Global Supermarket Analysis

Q1. 해당 데이터의 EDA를 진행하시오.

Basic Data Information

Data Shape : 26개의 특성(Features) , 총 51290개의 Data

Data Summary

- Missing Value : 0개
- 중복된 값: 0개

	data type	#missing	%missing	#unique	min	max
customer_id	object	0	0	4873	NaN	NaN
customer_name	object	0	0	795	NaN	NaN
customer_segment	object	0	0	3	NaN	NaN
order_id	object	0	0	25035	NaN	NaN
order_city	object	0	0	3636	NaN	NaN
order_region	object	0	0	13	NaN	NaN
order_date	datetime64[ns]	0	0	1430	2019-01-01	2022-12-31
order_year	int64	0	0	4	2019.0	2022.0
order_weeknum	int64	0	0	53	1.0	53.0
quantity	int64	0	0	14	1.0	14.0
sales	int64	0	0	2246	0.0	22638.0
product_id	object	0	0	10292	NaN	NaN
product_name	object	0	0	3788	NaN	NaN
profit	float64	0	0	24575	-6599.978	8399.976
discount	float64	0	0	27	0.0	0.85
category	object	0	0	3	NaN	NaN
sub_category	object	0	0	17	NaN	NaN
market_country	object	0	0	147	NaN	NaN
market_area	object	0	0	7	NaN	NaN
market_city	object	0	0	1094	NaN	NaN
ship_date	datetime64[ns]	0	0	1464	2019-01-03	2023-01-07
ship_mode	object	0	0	4	NaN	NaN
shipping_cost	float64	0	0	16877	0.002	933.57
row_id	int64	0	0	51290	1.0	51290.0
pre_sales	float64	0	0	6601	0.0	45276.0
uni_cost	float64	0	0	9285	0.0	7546.0

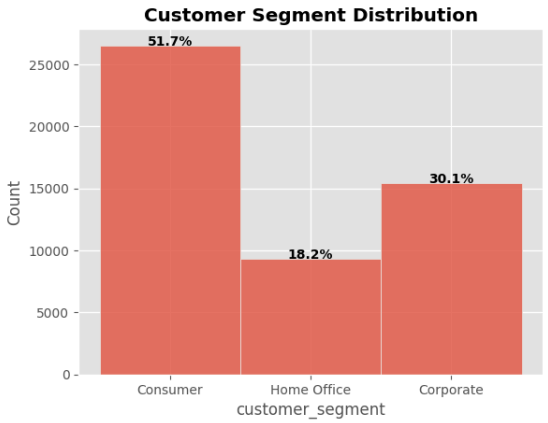
총 6개의 큰 소 주제로 나누어 EDA를 진행하였으며, 추가 Feature (`pre_sales,uni_cost`)를 생성하여 더욱 심도 있는 EDA를 진행하였습니다.

```
data['pre_sales'] = data['sales'] / (1-data['discount']) #Discount 적용 전 Sales
data['uni_cost'] = data['pre_sales'] / data['quantity'] # product 개당 원가
```

1. Customer Information (ID,name, Segement)

```
print('Unique 고객 ID:',data['customer_id'].nunique())
print('Unique 고객 이름:',data['customer_name'].nunique())
---------------------------------------
Unique 고객 ID: 4873
Unique 고객 이름: 795
```

슈퍼마켓을 이용하는 고객의 고유 ID는 총 4873개이며, 고유한 고객의 이름은 795개이다.



- Consumer: 개인 고객
- Corporate: 기업체
- Home Office: 재택 근무 사무실

- 각 고객 Segment의 비율을 나타낸 그래프이다. 개인 고객의 비율이 절반 이상을 차지하며, 그 다음은 기업체가 30% 정도의 비율을 차지한다.

2. Order Information

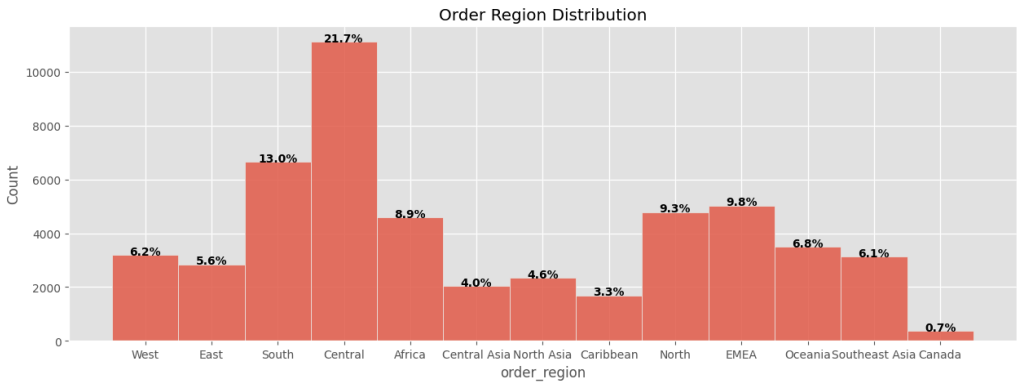
- Order와 관련된 정보는 크게 3가지로 나뉘어서 분석을 진행한다.

```
print('Unqiue order_id:',data['order_id'].nunique())
print('중복된 데이터 갯수: ',data.duplicated().sum())
-----------------------------------------------------
Unqiue order_id: 25035
중복된 데이터 갯수: 0
```

기록된 고유한 주문 ID는 총 25035개로, 중복되어 기록된 Data는 존재하지 않는다. 따라서 동일한 Order_ID 안에 서로 다른 주문 품목이 기록되어 있음을 알 수 있다.

