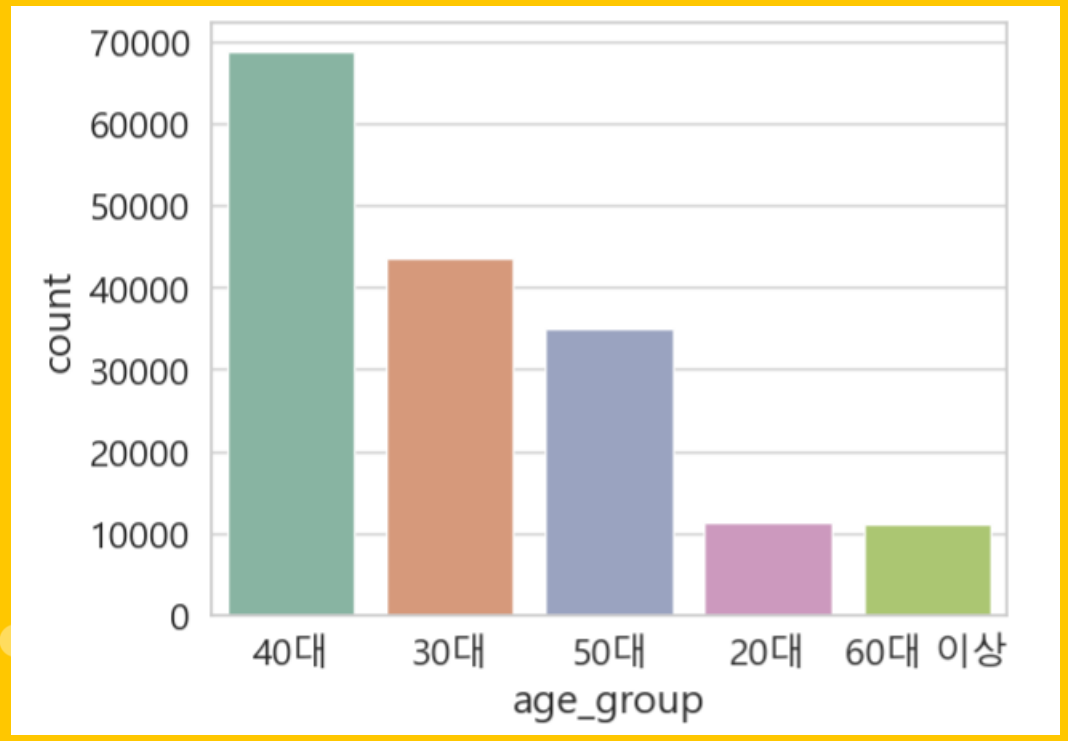
고객



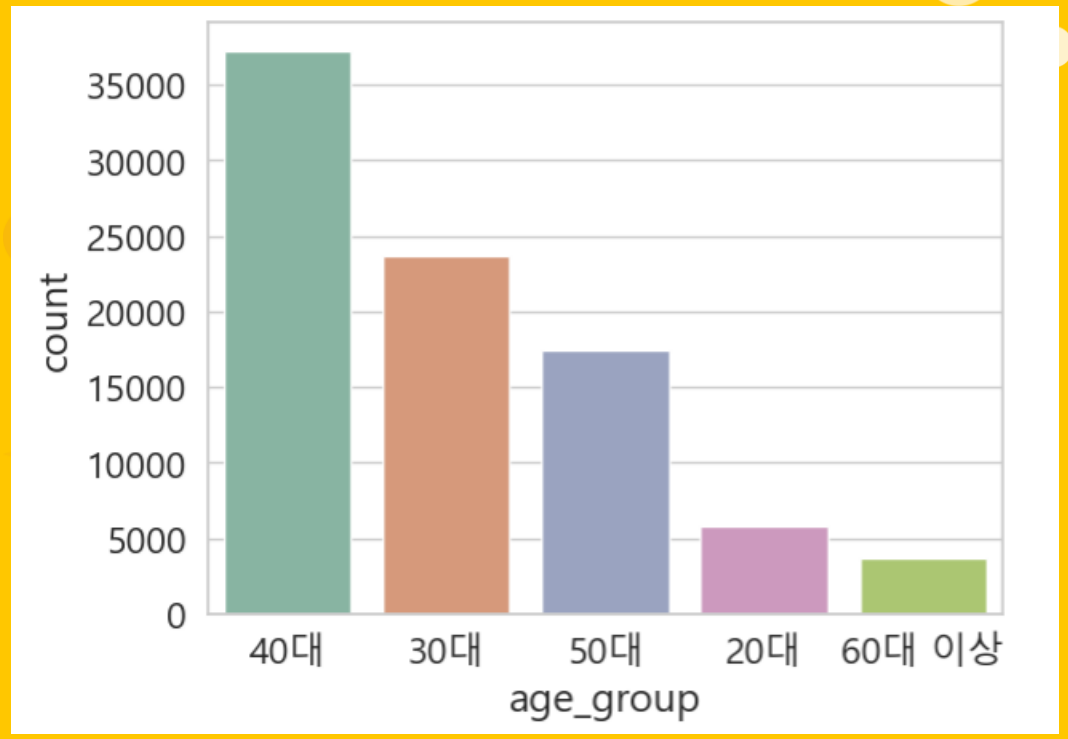
여성의 주류 구매량
-> 1위 국산맥주
2위 수입맥주

분석과 인사이트

고객



주류를 구매한 연령대



맥주를 구매한 연령대

40대 > 30대 > 50대 > 20대 > 60대 이상

분석과 인사이트

수량

〈주류 중분류별 1년 총 판매량〉

```
# 주류 중분류별 총 구매량
count_df = df.groupby(['mid_product_cat']).agg({'buy_count':'sum'});count_df
```

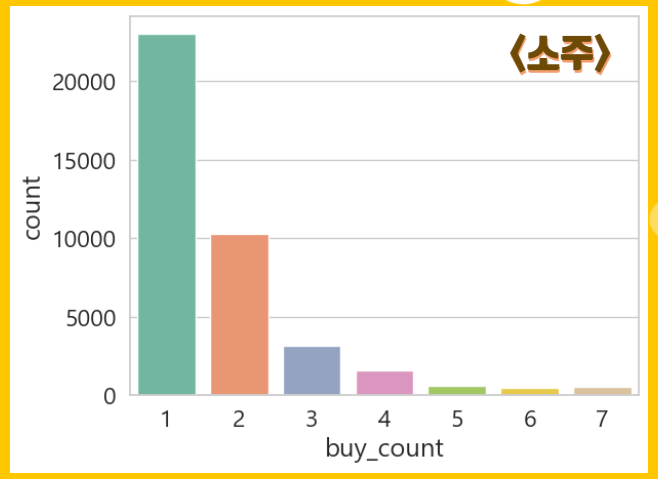
mid_product_cat	buy_count
맥주	160318
소주	69060
양주	4086
와인	21764
전통주	31565
주류세트	1833

