DArt-B 5기 사전과제

Global_Supermarket Data

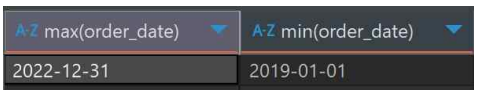
소프트웨어학부
20210591
조상원

목차

(A) EDA(Exploratory Data Analysis)

(a) 데이터 이해

주어진 데이터셋은 SQL문을 활용하여 확인한 결과 2019년부터 2022년까지의 데이터임을 확인하였다.



<해당 데이터셋의 전체 기간 확인>

Python을 사용하여 각 데이터가 가진 정보들과 Data Type을 확인하였다.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customer_id       51290 non-null  object
 1   customer_name     51290 non-null  object
 2   customer_segment  51290 non-null  object
 3   order_id          51290 non-null  object
 4   order_city        51290 non-null  object
 5   oreder_region     51290 non-null  object
 6   order_date        51290 non-null  object
 7   order_year        51290 non-null  int64
 8   order_weeknum     51290 non-null  int64
 9   quantity          51290 non-null  int64
 10  sales             51290 non-null  int64
 11  product_id        51290 non-null  object
 12  product_name      51290 non-null  object
 13  profit            51290 non-null  float64
 14  discount          51290 non-null  float64
 15  category          51290 non-null  object
 16  sub_category      51290 non-null  object
 17  market_country    51290 non-null  object
 18  market_area       51290 non-null  object
 19  market_city       51290 non-null  object
 20  ship_date         51290 non-null  object
 21  ship_mode         51290 non-null  object
 22  shipping_cost     51290 non-null  float64
 23  row_id            51290 non-null  int64
dtypes: float64(3), int64(5), object(16)
```

<info()>

(b) 가설 설정 및 검증

우선 Consumer(개인 고객), Corporate(기업체), Home Office(재택근무 사무실)로 나누어진 customer_segment(고객 유형)가 '각각의 고객 유형별로 뚜렷한 차이를 보일 것이다'라는 가설을 세우고 이를 확인하기 위해 고객들을 나눈 후 각각 SQL을 사용해 분석해 보았다.
각각의 고객들의 수는 다음과 같았다.

	A-Z customer_segment	123 count(*)
1	Consumer	26,518
2	Corporate	15,429
3	Home Office	9,343

<각각의 customer_segment의 수>

총 13개의 지역으로 이루어진 order_region을 각각의 고객 유형들의 상위 3개 지역을 출력한 결과 뚜렷한 차이를 보이지 않았다.

	A-Z customer_segment	A-Z oreder_region	123 count
1	Consumer	Central	5,782
2	Consumer	South	3,479
3	Consumer	EMEA	2,538
4	Corporate	Central	3,321
5	Corporate	South	1,998
6	Corporate	EMEA	1,574
7	Home Office	Central	2,014
8	Home Office	South	1,168
9	Home Office	EMEA	917

<각각의 customer_segment의 가장 count가 높은 상위 3개의 order_region>

category를 기준으로 각각의 고객 유형을 분석해 본 결과, 역시 뚜렷한 차이를 보이지 않았다.

	A-Z customer_segment	A-Z category	123 count
1	Consumer	Office Supplies	16,151
2	Consumer	Technology	5,272
3	Consumer	Furniture	5,095
4	Corporate	Office Supplies	9,364
5	Corporate	Technology	3,051
6	Corporate	Furniture	3,014
7	Home Office	Office Supplies	5,758
8	Home Office	Technology	1,818
9	Home Office	Furniture	1,767

<각각의 customer_segment의 가장 count가 높은 상위 3개의 category>

각각의 고객 유형별로 quantity(주문량), sales(주문 총 금액), profit(주문 이익), discount (해당 주문에 대한 할인)의 합과 평균을 구해보았다.

각각 Consumer, Corporate, Home Office 순으로 합의 값이 줄어들었지만, 이는 전체 고객의 비율을 고려해보았을 때 유의미한 결과가 아니라고 볼 수 있다. 평균값 또한 3개의 고객 유형에서 비슷한 값을 보이기 때문에 합(SUM)의 값은 전체 비율에 따른 차이라고 볼 수 있다.