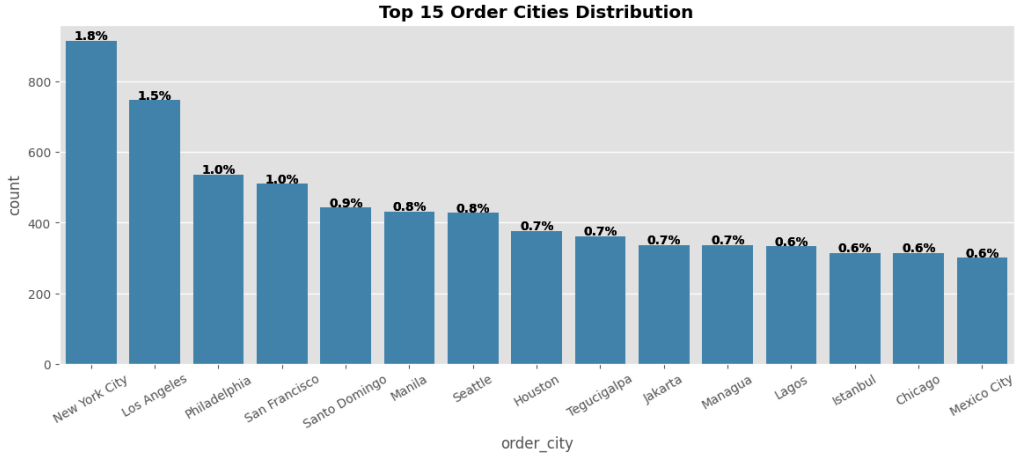
2-1. 지역별 특성 (Region, Cities)

Order Region



주문이 기록된 구역은 **Central의 비율이 가장 높고**, Canada에서 주문한 건수가 가장 낮다. 여기서 "Central,West,East",등 원래는 USA 내의 값으로 해석했으나, Berlin의 값이 Central로 기록되는 등, 미국 주위의 지역까지 포함하는 포괄적 개념으로 해석된다.

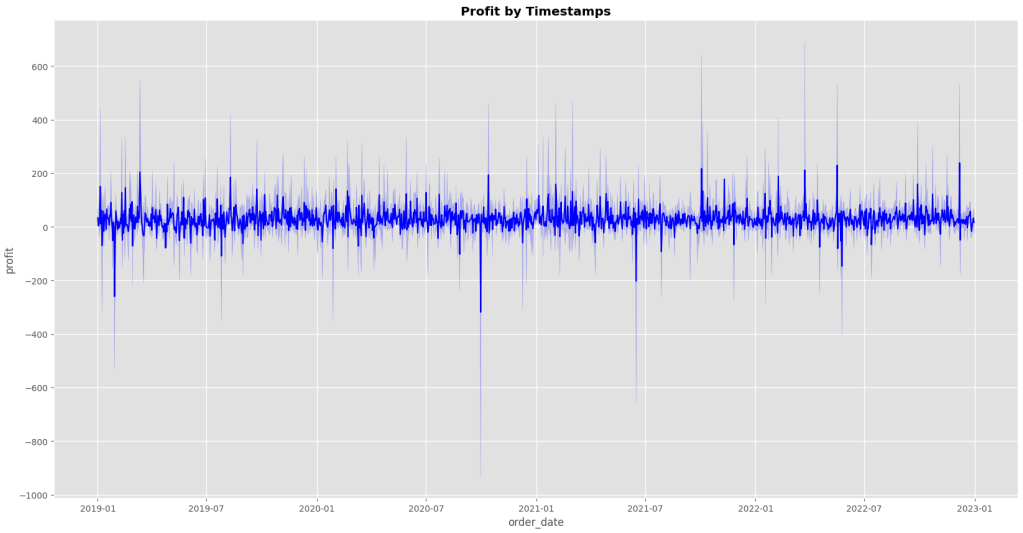
Order City



주문 도시 중 가장 많은 주문이 있었던 15개의 도시에 대한 그래프이다. 가장 많은 비율이 1.8%의 주문량으로 볼 때, 거의 대부분의 도시가 비슷한 비율로 주문한 것을 알 수 있다.

2-2. 시간적 특성 (Year,Number,Week,etc...)

Order Date



2019년 1월 1일 ~ 2022년 12월 31일 까지 시간의 순서에 따라 'Profit'의 변동을 나타낸 그래프이다. 중간 중간 급격한 수익 변화가 보이지만, 대부분 0을 기준으로 크게 벗어나지는 않는 경향성을 보인다.

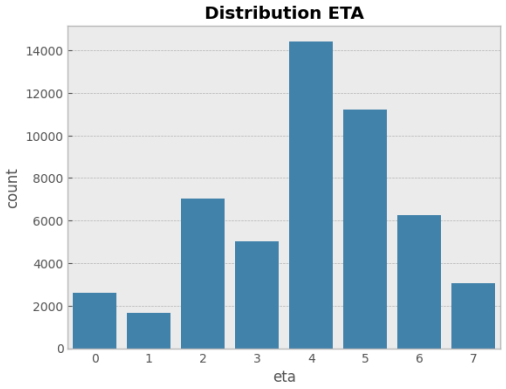
Order Year / Week Number

Order Year의 경향성으로 미뤄 보아, 2022년에 34.2%의 최대 주문 건수가 발생하였다. 또한, 연초 보다는 연중이나 연말에 주문 건수가 증가하는 추세를 보임을 알 수 있다.

ETA (Estimated Time Arrival)

Order_Date, Shipping_Date의 두 가지 Feature를 통해서 파생 Feature를 만들어 분석하였다.

```
data['eta'] = (data['ship_date'] - data['order_date']).dt.days
```



- 4일이 지나서 도착하는 상품이 가장 많았고, 아마 이는 Shipping_Class의 영향을 많이 받았으리라고 추측된다.

3. Product (ID,Categories)

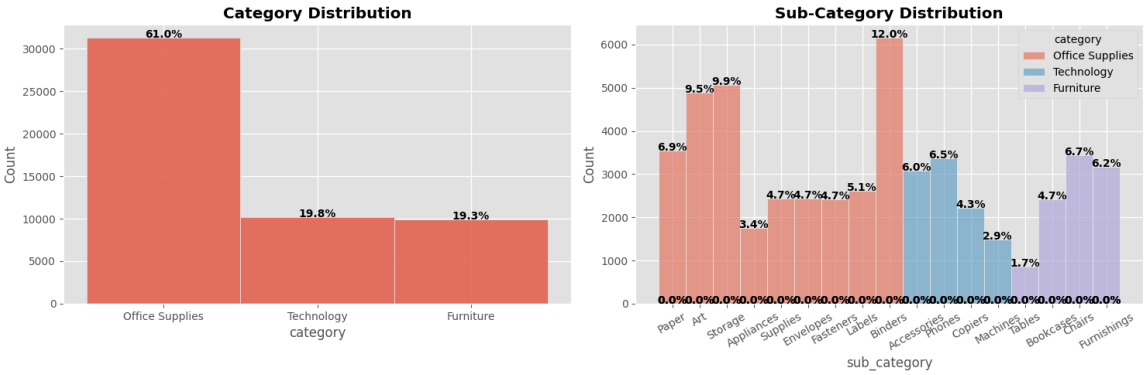
- 상품의 정보는 기업에 있어 중요한 정보이다. 특정 상품마다 다른 수익성을 지니고, "효자"노릇을 하는 상품이 있는 반면, 기업에 손해를 끼치는 상품도 있다. 기업은 손해가 있는 상품을 보완,개선하여 재정적으로 안정된 상태를 유지해야 한다.