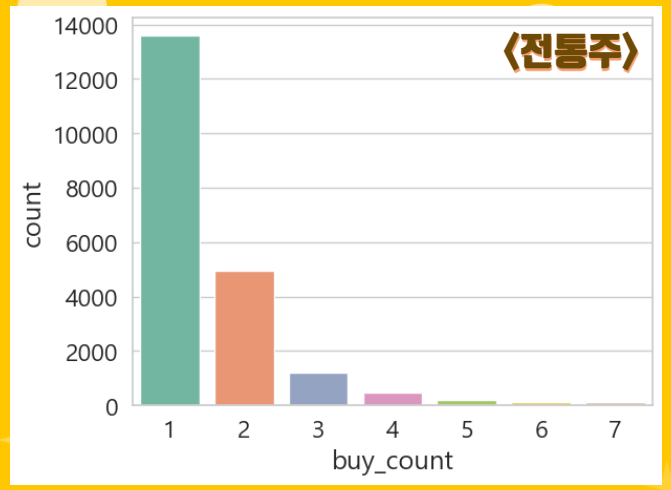
TOP3

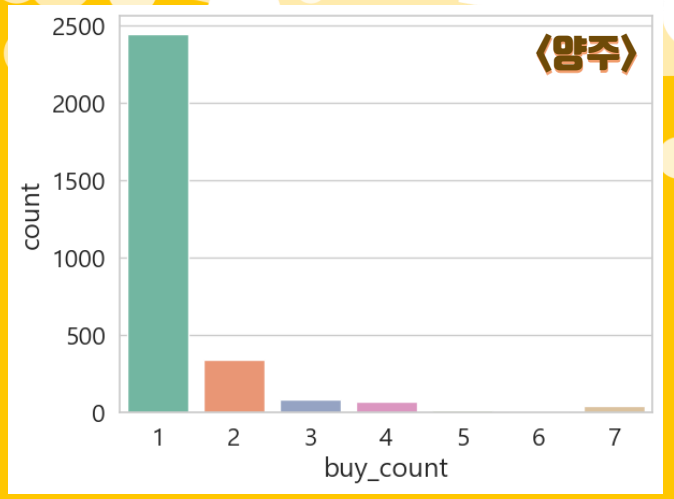
BOTTOM3

분석과 인사이트

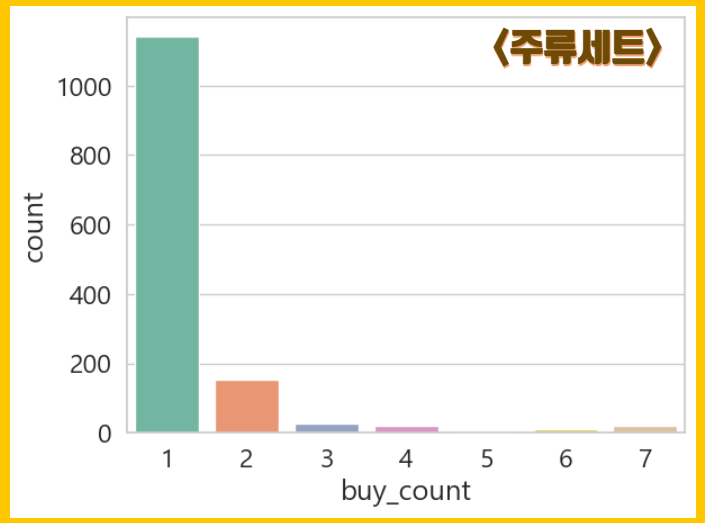
수량

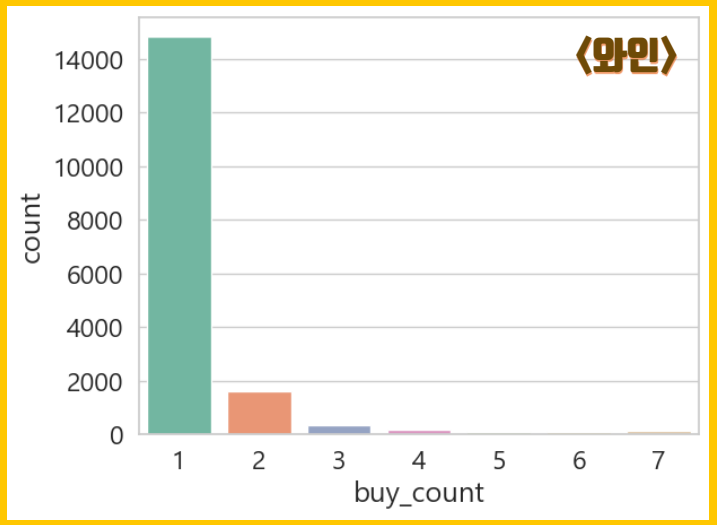






공통된 특징 : 한번에 한개씩 가장 많이 사가고
두개씩, 세개씩, 네개씩 ... 순서대로 한번에 사가는 고객 수가 줄어든다.

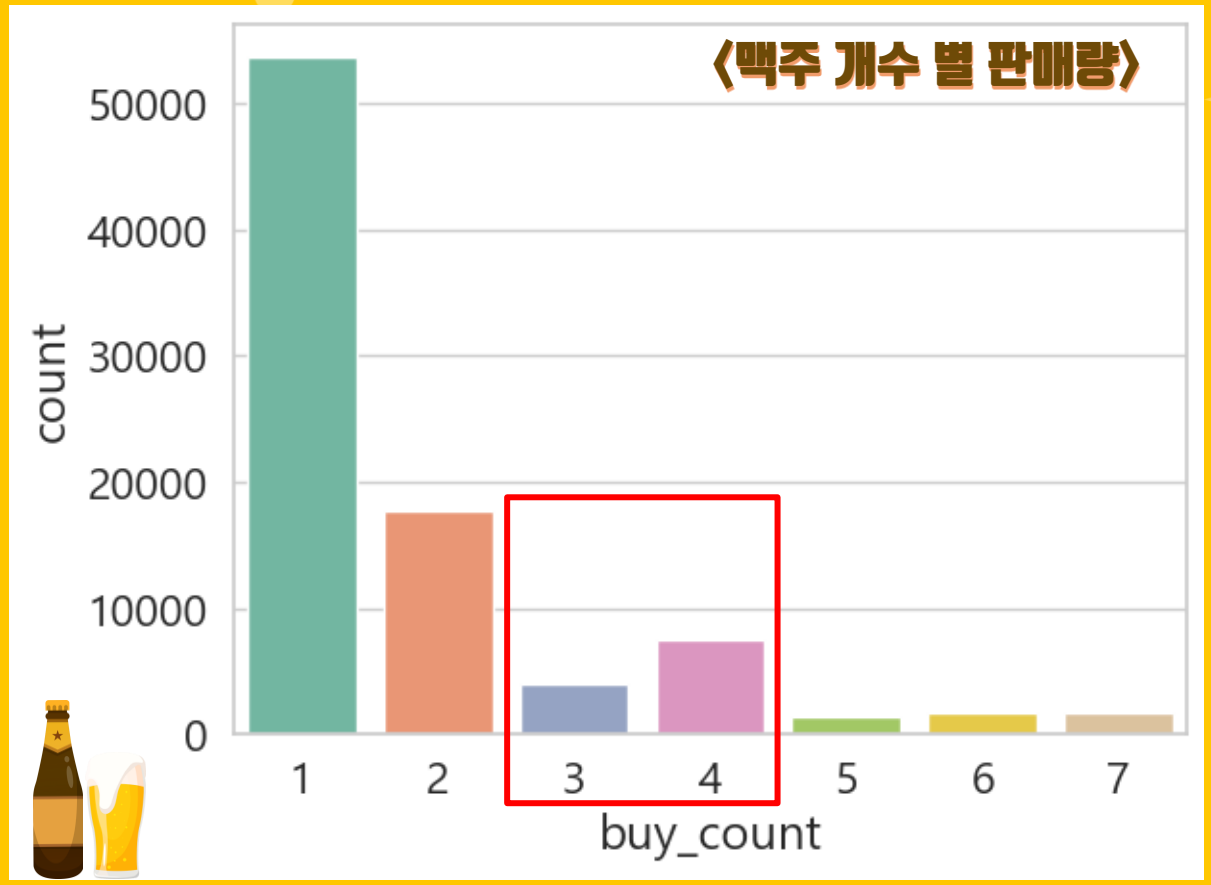




분석과 인사이트

수량

고객들이 한번에 맥주 3개보다 4개를 더 많이 구매한다.
맥주의 경우 4개를 사면 할인해주기 때문이라고 원인 유추

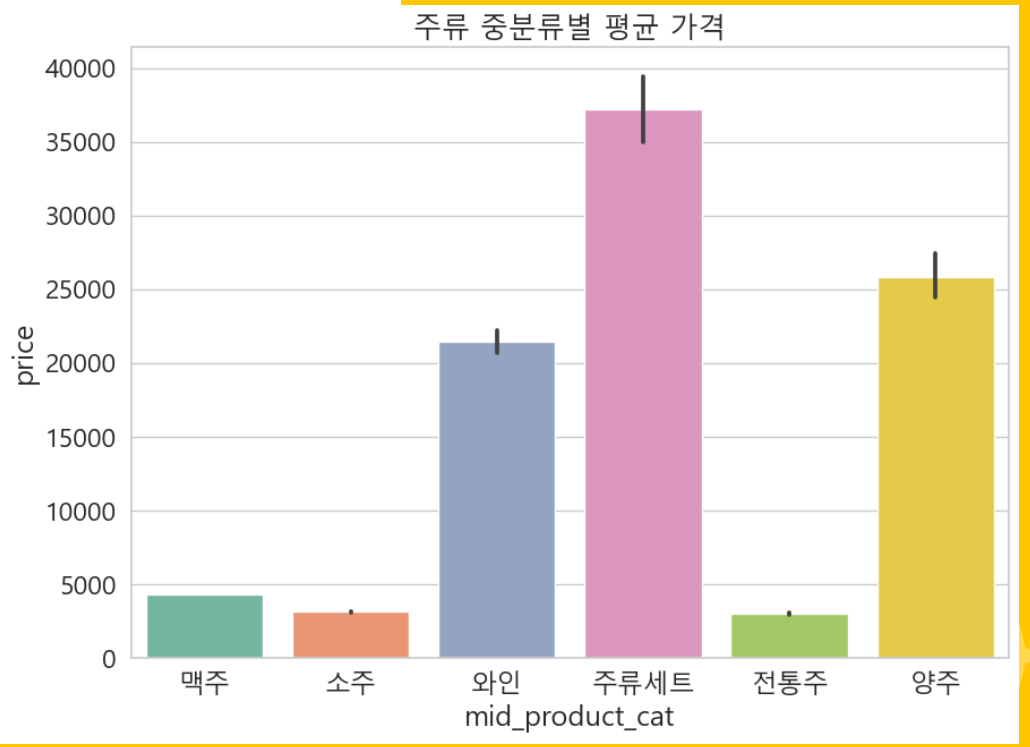


분석과 인사이트

가격

```
# 데이터 분석
money_df=df.groupby(['mid_product_cat']).agg({'price':'mean','buy_amount':'sum','buy_count':'sum'})
money_df=money_df.rename(columns={'price':'평균가격','buy_amount':'매출','buy_count':'판매량'})
money_df.sort_values('매출', ascending=False)
```

mid_product_cat	평균가격	매출	판매량
맥주	4324.559548	571956162.0	167338
와인	21422.899396	454916159.0	22491
소주	3136.505076	174067040.0	71676
양주	25829.061806	91639540.0	4450
전통주	2983.926755	83096560.0	32027
주류세트	37177.673675	66907020.0	1960