	customer_segment	SUM(quantity)	AVG(quantity)
1	Consumer	92,157	3.4752620861
2	Corporate	53,565	3.4717091192
3	Home Office	32,590	3.4881729637

<각각의 customer_segment의 quantity의 합과 평균>

	customer_segment	SUM(sales)	AVG(sales)
1	Consumer	6,508,141	245.4235236443
2	Corporate	3,824,808	247.8973361851
3	Home Office	2,309,956	247.2392165257

<각각의 customer_segment의 sales의 합과 평균>

	customer_segment	SUM(profit)	AVG(profit)
1	Consumer	749,239.78206	28.2540079214
2	Corporate	441,208.32866	28.5960417824
3	Home Office	277,009.18056	29.6488473253

<각각의 customer_segment의 profit의 합과 평균>

	customer_segment	SUM(discount)	AVG(discount)
1	Consumer	3,808.042	0.143602157
2	Corporate	2,205.284	0.1429311038
3	Home Office	1,316.402	0.1408971422

<각각의 customer_segment의 discount의 합과 평균>

즉 고객 유형별로 구매 경향에 차이를 보일 것이라는 가설은 틀림을 확인하였다. 그렇다면 클러스터링을 진행하여 고객들을 Grouping한 다음 각 고객군에 따라 구매 경향에 차이를 보일 것이라는 새로운 가설을 설정할 수 있다.

(c) 데이터 전처리

Data type 확인 및 변환

(a)에서 각 Column의 데이터 타입을 확인하였다. order_date의 경우 '2019-01-07'와 같이 날짜 형식으로 데이터가 저장되어 있다. 이러한 경우 정수형으로 변환하여 데이터를 다루는데 효율적으로 바꿀 수 있다.

범주형 데이터 처리

category, sub_category, market_country, market_area, quantity의 경우 범주형 데이터이다. OneHotEncoder를 사용해 범주형 변수들을 수치형 데이터로 변환하기 위해 One-Hot 인코딩을 진행하였다.

수치형 Data Type 변경 및 스케일링

수치형 컬럼 sales와 days_since의 데이터 타입을 np.float32로 변환하여 메모리 사용량을 줄이고, 일관된 데이터 타입으로 만든다. 이후, StandardScaler를 이용하여 각 수치형 데이터를 평균 0, 분산 1로 스케일링한다.

최종 변수 선정

주어진 데이터셋은 총 23개의 column으로 이루어짐을 확인하였고 각각의 Data Type을 확인할 수 있었다. 우리의 목적은 EDA를 진행 후 이 결과를 바탕으로 기업에 도움이 될 수 있는 문제 해결 방안이나 마케팅 전략 등 비즈니스 인사이트를 도출해내는 것이다. 이를 위해서는 모든 column이 필요하다고 판단하지 않았으며 필요한 column들을 추려보았다.

최종 선정 column:

customer_segment, order_id, order_region, order_date, order_year, quantity, sales, profit, discount, category, sub_category, market_country, market_area, ship_mode

(d) 클러스터링

전처리를 진행 후 클러스터링을 진행하였다.

총 3가지 방법으로 클러스터링을 진행하였다.

1. Elbow Method를 활용한 최적 K 찾기
2. K-Means를 활용한 클러스터링
3. PCA를 활용한 클러스터링

클러스터링을 진행할 column들을 선정하기 위해 RFM (Recency, Frequency, Monetary) 분석을 활용하였다. 이는 고객 Data를 기반으로 한 마케팅 기법 중 하나이다. 고객의 구매 행동을 3가지 요소로 분류하여 고객의 가치를 평가하는 방식이다, 이는 비교적 간단하면서도 효과적이어서 다양한 산업 분야에서 충성도 높은 고객을 식별하고, 개인화된 마케팅 전략을 수립하는 데 널리 사용된다.

RFM 요소

Recency (최근성)

- 고객이 가장 최근에 구매한 시점
- 일반적으로 '일수'로 측정하며, 최근에 구매할수록 점수가 높아진다.
- 최근성은 고객이 여전히 활동적이고 관심이 높다는 지표로 사용된다.

Frequency (빈도)

- 특정 기간 동안 고객이 구매한 횟수
- 빈도가 높을수록 고객 충성도가 높다고 간주하며, 이는 고객의 브랜드에 대한 관심과 만족을 반영한다.