ID, Names

```
print('Unique product ID:',data['product_id'].nunique())
print('Unique product names:',data['product_name'].nunique())
------------------------------------------------------------
Unique product ID: 10292
Unique product names: 3788
```

슈퍼마켓에서 판매하는 고유한 상품의 갯수는 총 10292개, 총 3788개의 서로 다른 이름이 있음을 알 수 있다.

Categories

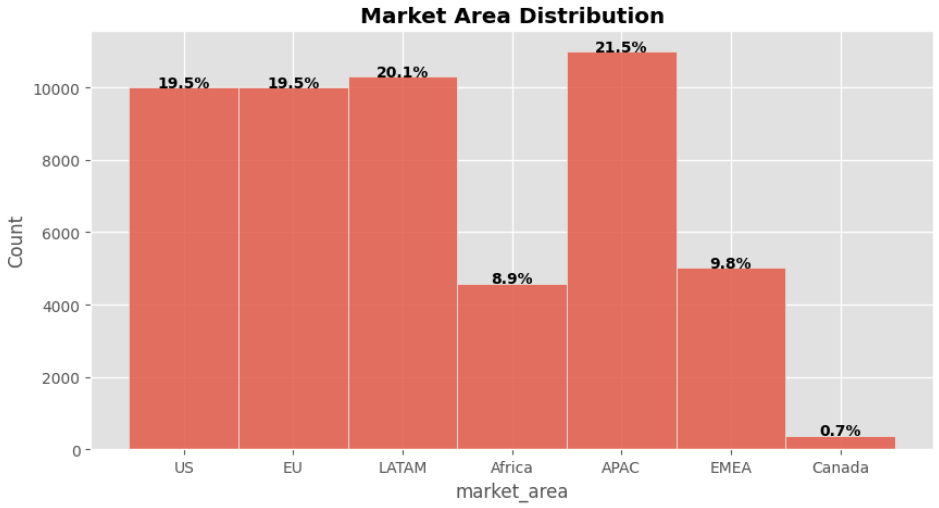


상품의 카테고리에서는 Office, 사무 용품이 대부분의 대부분의 비율을 차지하고, 그 중에서도 Binder, Art, Storage의 품목이 상당 수 많은 부분을 차지하고 있음을 알 수 있다.

4. Supermarket (Area,Country,City)

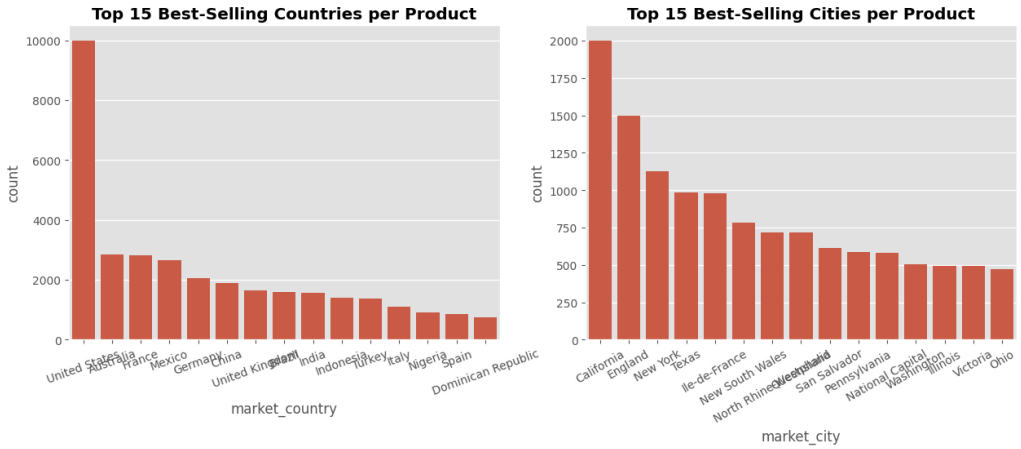
- 해당 Supermarket이 어디에 존재하는지는 기업에서 해당 국가, 도시에 알맞은 상품의 양/질을 결정하여 효과적인 판매 전략 형성에 중요한 요인이다.

Area



Market Area는 US,EU,LATAM(라틴아메리카), APAC(아시아-태평양) 모두 주요 대륙에 비슷한 분포로 존재함을 알 수 있다. US의 경우에는 캐나다를 제외한 하나의 국가이기에 주목해야할 필요성이 제기된다.

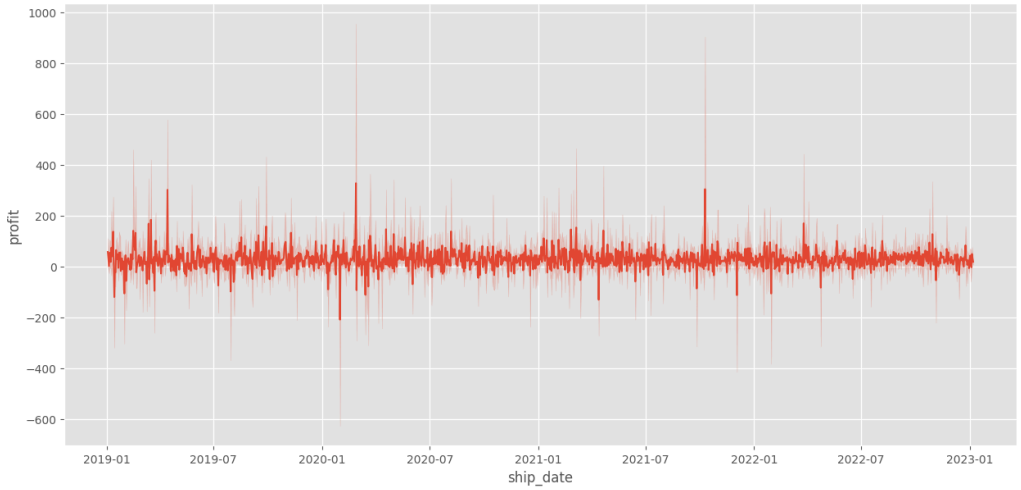
Country & City (Top 15)



- 앞에서 언급한 바와 같이, US(미국) 점포가 압도적으로 가장 많은 거래량(약 20%)을 보유하고 있음을 확인 할 수 있다. 이 중에서도 **CA(서부),NY(동부), TX(중부)**가 나란히 1,3,4등을 차지한다는 점 또한 특징이다.
- 점포 당 거래 수 2위를 차지한 **"England"** 또한 주목 해야 한다. England의 점포의 거래량은 웬만한 다른 국가 규모에 맞먹는 숫자를 갖고 있다는 점이 특징이다.
    - 번외로 United Kingdom(영국)의 도시(City)로 기록된 값은 **"England", "Scotland", "Wales"** 3가지로 추정된다.