맥주, 소주, 전통주 : 가격대가 낮고 평균가격에 가깝게 분포
와인, 주류세트, 양주 : 가격대가 높고 넓게 분포

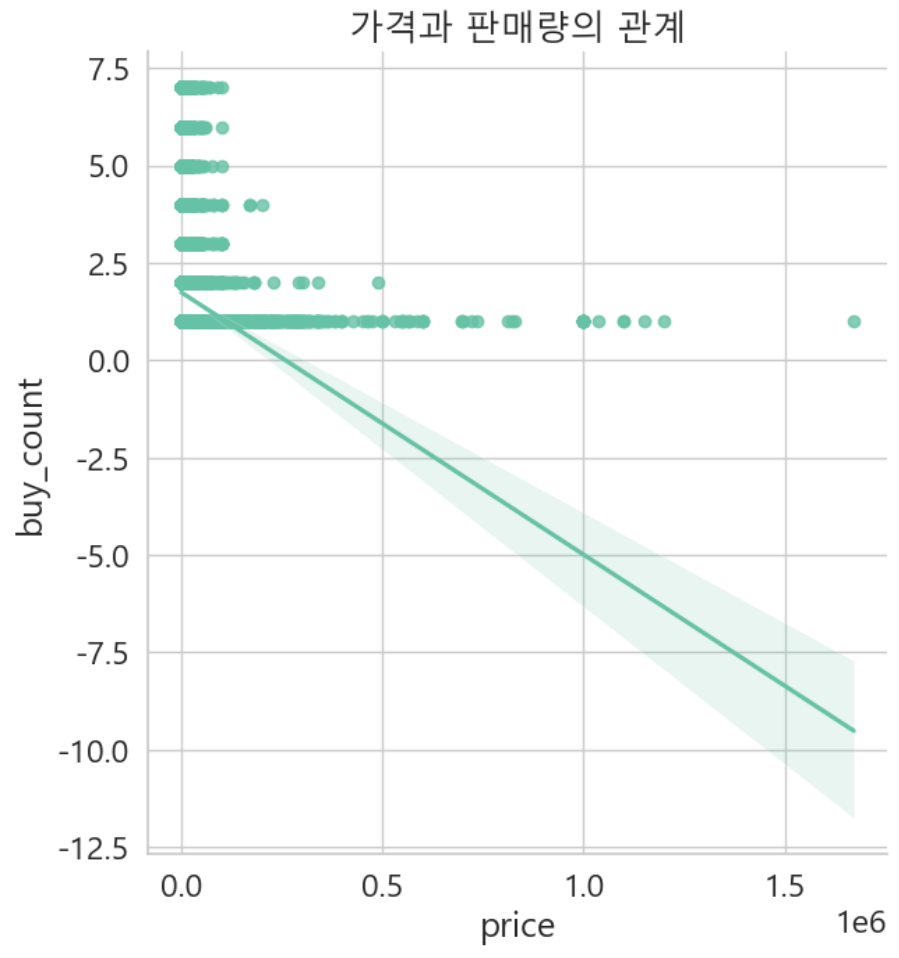
가격대가 낮은 품목 = 판매량 & 매출 TOP3

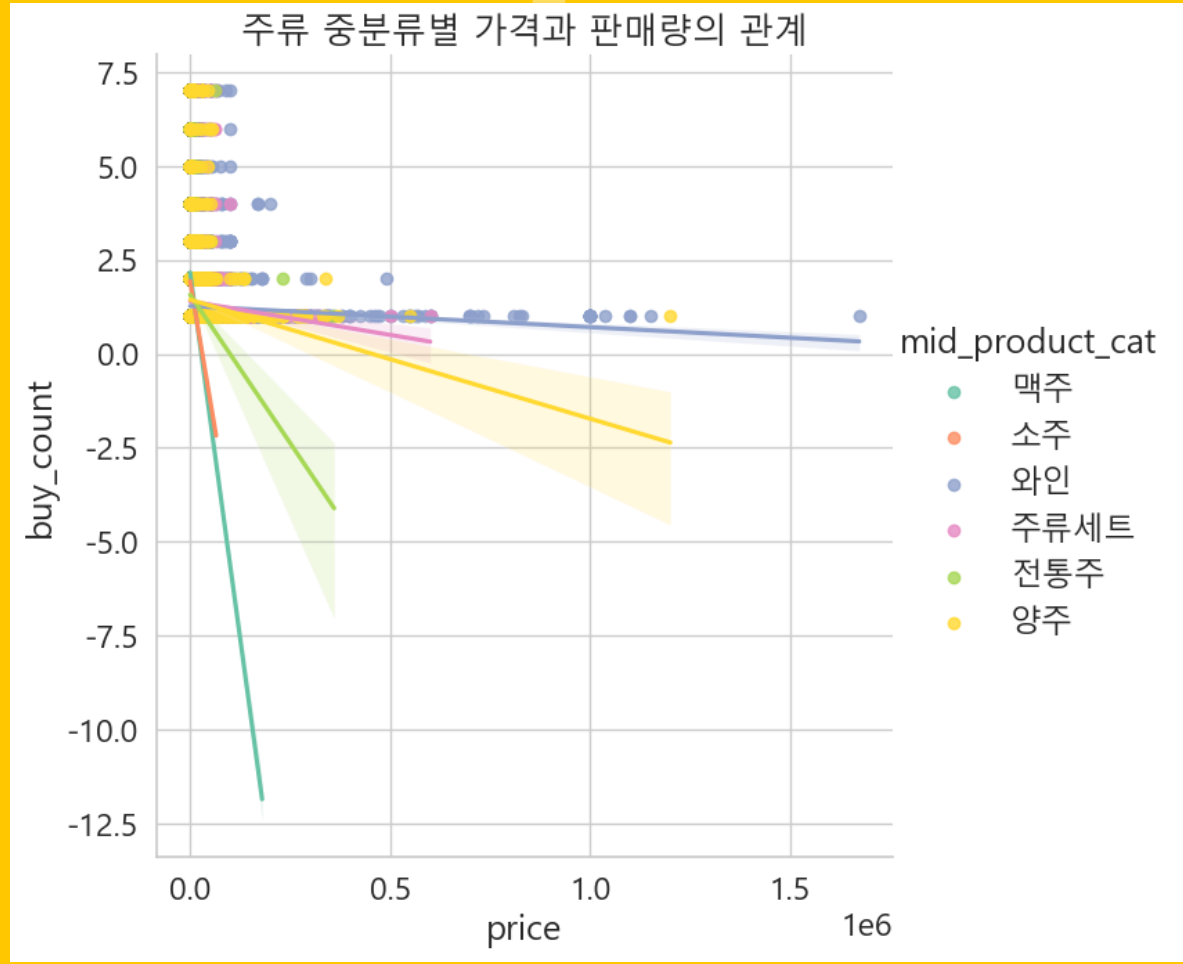
가격이 낮을수록 더 많이 팔리나?