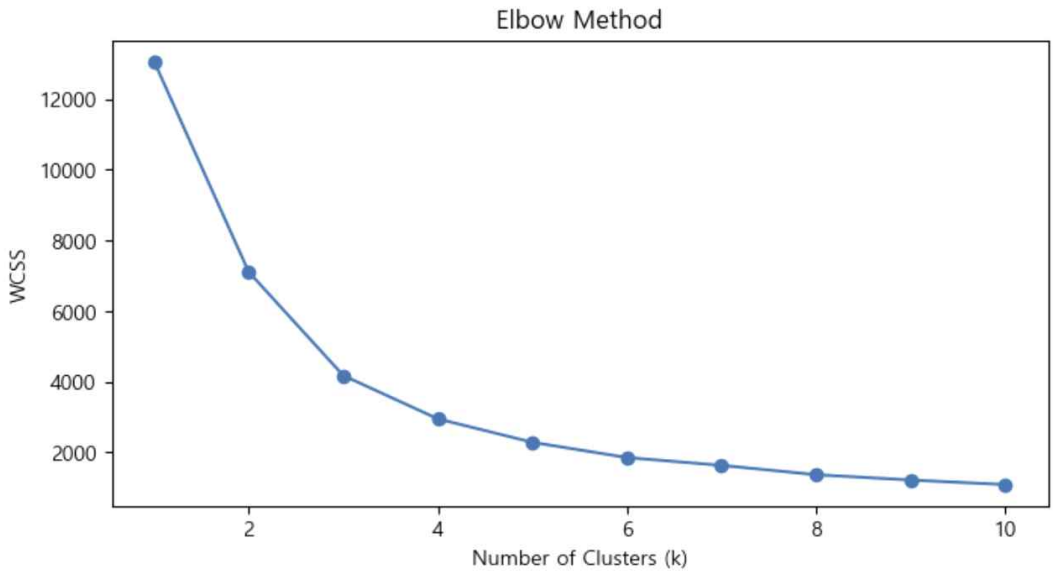
Monetary (금액)

- 고객이 특정 기간 동안 총 지출한 금액
- 금액이 클수록 고객의 경제적 가치가 높다고 평가한다.

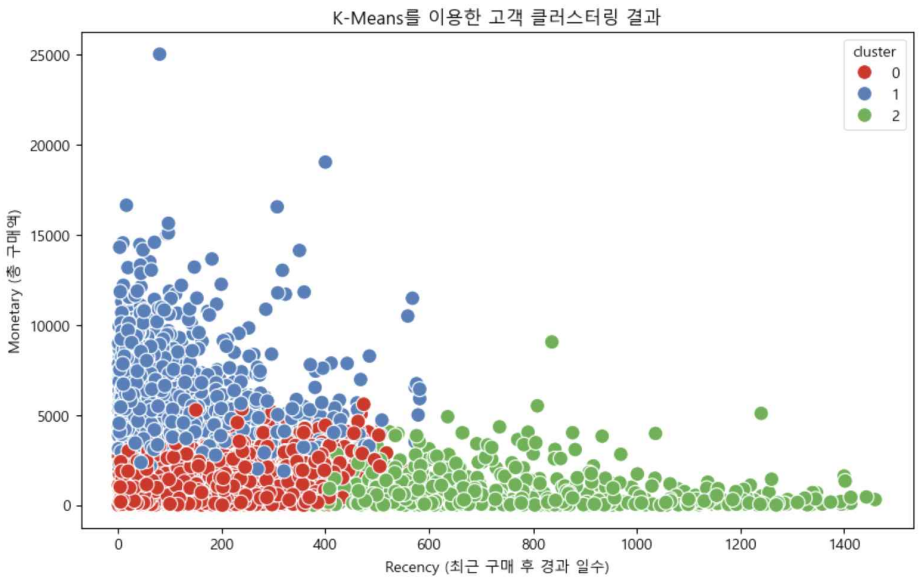
주어진 Global_Supermarket.csv 데이터에서는 Recency (최근성), Frequency (빈도), Monetary (금액) 대신 order_date (주문 날짜), order_id (order_id의 고유값 개수), sales (sales 합계)를 사용하였다.

(d)-1 Elbow Method를 활용한 최적 K 찾기

Elbow Method를 이용해 최적의 클러스터 개수를 선정하였다. 그래프에서 감소폭이 줄어드는 Elbow Point(K=3)을 확인하였다. K=3을 최적의 클러스터 개수로 판단하였다.