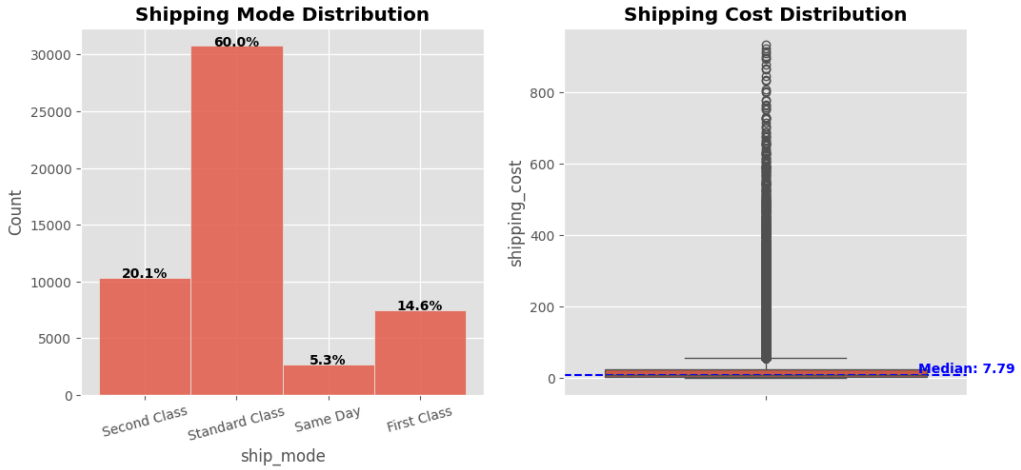
5. Shipping

Shipping Date



도착 날짜는 대체로 발송 날짜와 비슷한 경향성을 보여준다.

Shipping Mode / Cost

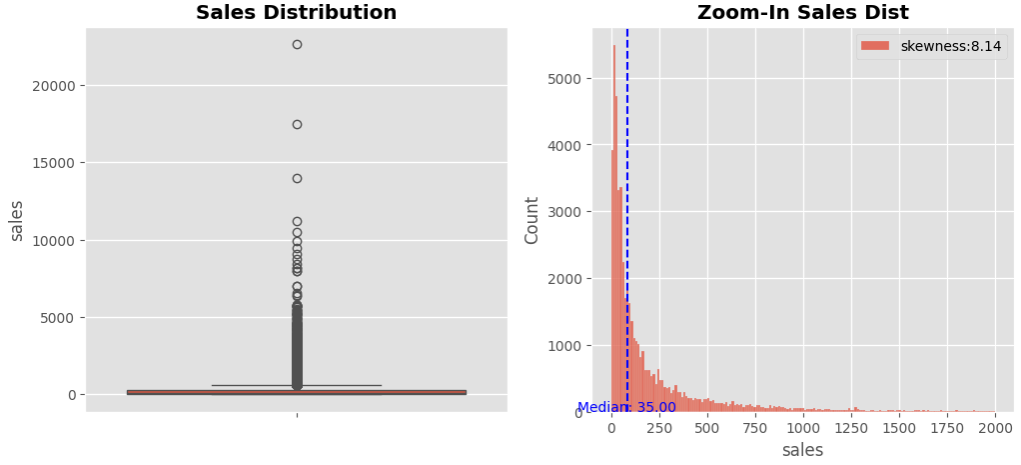


- 운송하는 등급에 있어서는 “Standard Class”가 압도적으로 많은 비율을 차지했고, 비용적인 측면에서는 대체로 0~50 사이의 값을 지니고 있음을 알 수 있다. Boxplot을 보면 굉장히 높은 Shipping Cost를 지닌 물건들도 확인할 수 있다.

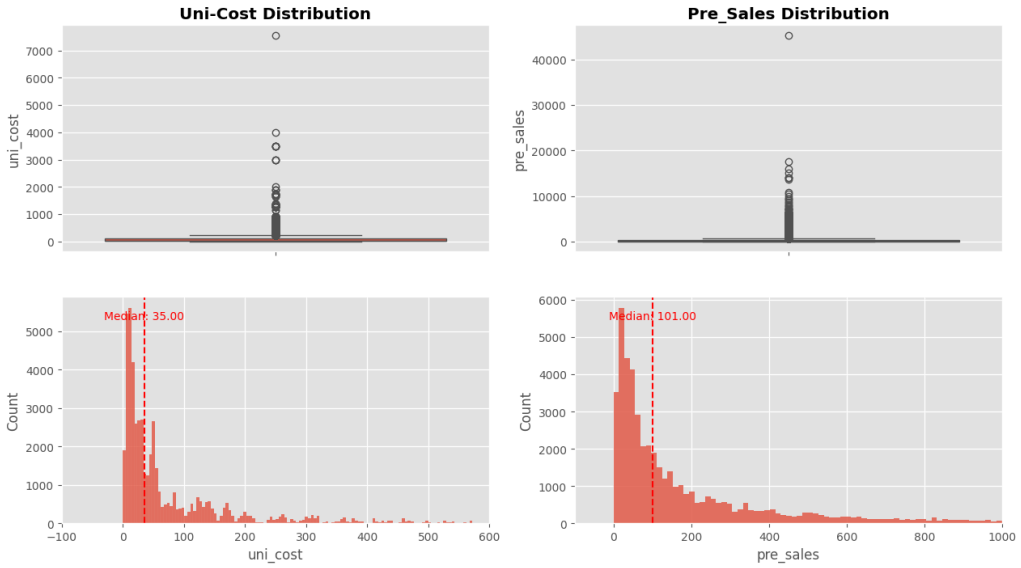
6. Prices (Sales,Profit,etc..)

- 기업에 있어 재무적 사항은 매우 중요하다. 기업은 **이익 창출**을 목표로 하며, 이익의 여부에 따라 기업의 흥망성쇠가 결정된다.