분석과 인사이트

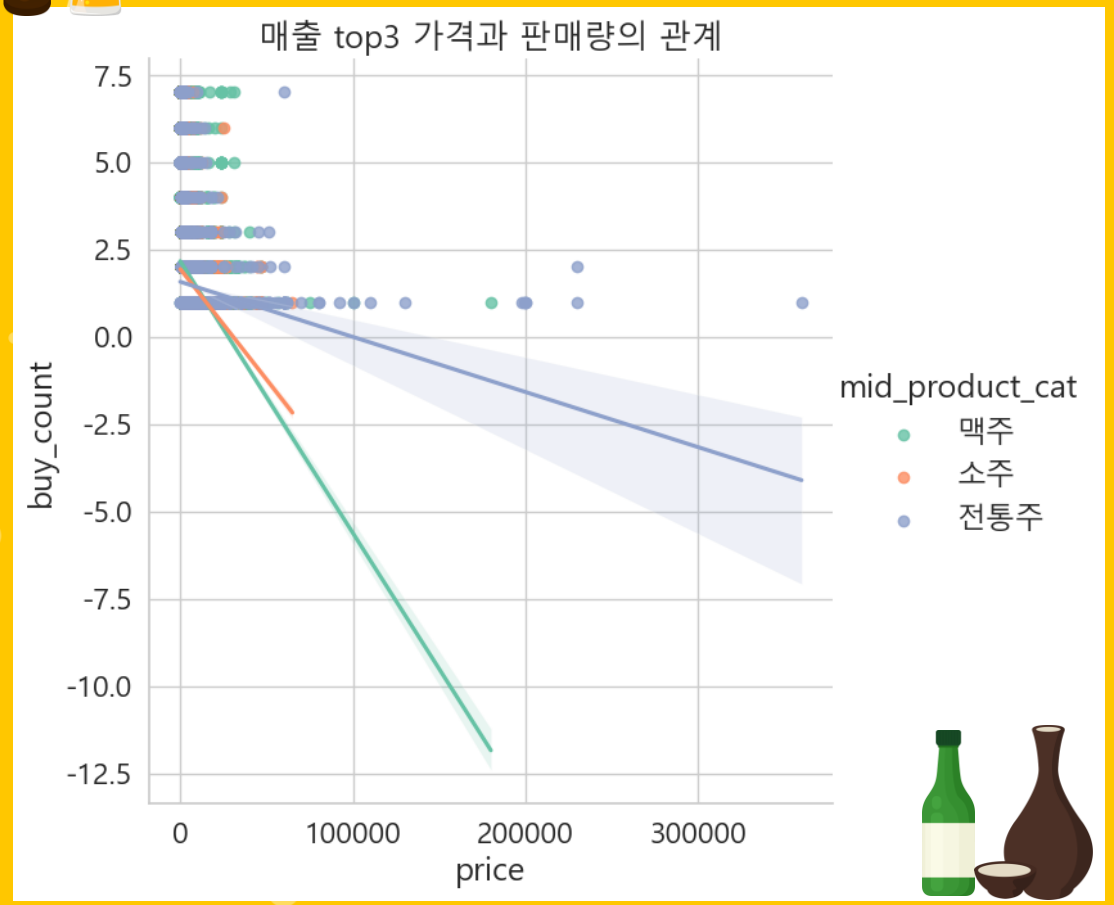
가격

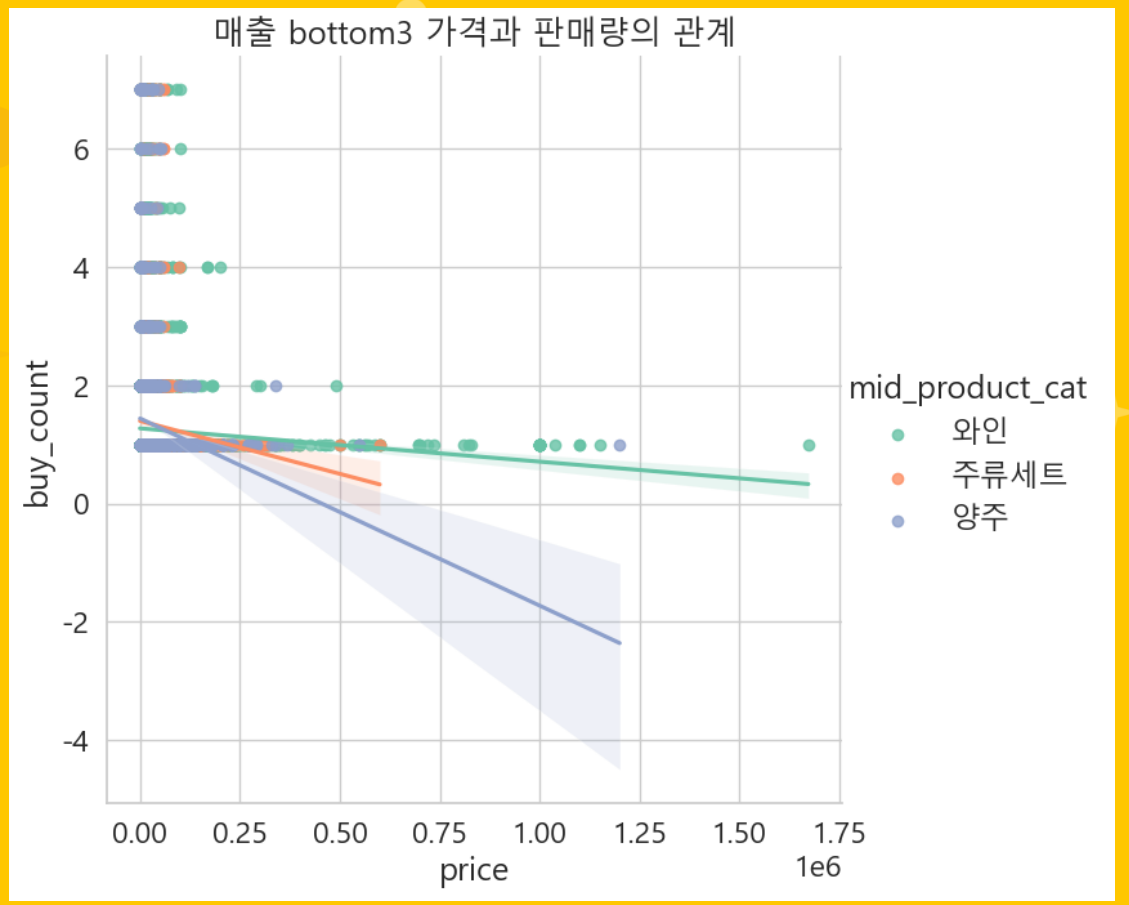
우하향 그래프 -> 가격이 낮을수록 역시 잘 팔리는군 !





매출 TOP3 더 가파르게 우하향하는 그래프 -> 가격이 낮을수록 더더더 많이 산다

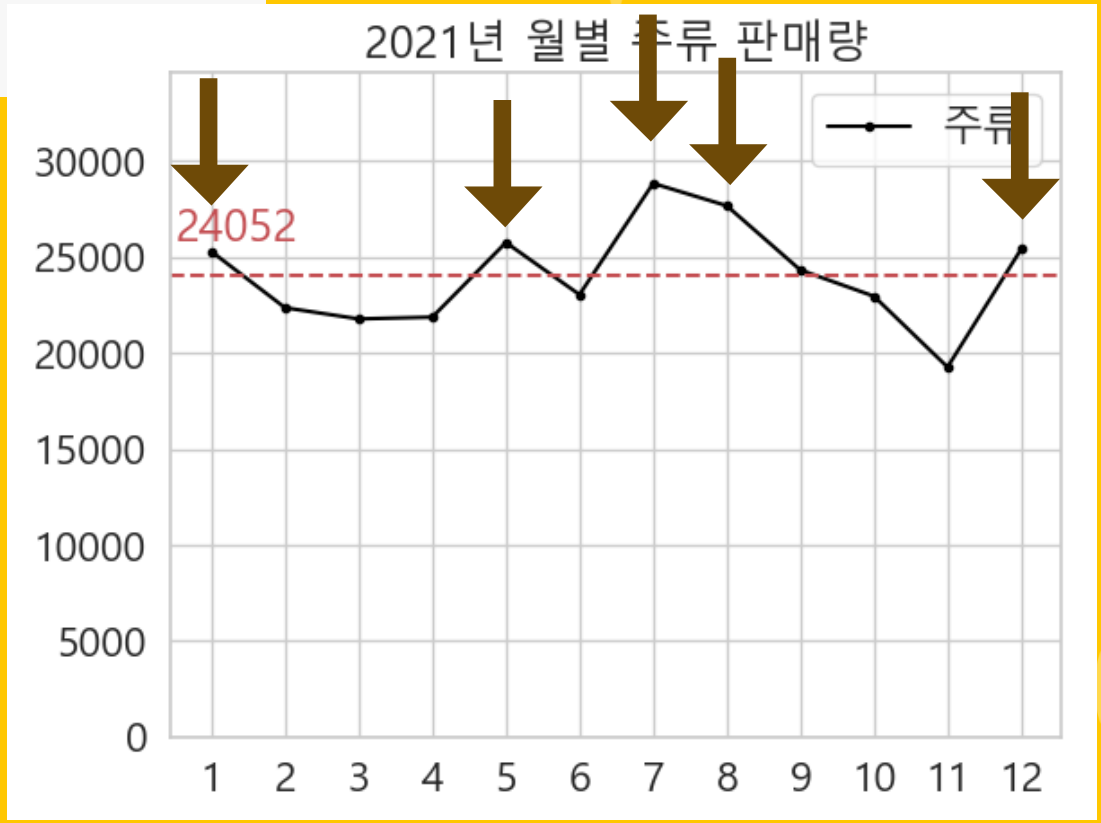




분석과 인사이트

시기

```
###상품의 월별 판매량 추세 함수
def show_large_product_month(product):
    ex_df = df.query(f'large_product_cat == "{product}"')
    monthly_count = ex_df.groupby(['month'])['buy_count'].sum()

    return monthly_count
```



'2021년 월 평균 약 24000개씩 판매'

&

1, 5, 7, 8, 12월 판매량

연초, 연말, 연휴 주류 판매량

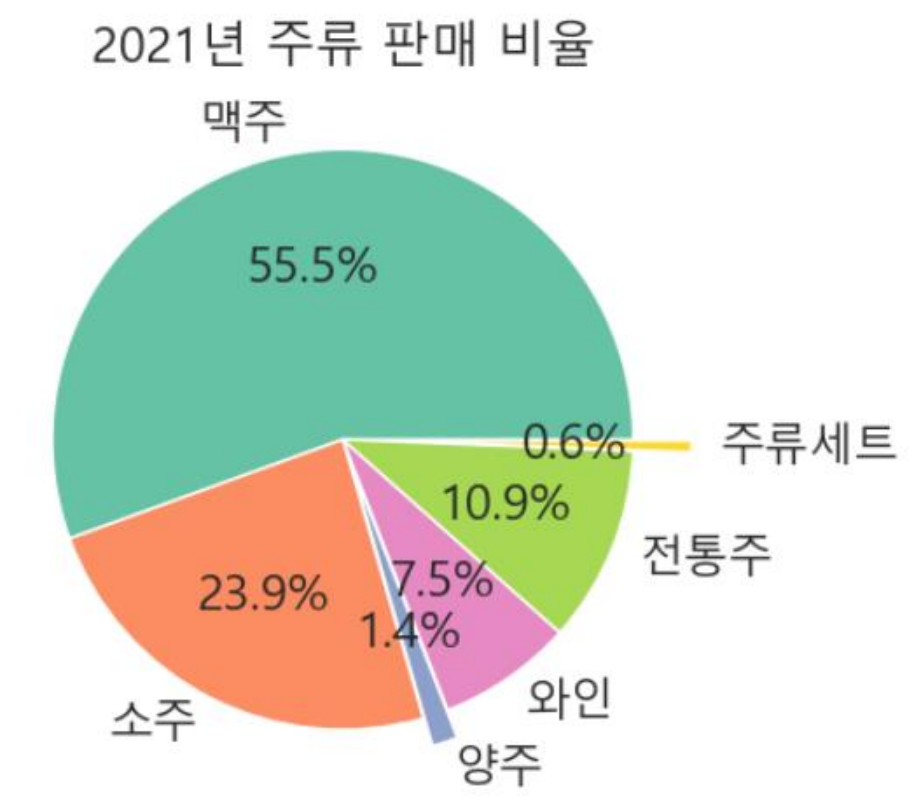
7, 8월은 '왜' 높을까??