(d)-2 K-Means를 활용한 클러스터링



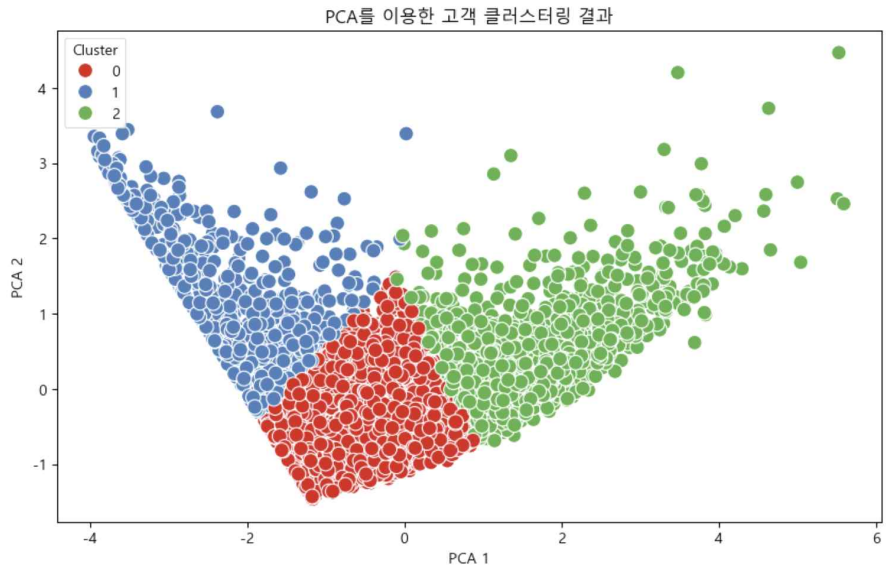
클러스터별 평균 RFM 값:

cluster	recency	frequency	monetary
0	140.778741	4.273276	1511.757256
1	95.143964	8.369916	5238.622983
2	730.116244	2.162444	771.365127

클러스터별 RFM분석

- **클러스터 0**: 상대적으로 최근에 구매한 고객들로 구성되어 있지만, 구매 빈도는 평균적이며 구매 금액도 중간 수준이다.
- **클러스터 1**: 빈도가 높고 구매 금액도 매우 높다. 이 클러스터는 가장 가치 있는 고객을 포함할 가능성이 높다.
- **클러스터 2**: 구매한 지 가장 오래된 고객들로 구성되어 있으며, 구매 빈도와 금액 모두 낮다. 이 클러스터는 잠재적으로 이탈할 위험이 있는 고객들로 구성될 수 있다.

(d)-3 PCA를 활용한 클러스터링



클러스터별 평균 RFM 값:

cluster	recency	frequency	monetary
0	140.067466	4.336582	1516.020630
1	721.982533	2.141194	799.173218
2	92.931489	8.374177	5302.457031

클러스터별 RFM분석

- **클러스터 0**: 중간 정도의 최근성, 빈도가 평균적이며, 구매 금액도 평균적인 고객들로 구성되어 있다.
- **클러스터 1**: 구매한 지 가장 오래된 고객들로 구성되어 있으며, 구매 빈도와 금액이 가장 낮습니다. 이들은 잠재적으로 마케팅 개입이 필요한 고객층일 수 있다.
- **클러스터 2**: 최근에 구매를 많이 하고, 많은 금액을 지출한 고객들로 구성되어 있다.

(d)-4 클러스터링 성능 평가

클러스터링 성능 평가를 위해 실루엣 점수를 활용하였다.

실루엣 점수는 클러스터링의 품질을 평가하는 지표 중 하나로, 클러스터링 결과가 얼마나 잘 수행되었는지를 수치적으로 나타낸다. 이 점수는 각 데이터의 클러스터 내 응집도와 클러스터 간 분리도를 측정하여 계산됩니다. 실루엣 점수는 -1에서 1 사이의 값을 가지며, 이 값이 높을수록 클러스터링 결과가 더 좋은 것으로 해석할 수 있다.

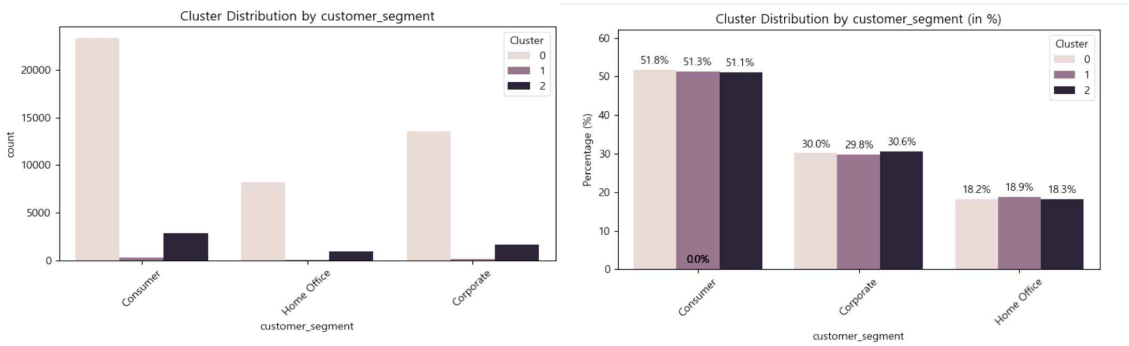
K=3에서 0.5308로 가장 높음을 알 수 있으며 이는 적절한 클러스터 구조를 반영한다고 볼 수 있다.