Sales (inc. Pre / Unit Sales)

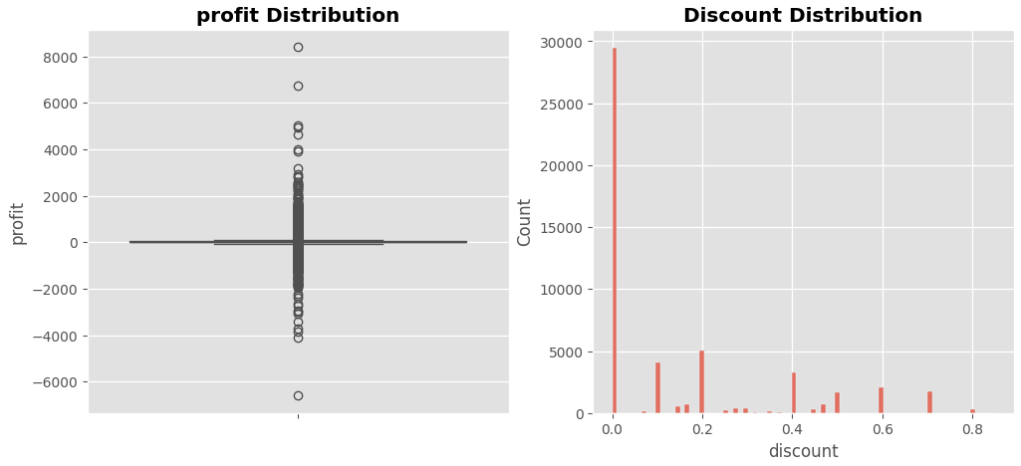


기본적으로 주어진 Sales의 경우 매우 넓게 분포하고 있다. 상당히 가격이 높게 팔린 상품들도 보이며, 대부분 0~250의 범위 내에서 판매된 상품이 많다.

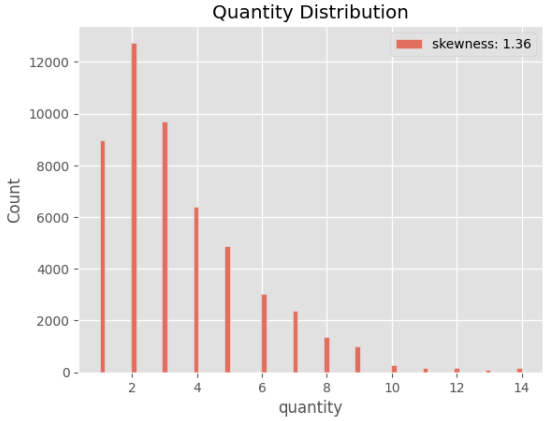


Pre_Sales와 Unit-Cost의 경우에도 Sales와 비슷한 경향성을 가지고 있음을 확인할 수 있다.

Profit, Discount, Quantity



profit의 경우, 0을 중심으로 많은 값들이 분포해 있다. 또한, Discount의 경우에도 0이 상당 부분을 차지하고 있다. 뒤에서 나오겠지만, 수익(profit>0)을 낸 데이터의 수는 약 80%정도이고, 손실이 난 데이터는 약 20%정도를 차지하고 있다.



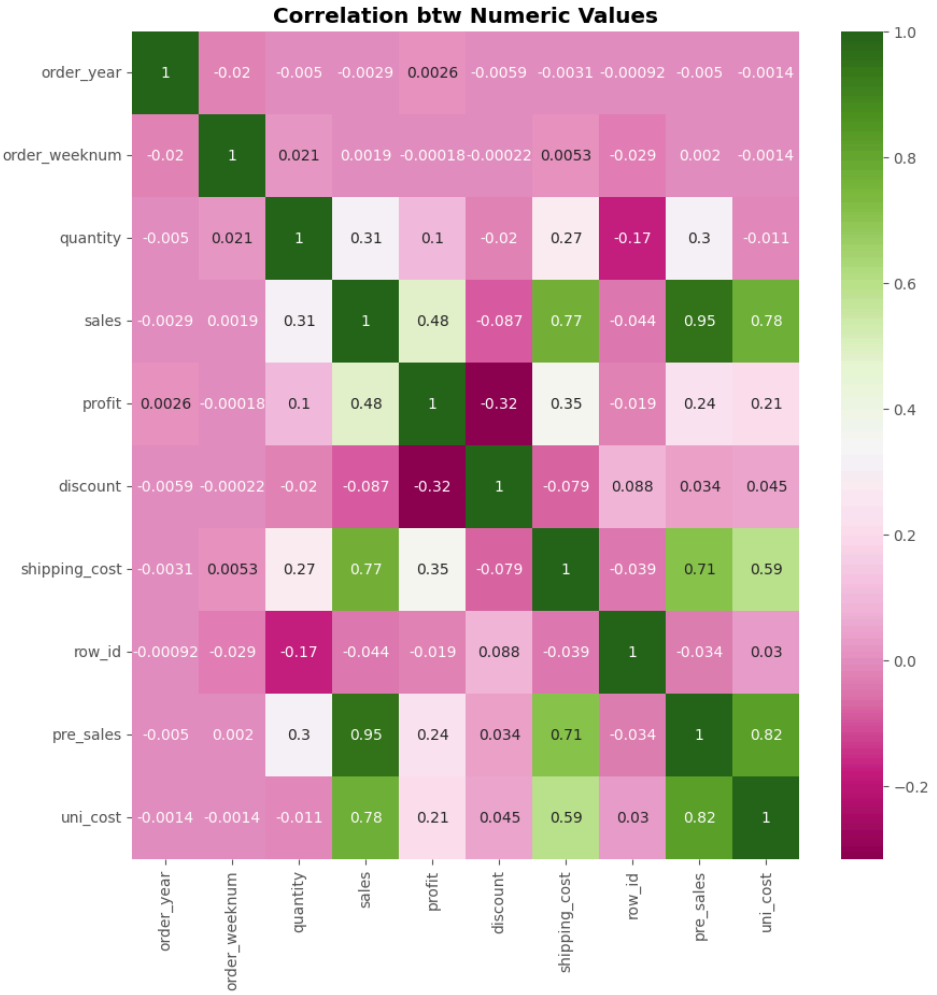
- Quantity의 경우에는 2개로 파는 상품이 가장 많았고, 다른 연속형 변수형 값들에 비해서 정규화된 분포를 보여준다.

7. Correlation Anaylsis

Data의 Feature들을 모두 다음과 같이 분류하였다.