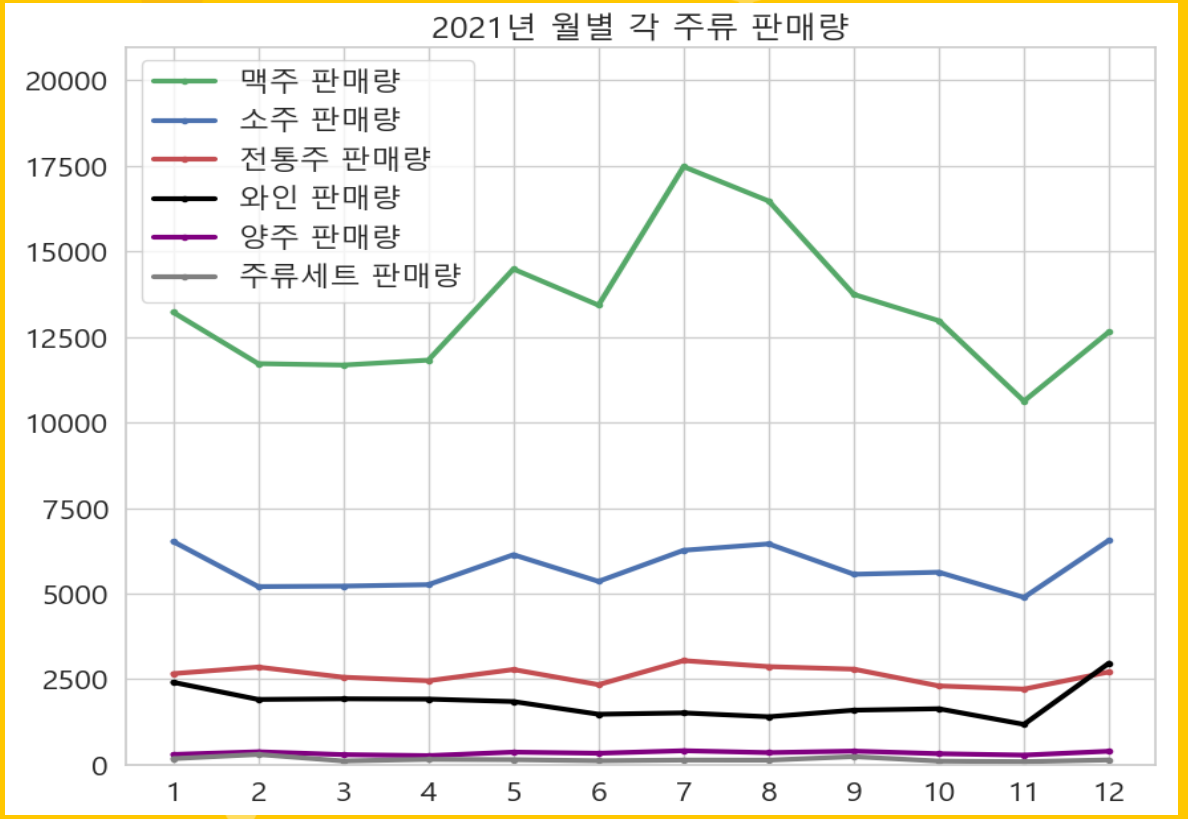
분석과 인사이트

시기

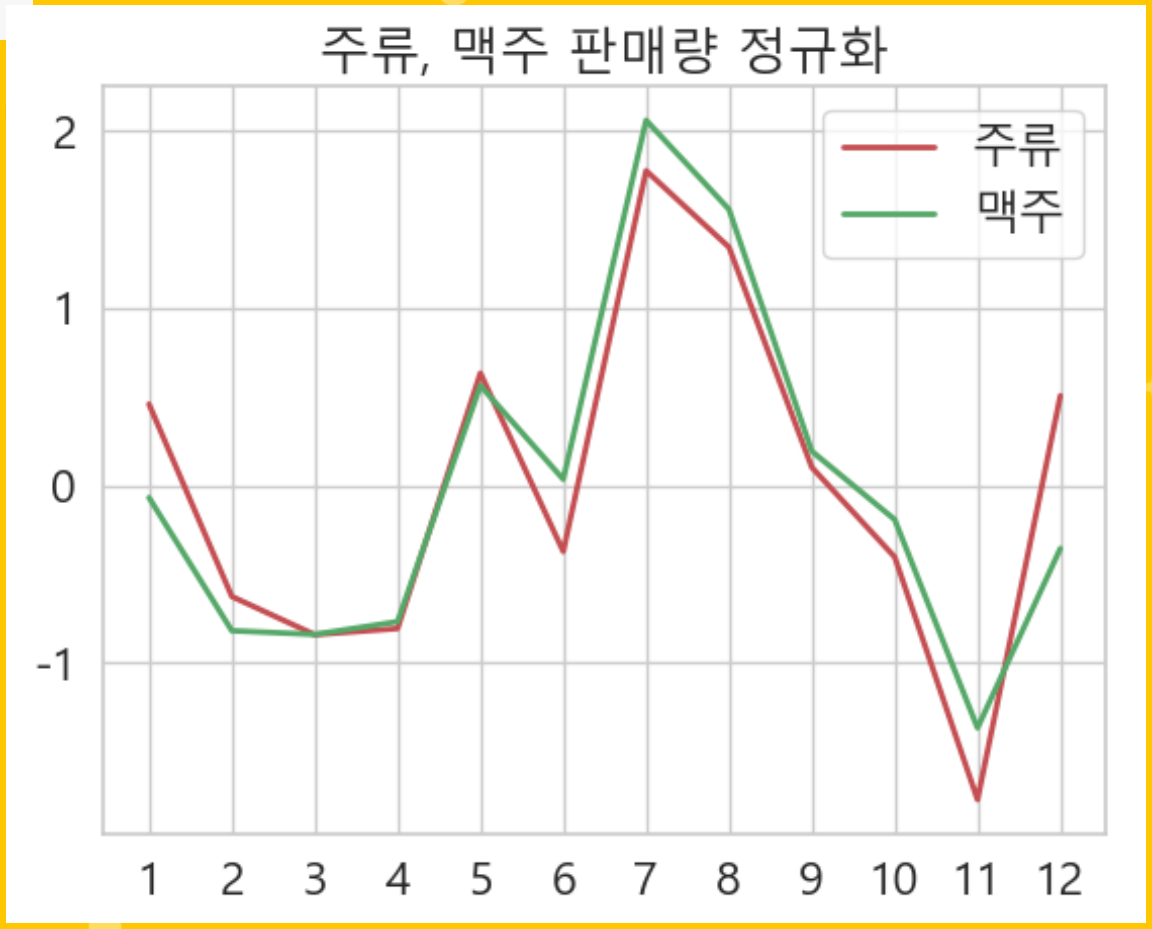
```
# 2021년 주류 판매 비율 파이차트
buy_count_mid = df.groupby(['mid_product_cat']).sum()['buy_count'].reset_index()

plt.pie(buy_count_mid['buy_count'].tolist(), labels=buy_count_mid['mid_product_cat'].tolist(), explode = [0, 0, 0.1, 0, 0, 0.2])
plt.title('2021년 주류 판매 비율')
plt.show()
```

맥주의 비율이 55.5%로 매우 높다!



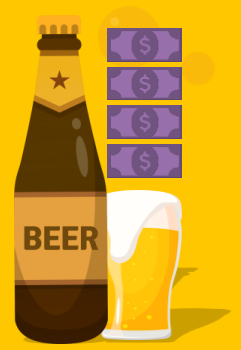


```
def normalize(x):
    return (x - x.mean())/x.std()
```



정규화한 주류와 맥주의
판매량 그래프가 매우 유사!

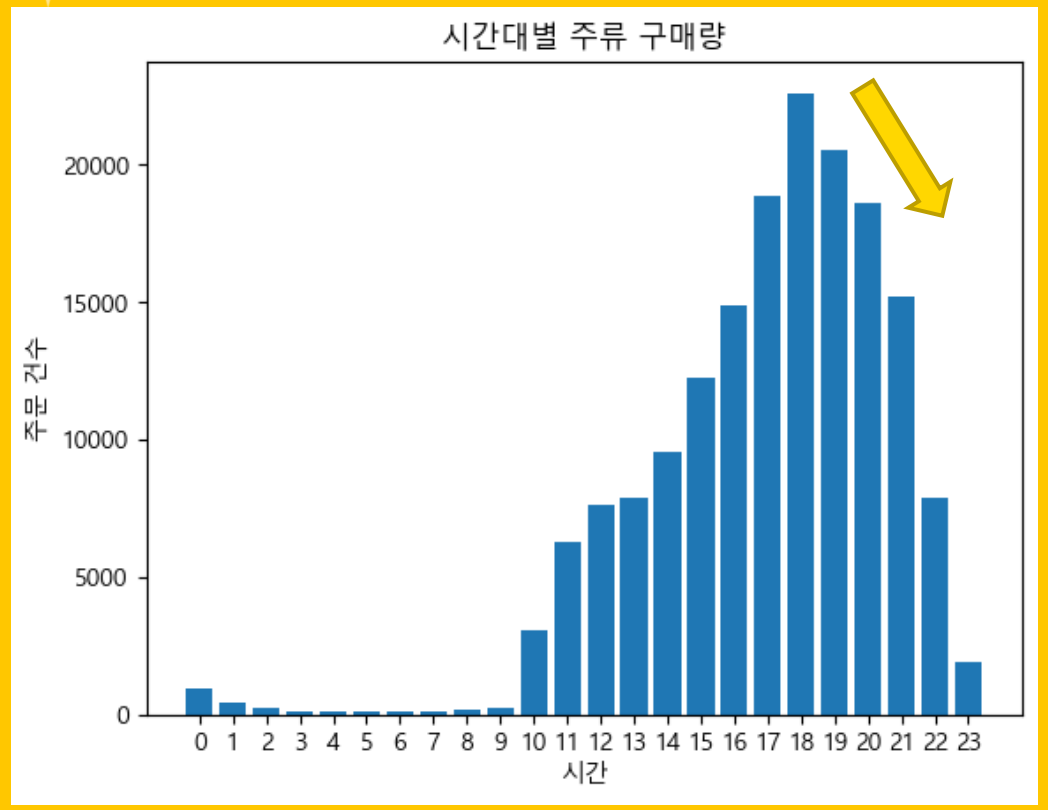
-> 7, 8월 맥주의 판매량이
주류 판매량에 영향



술 ,시간대별 얼마나 판매될까?