```
클러스터개수: 2, 실루엣점수: 0.4835
클러스터개수: 3, 실루엣점수: 0.5308
클러스터개수: 4, 실루엣점수: 0.4306
클러스터개수: 5, 실루엣점수: 0.4231
클러스터개수: 6, 실루엣점수: 0.4292
클러스터개수: 7, 실루엣점수: 0.4500
```

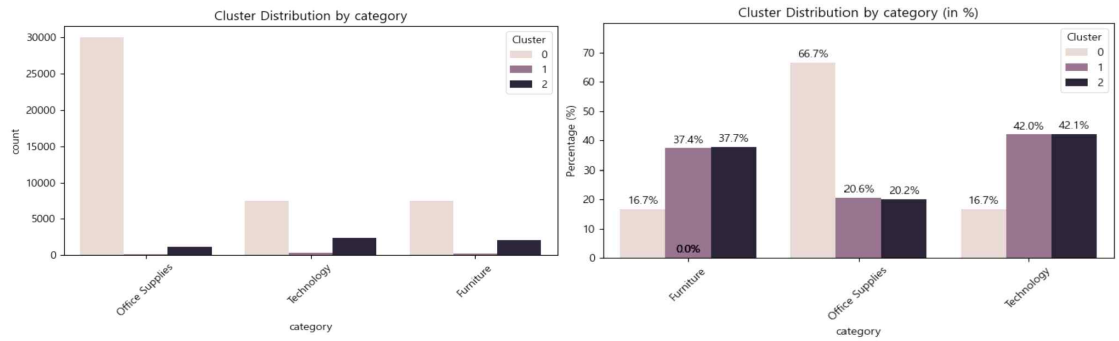
<클러스터 개수별 실루엣 점수>

(e) 결과 시각화

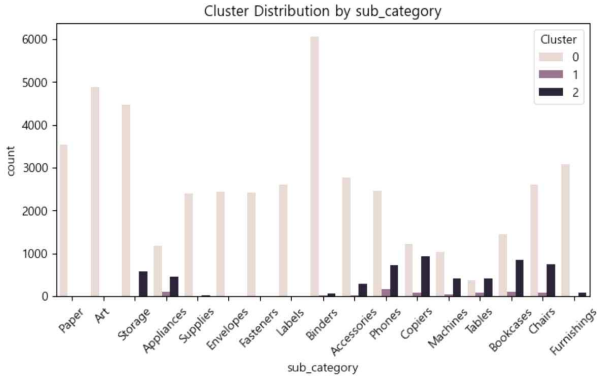
상대적으로 고객들의 특징이 잘 나타난 K-Means를 이용해 구분한 클러스터들을 각 항목별로 분류하여 막대그래프로 시각화하였다.



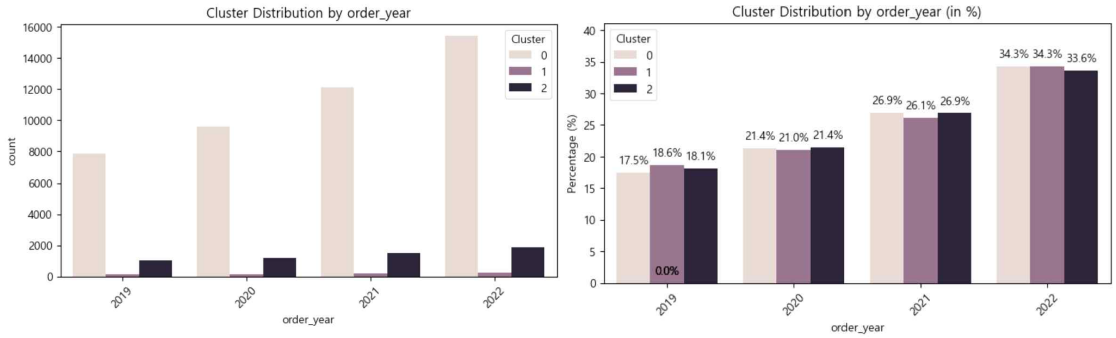
<고객 유형별 클러스터 분포와 각 클러스터별 차지하는 비율>



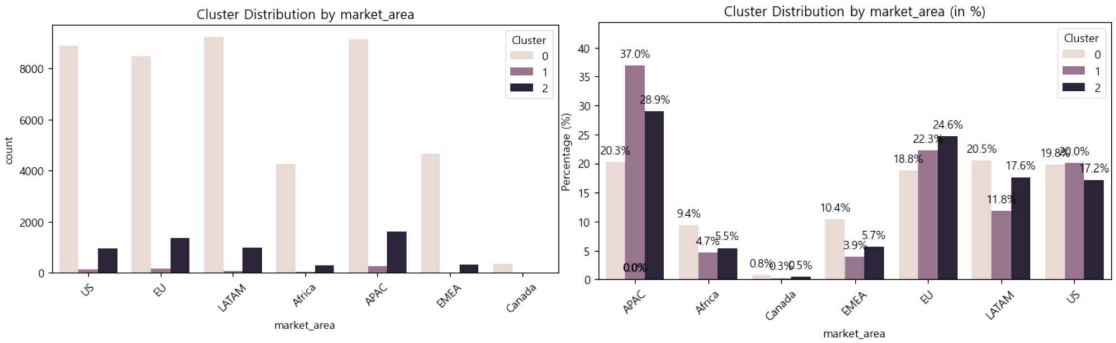
<카테고리별 클러스터 분포와 각 클러스터별 차지하는 비율>



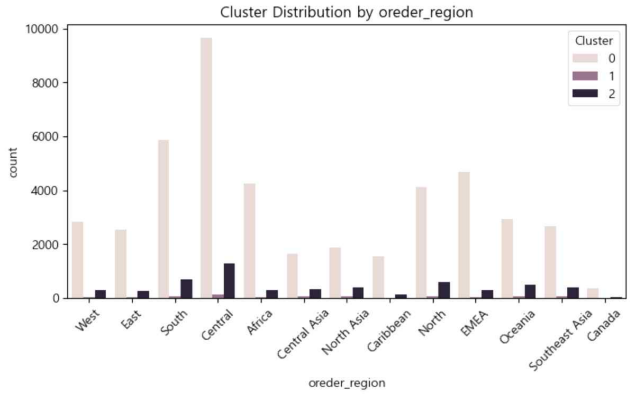
<상품의 서브 카테고리별 클러스터 분포와 각 클러스터별 차지하는 비율>



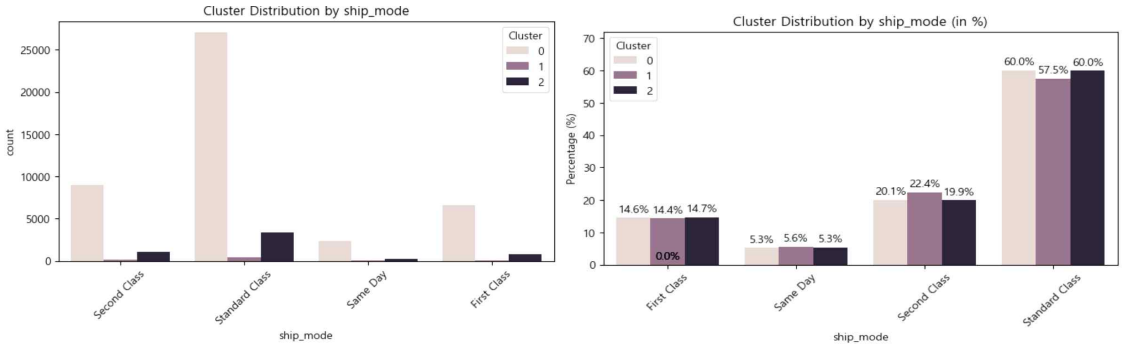
<주문 연도별 클러스터 분포와 각 클러스터별 차지하는 비율>