Column Types	Column Names
개별 식별자 (ind_col)	customer_id, customer_name, order_id
범주형 데이터 (cat_col)	customer_id, customer_name, customer_segment, order_id, order_city, order_region, order_date, order_year, order_weeknum, product_id, product_name, market_area, market_city, ship_date, ship_mode, category, sub_category, market_country
연속형 데이터 (con_col)	quantity, sales, shipping_cost, profit, discount, row_id, uni_cost, pre_sales

변수 간 상관 관계를 파악하기 위해 heatmap으로 Correlation Coefficient를 시각화 하였다.



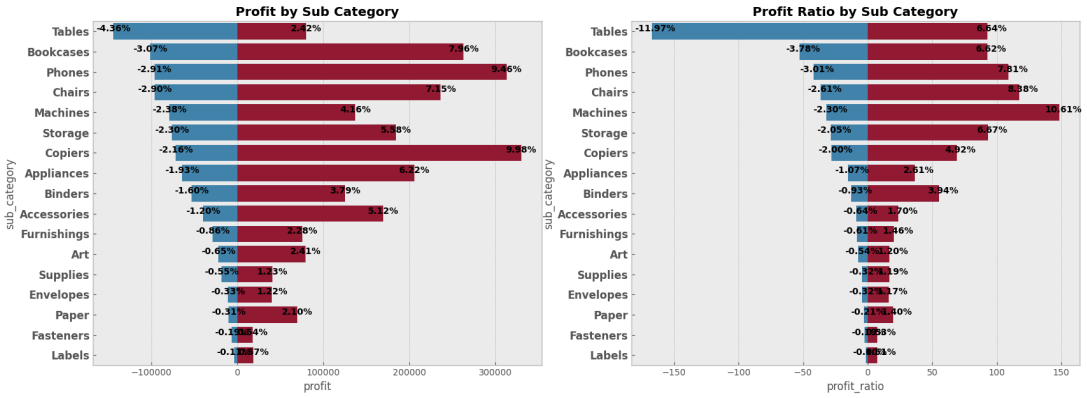
- **Discount - Profit (0.32)** : 할인을 많이 하면 할 수록 수익성이 떨어지는 음(−)의 상관 관계를 띄고 있다. 추가적인 분석이 필요하지만, 이러한 관계를 파악하고 개선하는 것은 당장의 수익성을 높이는데 효과적일 수 있다.
  - 단적인 예시로, Discount의 요인을 제거한 Pre_Sales의 경우에는 Discout와는 반대로 양(+)의 상관 관계를 띄고 있음을 알 수 있어 **수익성을 위해 할인률을 줄이는 방법을 고려해볼 수 있다**.
- **Sales** : Sales(매출)와 연속형 변수들 간의 상관 관계가 눈에 띈다. 특히 운송 비용(**0.77**)이 많이 드는 상품의 경우 많은 매출이 발생했고, 또한 많은 양(Quantity,**0.31**)를 팔면 팔 수록 매출에 긍정적인 상관 관계가 있었다.

Loss(손실) Analysis

아래에서는 기업의 영업이익 개선을 위해 손실을 분석하고, 이를 최소화 하고자 하는 방법을 찾아내기 위한 탐색을 진행한다. 주로 Profit Loss와 관련되어 이변량(BiVariate), 다변량(MultiVariate) 분석을 진행한다.

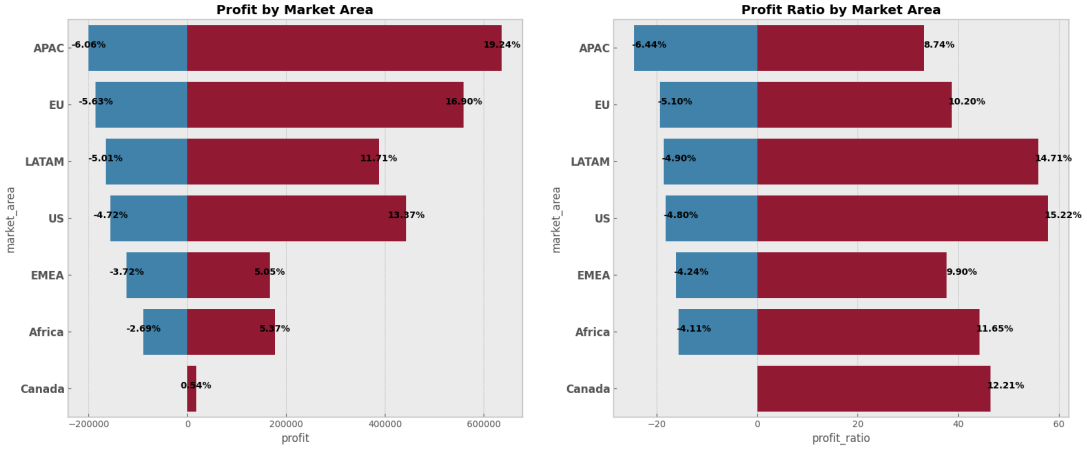
이를 위해 우선 Profit이 음수인 데이터를 추출하여 '**Minus**'라는 `DataFrame` 으로 저장하였다.

```
minus = data.loc[data['profit']<0]
print(minus.shape)
---------------
(12544,26)
```

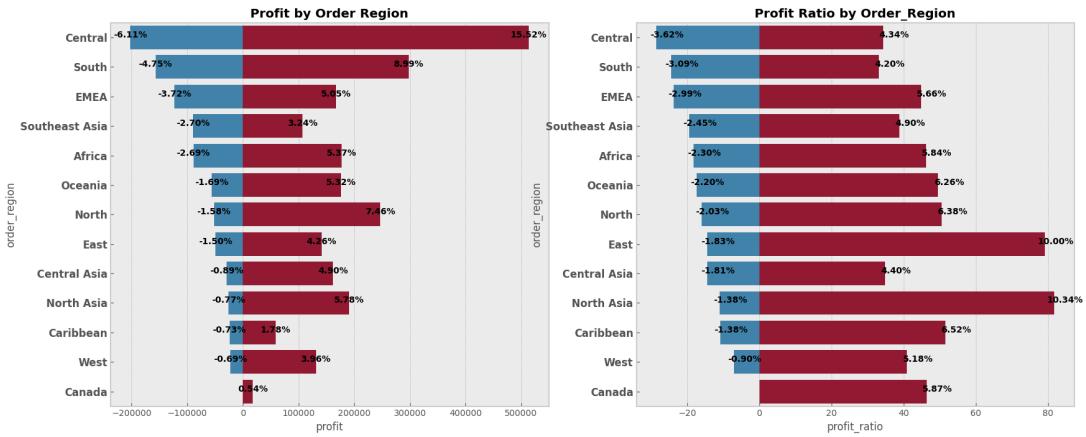


위의 두 그래프는 이익 / 손실이 난 비용과 거래량당 비용 비율을 Sub_Category에 따라 시각화한 그래프인다. 여기서 “Tables”의 경우, 비용,비율적인 측면에서 가장 손해가 큰 부분임을 확인 할 수 있다.