분석과 인사이트

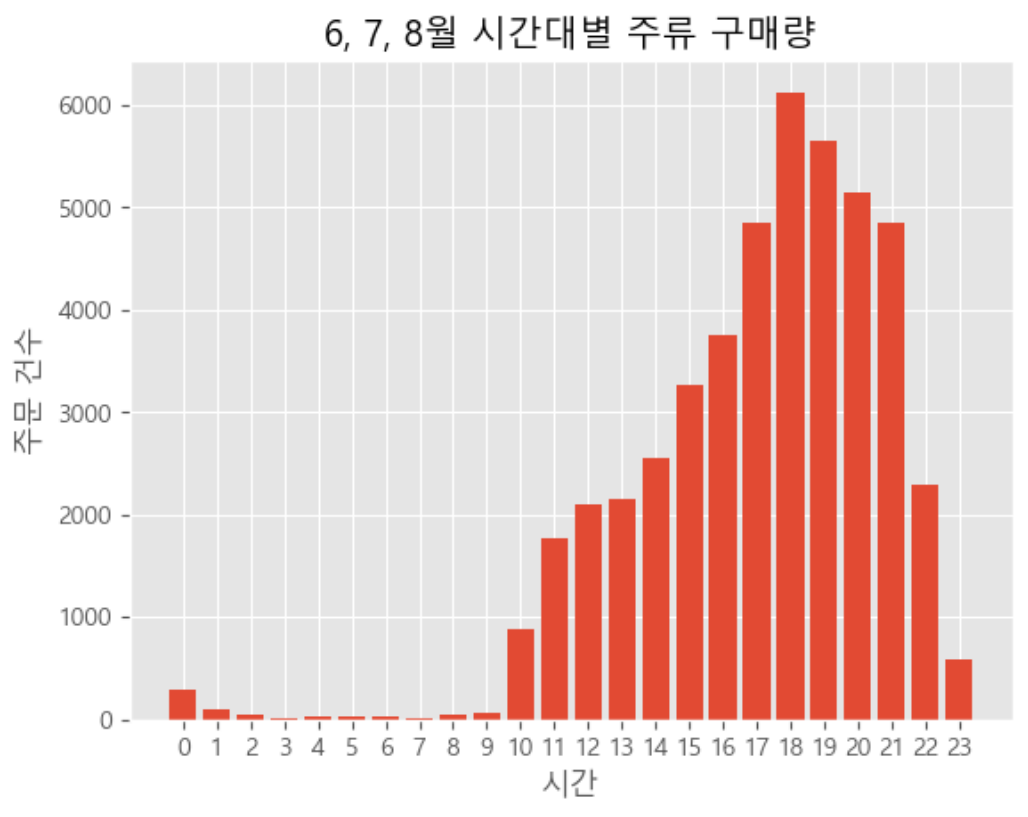
시간




```
liquor_orders_by_hour = df.groupby('buy_hour').size().reset_index()
liquor_orders_by_hour.columns = ['buy_hour', 'order_count']

plt.bar(liquor_orders_by_hour['buy_hour'], liquor_orders_by_hour['order_count'])
plt.xticks(range(24), range(24))
plt.xlabel('시간')
plt.ylabel('주문 건수')
plt.title('시간대별 주류 구매량')
plt.show()
```

분석과 인사이트

시간

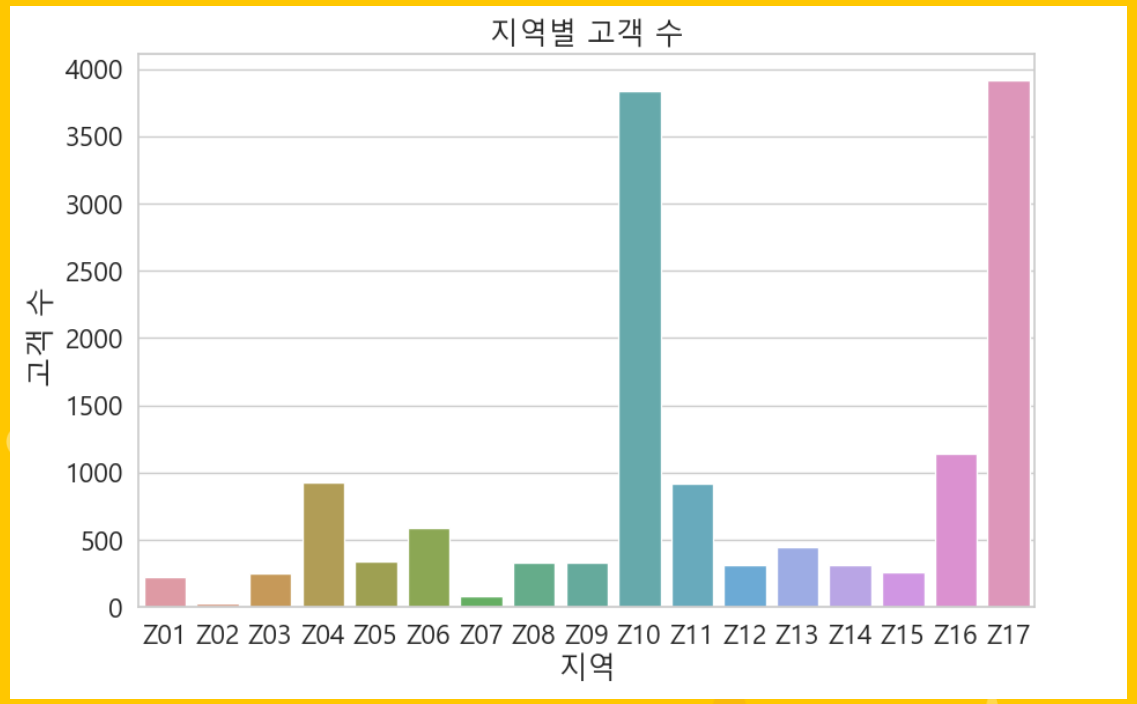



```python
# datetime
df['buy_date'] = pd.to_datetime(df['buy_date'], format='%Y%m%d')
# 6월부터 8월까지..
summer_df = df[(df['buy_date'].dt.month >= 6) & (df['buy_date'].dt.month <= 8)]

liquor_orders_by_hour = summer_df.loc[summer_df['large_product_cat'] == '주류'].groupby('buy_hour').size().reset_index()
liquor_orders_by_hour.columns = ['buy_hour', 'order_count']

plt.bar(liquor_orders_by_hour['buy_hour'], liquor_orders_by_hour['order_count'])
plt.xticks(range(24), range(24))
plt.xlabel('시간')
plt.ylabel('주문 건수')
plt.title('6, 7, 8월 시간대별 주류 구매량')
plt.show()
```

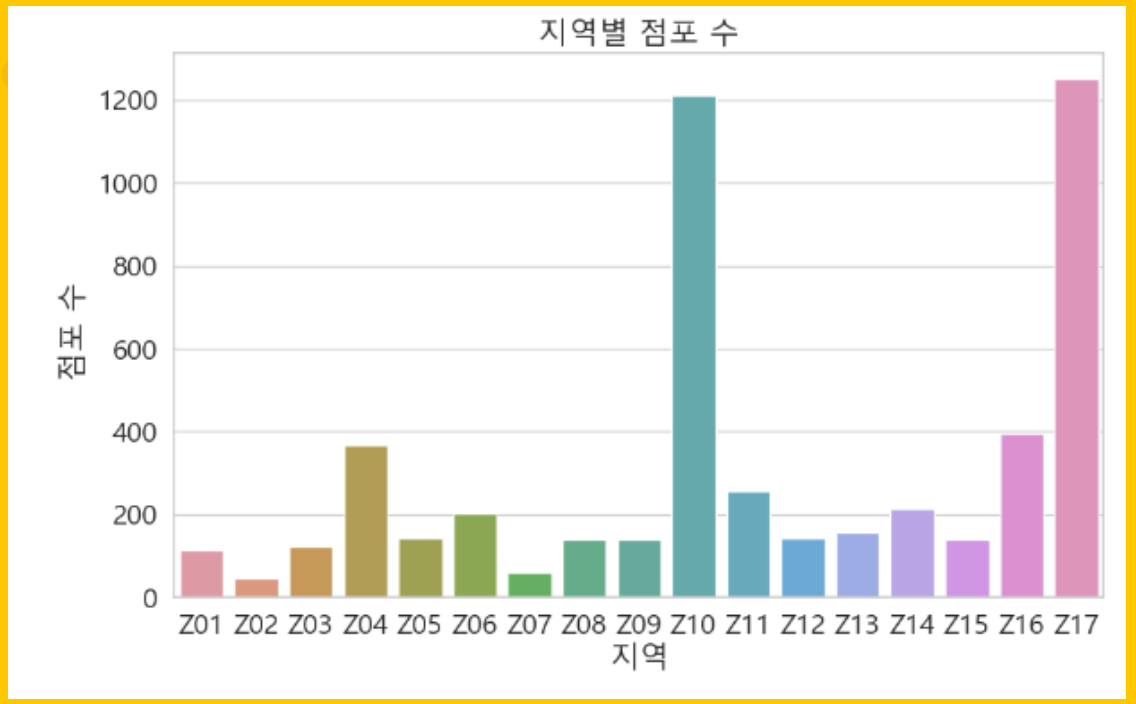
분석과 인사이트

지역

```
# 지역별 고객 수
cus = df.groupby('location')['customer_id'].nunique()
```




```
# 지역별 점포 수
store = df.groupby('location')['market_code'].nunique()
```