<market_area별 클러스터 분포와 각 클러스터별 차지하는 비율>

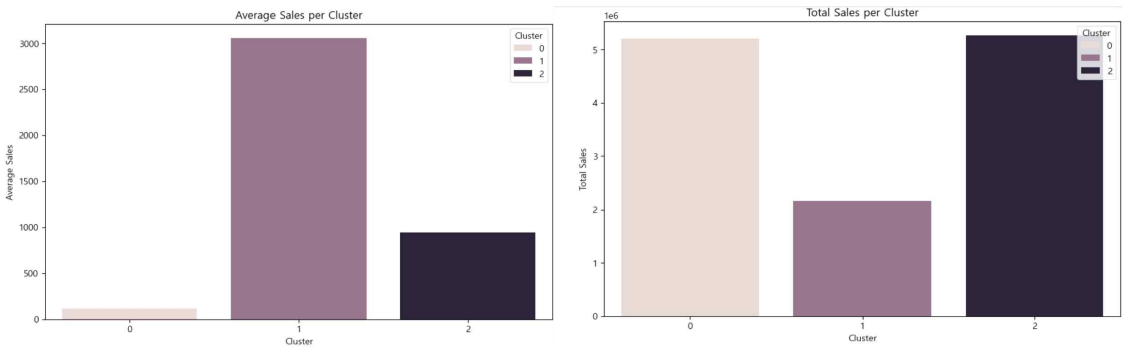


<order_region별 클러스터 분포와 각 클러스터별 차지하는 비율>

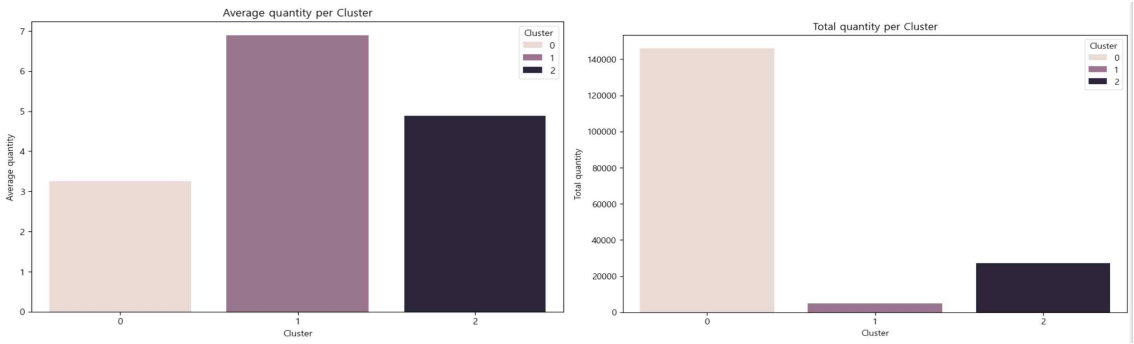


<운송 종류별 클러스터 분포와 각 클러스터별 차지하는 비율>

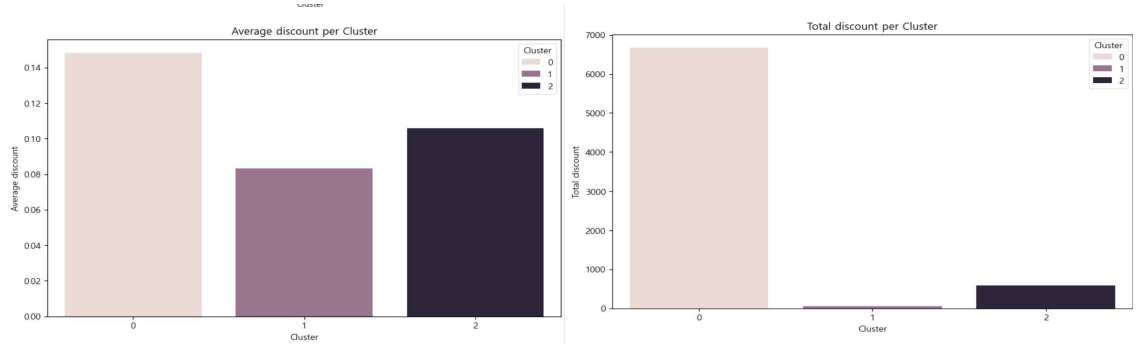
sales, quantity, discount, profit 같은 경우에는 평균값과 총합계를 구해 비교해보았다.



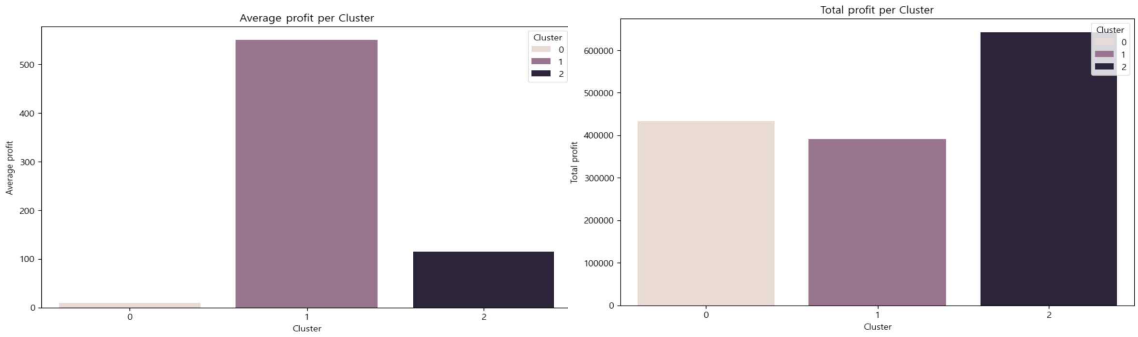
<클러스터별 주문 총 금액의 평균값과 총합계>



<클러스터별 주문량의 평균값과 총합계>

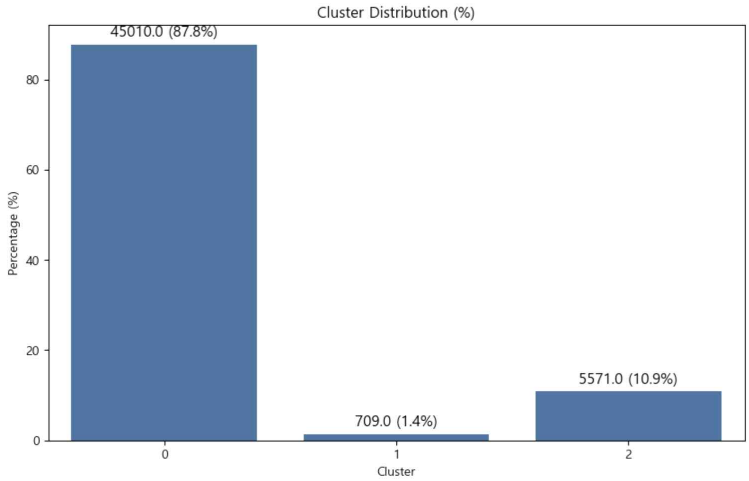


<클러스터별 해당 주문에 대한 할인의 평균값과 총합계>



<클러스터별 주문 이익의 평균값과 총합계>

다음은 전체 데이터 중 해당 클러스터가 차지하는 비율과 실제 수치를 나타낸 그래프이다.



<전체 클러스터의 비율 및 실제값 비교>

(B) 비즈니스 인사이트 도출

(a) 클러스터별 특성 비교

클러스터 0	클러스터 1	클러스터 2