그에 반해 “Copiers”의 경우에는 비용,비율적인 측면에서 가장 수익성이 뛰어난 상품 sub_category임을 확인 할 수 있다.



Market Area에 따른 Profit을 보면 모든 지역(대륙)에서 (+) 이익을 내고 있음을 알 수 있다. 손실 비용, 비율측면에서 많은 지역은 APAC임을 알 수 있지만, 그만큼 수익성도 많이 뛰어남을 알 수 있다.

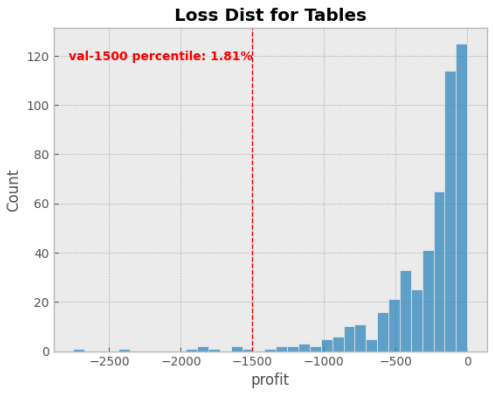


Region에 따라서는 Central 지방의 손실도 가장 높지만, 수익도 가장 많이 나서 가장 수익성이 좋은 지역 중 하나이다. 그 이외에 동북아시아 지방 또한 거래 당 수익이 가장 많이 발생하는 지역임을 확인 할 수 있다.

1. Product “Table”

```
mt = minus.loc[minus['sub_category']=='Tables']
mt['sub_category'].min()
-------------------------------
-2750.28
```

“mt”라는 데이터 프레임을 만들어 손실이 난 부분을 분석하고자 한다.

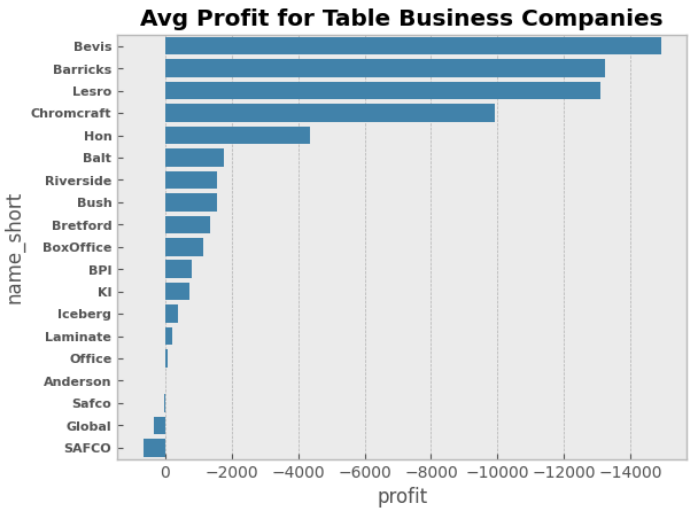


- 옆의 그래프는 Tables data 중 Loss의 Distribution을 나타내었다. 대부분 -1000~0 사이의 손실이 많다.
- 손실이 작은 값들을 보완하는 것이 중요하지만, 비용적으로 가장 큰 타격을 입힌 거래, 실수들을 방지하는 것이 우선이다.
- 아래에는 mt 데이터에서 심각한 피해를 입힌 데이터의 정보를 요약한 DataFrame이다. (uni_cost 반올림)

index	order_region	order_city	market_area	product_id	discount	uni_cost	profit
29652	EMEA	Vilnius	EMEA	FUR-BAR-10003532	0.7	904.5	-2750.28
30191	Central Asia	Lahore	APAC	FUR-TA-10002172	0.8	919.28	-2380.35
29974	Central	Hanover	EU	FUR-TA-10003963	0.85	925	-1924.54
29704	North	Stockholm	EU	FUR-TA-10003354	0.7	909	-1864.09
47284	South	Concord	US	FUR-TA-10000198	0.4	551.02	-1862.31
29390	EMEA	Ankara	EMEA	FUR-BEV-10002193	0.6	520.42	-1779.76

index	order_region	order_city	market_area	product_id	discount	uni_cost	profit
38582	South	Barcelona	EU	FUR-TA-10004054	0.6	857.5	-1629.54
40773	Africa	Zaria	Africa	FUR-CHR-10002278	0.7	469.2	-1576.82
29693	North	Stockholm	EU	FUR-TA-10004371	0.7	454.28	-1557.99

- 'name_short' 변수는 'product_name'에서 공백 이전의 문자를 추출한 Feature이다. mt의 데이터 중 product_name앞에 회사 이름을 대표하는 문자가 있는 것을 발견하였다.
- 대체로 Discount 비율이 높고, EU 쪽 점포에 많은 손실을 끼쳤을 것으로 파악된다.



- 위 그래프는 Table 생산 기업들의 평균적인 손/이익을 나타내었다.
- 상위 5개의 평균 영업 손실을 일으킨 기업들의 이름을 위의 표에서도 볼 수 있다.
- 따라서 큰 금액의 손실을 남긴 거래들을 방지하는 것이 최우선적인 해결책으로 보인다.

특히, **Bevis, Barricks,Lesro,Chromcraft 사**의 경우에는 손실이 너무 많기 때문에 아예 Table 사업을 정리/ 축소 하는 것도 효과적으로 손실을 줄이는 방법이라 할 수 있겠다.