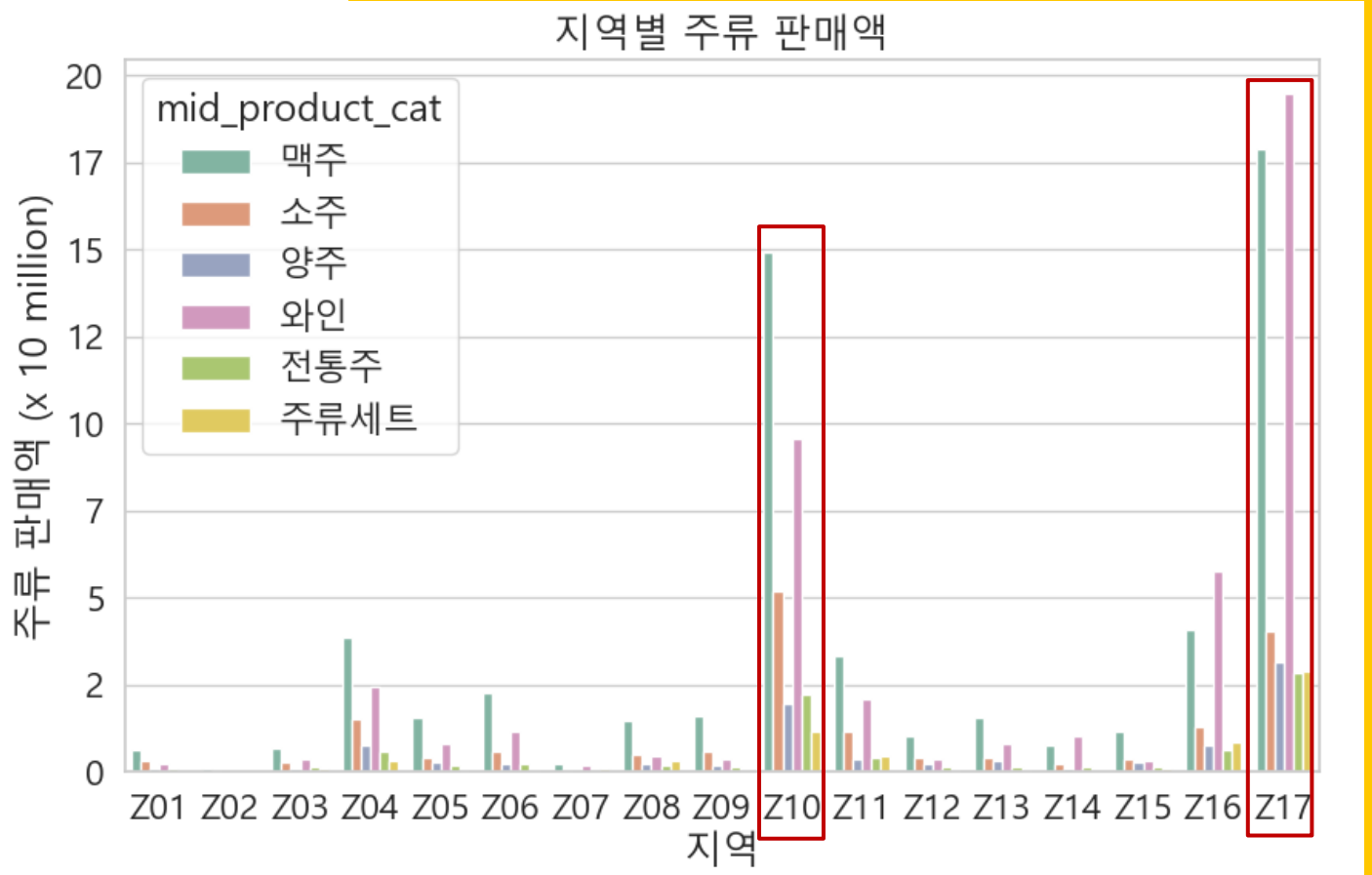


분석과 인사이트

지역

```
# 지역별 mid_product_cat 판매액
df_locammid = df.groupby(['location', 'mid_product_cat'])['buy_amount'].sum().reset_index()

df_locammid
```

	location	mid_product_cat	buy_amount
0	Z01	맥주	6356140.0
1	Z01	소주	3236330.0
2	Z01	양주	1213190.0
3	Z01	와인	2342530.0
4	Z01	전통주	975680.0
...	...	...	...
97	Z17	소주	40325510.0
98	Z17	양주	31693980.0
99	Z17	와인	194724380.0
100	Z17	전통주	28325700.0
101	Z17	주류세트	28722140.0



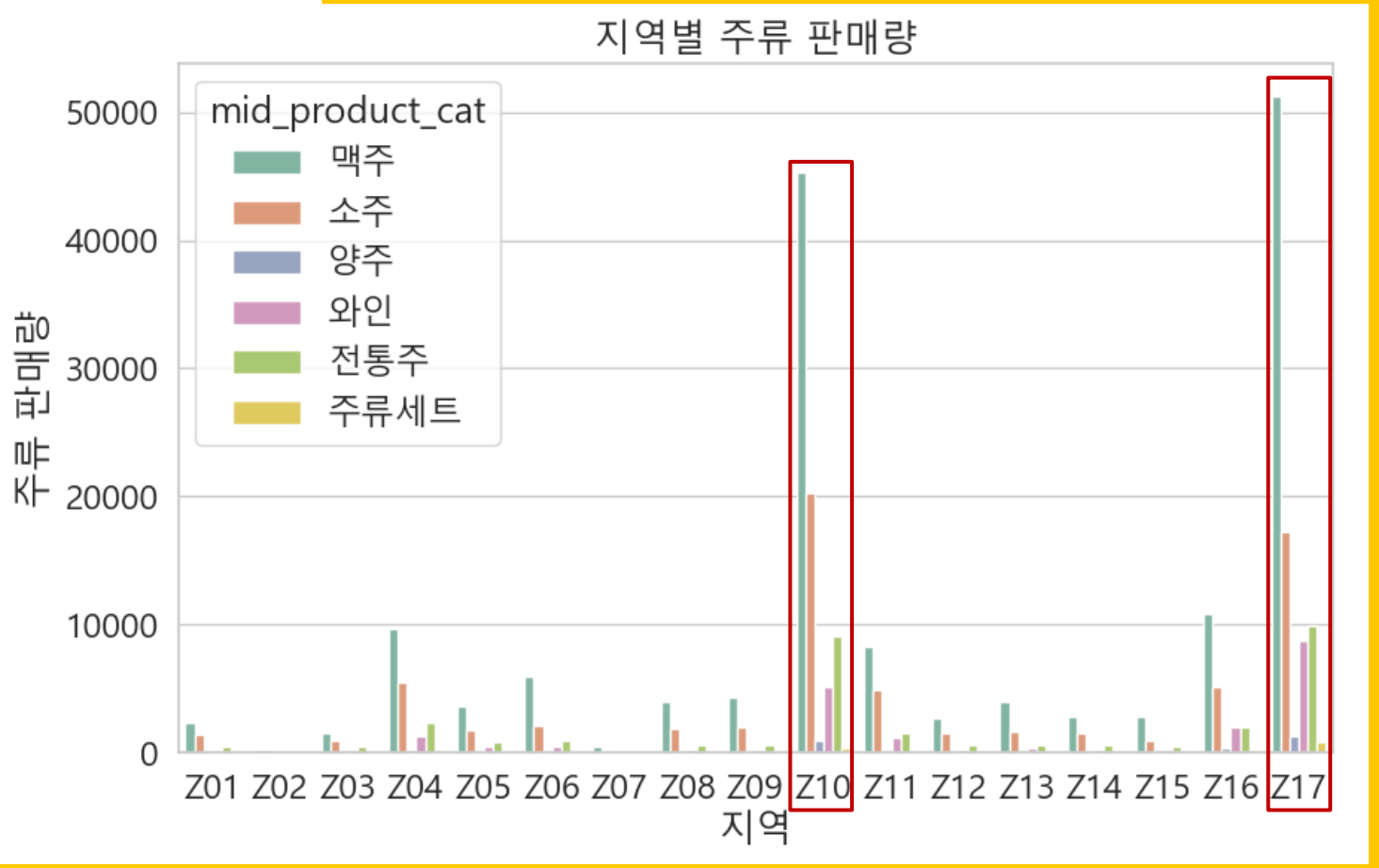
Z17, Z10에서 주류 판매액 높음
대부분 지역에서 맥주가 큰 판매액 비중 차지

분석과 인사이트

지역

```
# 지역별 mid_product_cat 판매량
df_loccmid = df.groupby(['location', 'mid_product_cat'])['buy_count'].sum().reset_index()

df_loccmid
```

	location	mid_product_cat	buy_count
0	Z01	맥주	2530
1	Z01	소주	1443
2	Z01	양주	70
3	Z01	와인	161
4	Z01	전통주	540
...	...	...	...
97	Z17	소주	17595
98	Z17	양주	1389
99	Z17	와인	9022
100	Z17	전통주	9908
101	Z17	주류세트	924



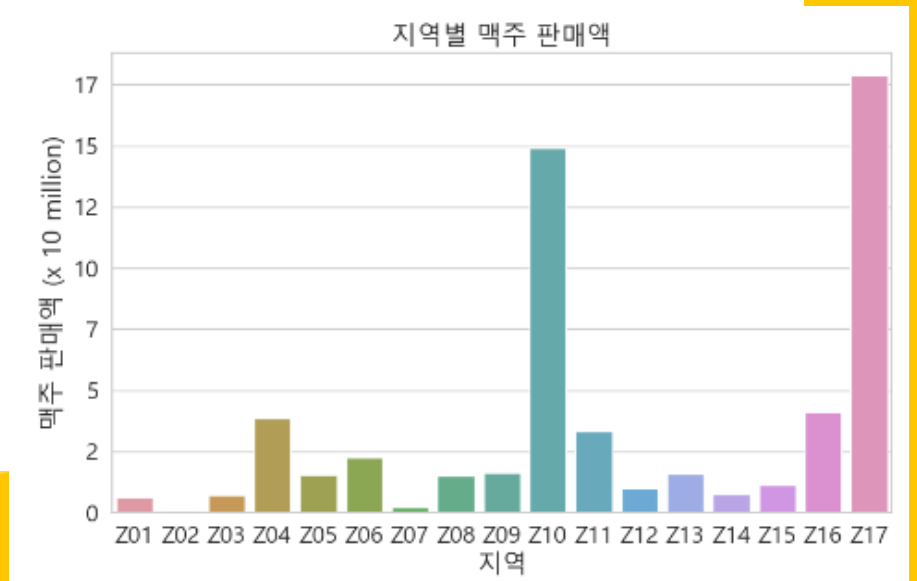
Z17, Z10에서 주류 판매량 높음
모든 지역에서 맥주의 판매량 가장 높음