87.8%	1.4%	10.9%
주문 총금액의 총합계 2위 주문 총금액의 평균값 3위 **주문량의 총합계 1위** 주문량의 평균값 3위 **할인의 총합계 1위** **할인의 평균값 1위** 주문 이익의 총합계 2위 주문 이익의 총합계 3위	주문 총금액의 총합계 3위 **주문 총금액의 평균값 1위** 주문량의 총합계 3위 **주문량의 평균값 1위** 할인의 총합계 3위 할인의 평균값 3위 주문 이익의 총합계 3위 **주문 이익의 평균값 1위**	**주문 총금액의 총합계 1위** 주문 총금액의 평균값 2위 주문량의 총합계 2위 주문량의 평균값 2위 할인의 총합계 2위 할인의 평균값 2위 **주문 이익의 총합계 1위** 주문 이익의 평균값 2위
#Office_Suplies #Standard_Class	#APAC #VIP	#Technology #주문량 증가폭이 상대적으로 작아지는 추세 #잠재적 이탈 위험

(b) 마케팅 전략 제안

클러스터 0

할인 시스템: 할인의 총합계와 평균값이 모두 1위인 집단이다. 할인을 활용하여 특정 개수 이상 구매 시 할인율 상승 등 할인율을 계단식으로 적용하여 주문량과 주문 금액을 높이는 방법을 적용한다.

Cross-selling: Office Supplies의 구매 빈도가 높은 집단이다. Office Supplies 구매 고객에게 Furniture나 Technology 제품을 추천하여 구매를 유도한다. Furniture나 Technology 제품을 구매시에는 Office Supplies 제품에 할인을 해주는 등의 방식으로 할인 연계 시스템을 적용한다.

재구매 촉진 시스템: 자주 주문하는 고객을 대상으로 재주문을 편리하게 할 수 있는 시스템을 적용한다. 구독 서비스를 진행하여 장기적인 고객 관계 유지에 도움을 준다.

클러스터 1

지역 맞춤형 프로모션: APAC 지역의 문화와 소비 패턴에 맞춘 특별 프로모션을 기획하여 개인화된 마케팅을 강화한다.

VIP 고객 관리 프로그램: 고가의 제품을 자주 구매하는 이 클러스터를 대상으로 프리미엄 고객 서비스를 제공한다.

고객 참여 이벤트: 온라인 및 오프라인 VIP 고객 이벤트를 통해 브랜드에 대한 충성도를 높인다.

클러스터 2

재활성화 캠페인: 구매 감소 추세에 있는 이 고객들을 대상으로 웰컴 쿠폰 등 맞춤형 이벤트를 통해 고객의 관심을 다시 끌고 재구매를 유도한다.

맞춤형 제품 세트화 및 할인: Technology의 구매율이 높으며 주문 이익 또한 높은 집단이다. Technology 제품들을 세트화시키거나 Technology 제품들을 연속적으로 구매시 할인율을 높이는 방식으로 Technology 제품들의 구매량을 더욱 높인다.