분석과 인사이트

지역

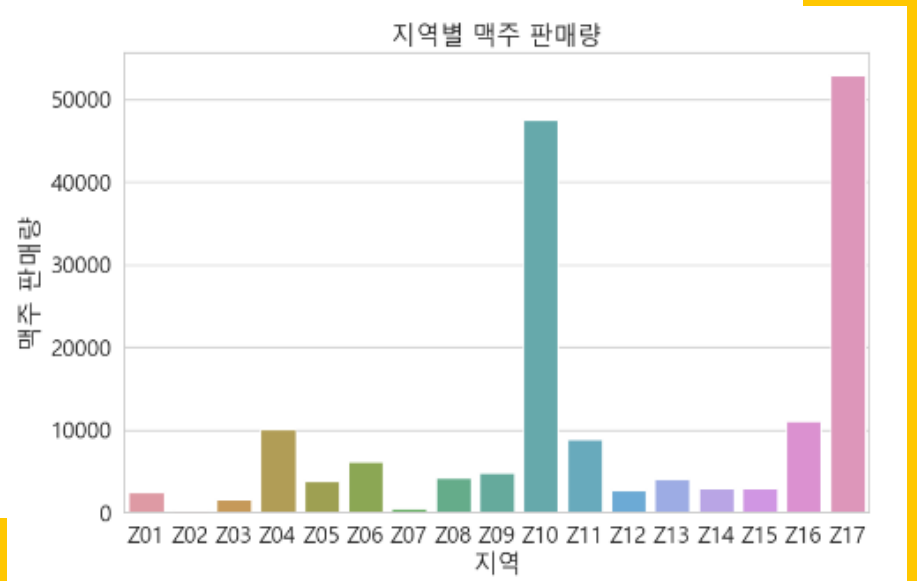
```
# 지역별 맥주 판매액
beer_locam = df[df['mid_product_cat'] == '맥주'].groupby('location')['buy_amount'].sum().reset_index()

beer_locam
```

	location	buy_amount
0	Z01	6356140.0
1	Z02	1129270.0
2	Z03	6951940.0
3	Z04	38611490.0
4	Z05	15593536.0
5	Z06	22698542.0
6	Z07	2130790.0
7	Z08	14922100.0
8	Z09	16160770.0
9	Z10	149117070.0
10	Z11	33315330.0
11	Z12	10097070.0
12	Z13	15783870.0
13	Z14	7698090.0
14	Z15	11498830.0
15	Z16	40967715.0
16	Z17	178923609.0




```
# 지역별 맥주 판매량
beer_locc = df[df['mid_product_cat'] == '맥주'].groupby('location')['buy_count'].sum().reset_index()

beer_locc
```

	location	buy_count
0	Z01	2530
1	Z02	310
2	Z03	1608
3	Z04	10069
4	Z05	3804
5	Z06	6214
6	Z07	559
7	Z08	4230
8	Z09	4810
9	Z10	47471
10	Z11	8876
11	Z12	2759
12	Z13	4132
13	Z14	2919
14	Z15	2986
15	Z16	11134
16	Z17	52927



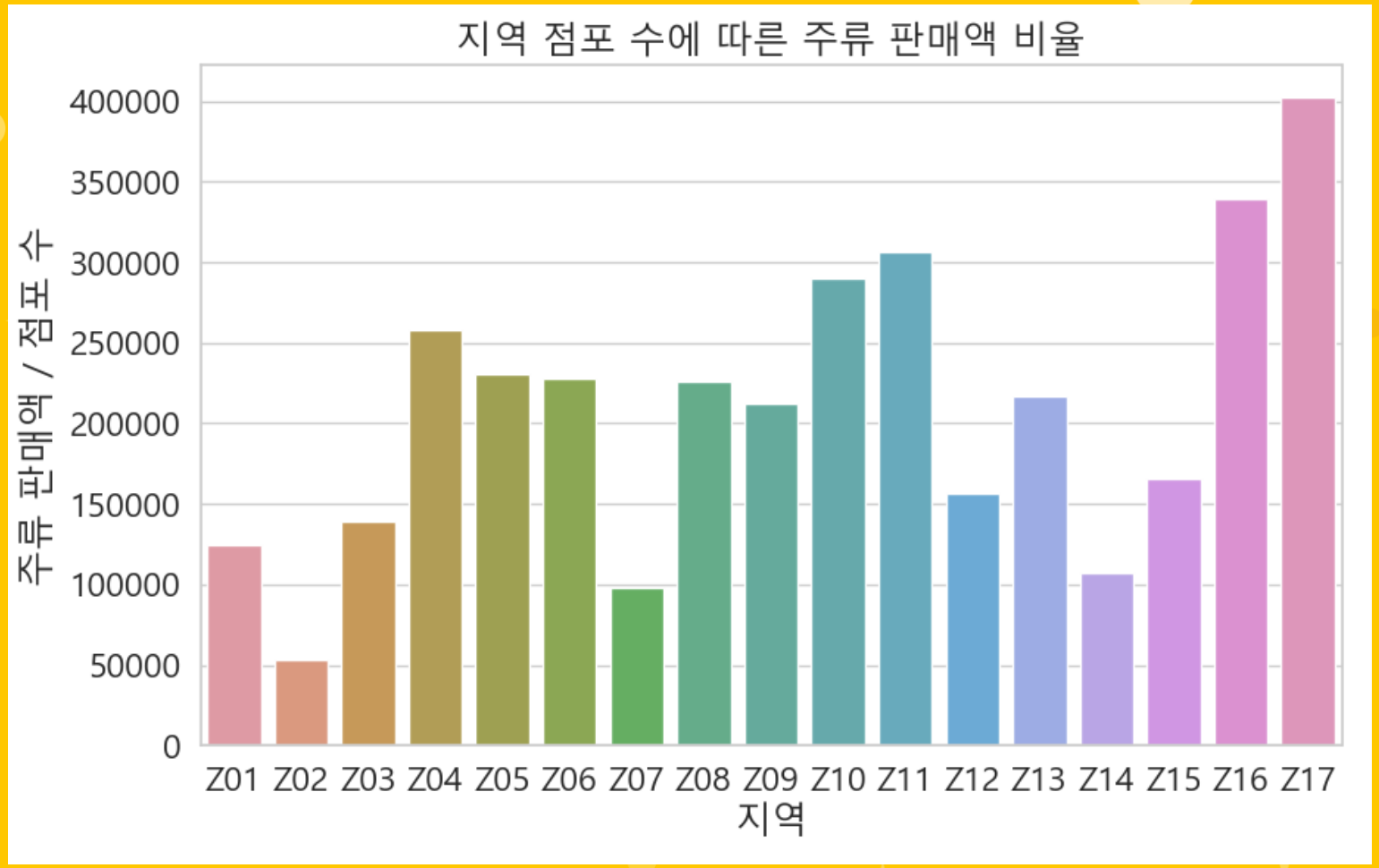
Z17, Z10에서 절반 이상의 판매량, 판매액 차지

분석과 인사이트

지역

```
# 지역별로 주류 판매액을 점포 수로 나눈 비율
store = df.groupby('location')['market_code'].nunique()
df_locam = df.groupby('location')['buy_amount'].sum()

loc_store = df_locam / store
loc_store
```

```
location
Z01    124092.719298
Z02     53421.914894
Z03    139090.650407
Z04    257549.630435
Z05    230748.492958
Z06    227734.437811
Z07     97898.000000
Z08    225817.535714
Z09    212660.571429
Z10    289781.809917
Z11    306777.500000
Z12    156549.929577
Z13    216913.461538
Z14    107443.906977
Z15    165985.428571
Z16    339097.734177
Z17    402172.255200
```



Z17에서 한 점포 당 주류 판매액 가장 높음

여름 신제품 활용

시기 고객 가격

수량 시간 지역

여름 신제품 활용

시기

'New'

1. 출시시기 -> '여름'

2. 판매량 1위!

=> 이번 신제품은 '맥주'

여름 신제품 활용

고객

✓ 타겟 -> 여성

+

=

여름 신제품 활용

고객


하이트진로
초깔끔한 맛
16.9도
眞露
진로
JINROSOJU
소주의 원조
진-로


7bräu
스위트하게~ 위트있게~
곰표
밀맥주
SINCE 1952
WHEATBEER
4.5% ALC / VOL. 500mL

여름 신제품 활용

가격

맥주는 특히나 가격과 판매량의 관계가 **강하다**
-> 95% 신뢰구간의 **하한** 신뢰경계로 가격 설정

```
def confidence_interval(data, confidence):
    data=np.array(data)
    mean=np.mean(data)
    n=len(data)
    stderr=stats.sem(data)
    interval=stderr*stats.t.ppf((1+confidence)/2, n-1)
    return (mean, mean-interval, mean+interval)
```

```
confidence_interval(beer['price'], confidence=0.95)
```

```
(4328.574466510286, 4299.144072888578, 4358.004860131993)
```

맥주 가격이 95% 확률로
약 4200원~ 약 4400원 사이에 존재
-> 신제품 가격 : **4200원**

4개 할인 판매

3개 사는 것보다 4개 사도록 유도 -> 더 많은 매출이익

주류 구매량은 6시 이후로 감소

타임세일
8시 ~ 10시

여름 신제품 활용

지역

맥주 신제품을 판매량이 높은
Z17, Z10순으로 차등 분배

많은 유동 인구, 점포 당 주류 판매액 높은
Z17 지역 집중공략

감사
합니다