- **Bevis**

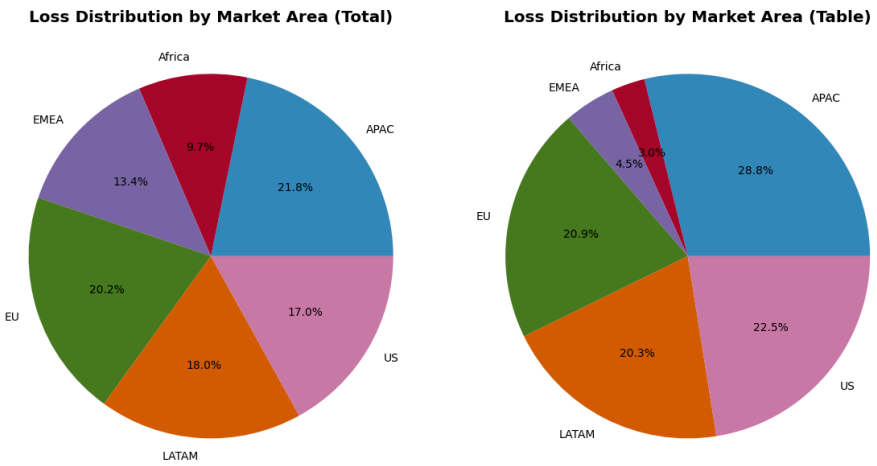
Bevis Training Table, Fully Assembled',
'Bevis Wood Table, Rectangular',
'Bevis Training Table, Rectangular',
'Bevis Computer Table, Fully Assembled',
'Bevis Computer Table, Adjustable Height',etc

- **Barricks Furniture Solutions**

'Barricks Computer Table, Fully Assembled',
'Barricks Training Table, with Bottom Storage',
'Barricks Conference Table, Rectangular',etc

- **Lesro Reception Furniture**

'Lesro Computer Table, Adjustable Height',
'Lesro Training Table, Fully Assembled',
'Lesro Wood Table, with Bottom Storage',
'Lesro Coffee Table, Adjustable Height',etc



*(좌) Market Area에 따른 전체 손실액 비율 (우) Market Area에 따른 Table 상품의 손실액 비율

APAC, US, LATAM 지역에서는 전체 손실액에서 차지하는 비율보다 **Table 상품 판매 시 차지하는 손실 비율이 더 높음**을 확인할 수 있다. 이는 해당 지역에서 Table 상품이 상대적으로 더 큰 손실을 발생시키고 있음을 의미한다. 따라서 위 지역에서 발생하는 문제를 추가적으로 분석해볼 필요성이 제기된다.

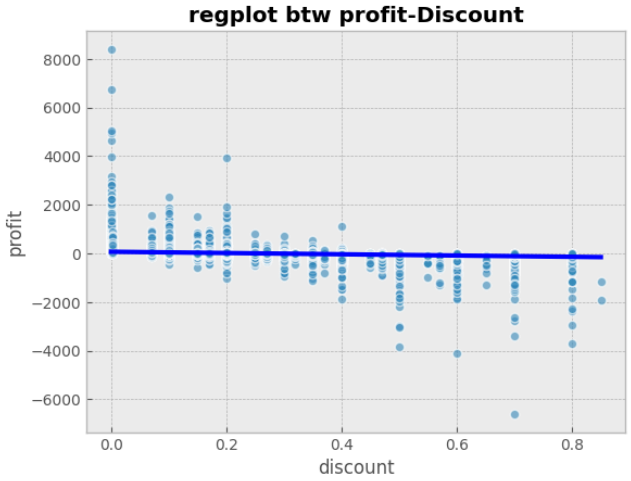
참고/고려사항

Table로 거래되어 기록된 Data가 매우 부족하다(861여개). 현재 데이터를 가지고는 개선이 시급하지만, 추후에 데이터를 더욱 추가하여 기업의 의사 결정에 이용하는 것이 권장된다.

---

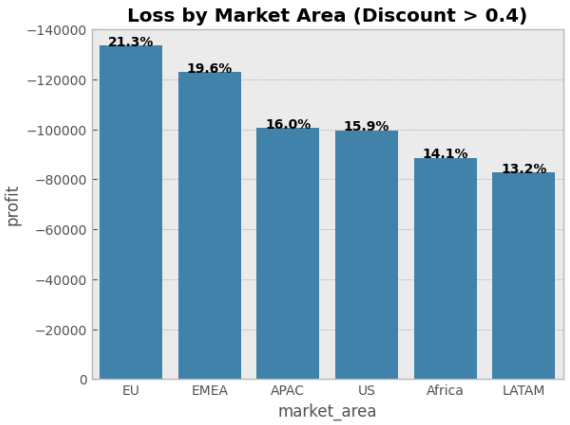
2. Discount

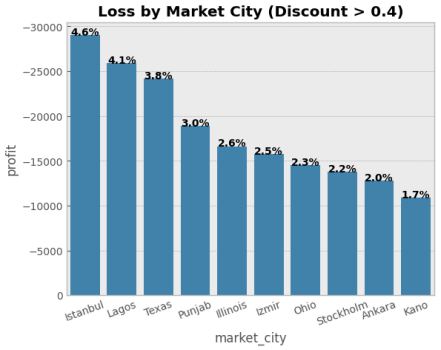
- 할인은 사람들이 물건을 구매하게 되는 중요한 요인 중 하나로 작용한다. 하지만 너무 무리한 할인으로 인해 기업에게는 손실로 작용할 가능성이 존재한다. 할인은 구매하는 **지역/물품**에 따라서도 달라지고, **고객의 등급**에 따라서도 달라질 수 있다.
- Unique Value : 27개 , 가장 많이 나타난 값 : 0 (29009개,약 56%)



- 옆의 그래프는 전체 Data에서 profit와 Discount를 regplot으로 시각화 한 그림이다.
- 할인률 0.4를 기준으로 거의 이익이 발생하는 주문 건수는 존재하지 않는다.
- 반면에 0.4를 넘는 주문 건수 중 다수가 회사에 손해를 끼쳤음을 알 수 있다.
- 또한 손해액의 규모 또한 커지고 있음을 알 수 있다.

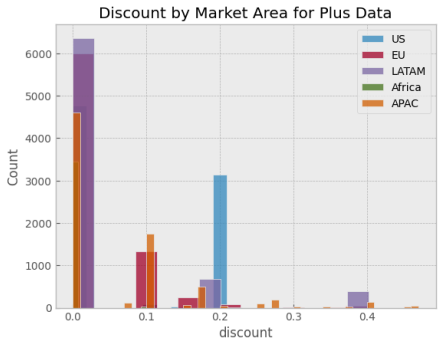
0.4보다 높은 할인률을 가지는 데이터의 비율은 약 13% 정도이다. 또한, 이 데이터는 전체 이익의 0.01%를 차지하지만, **전체 손해의 68% 정도**를 차지한다.

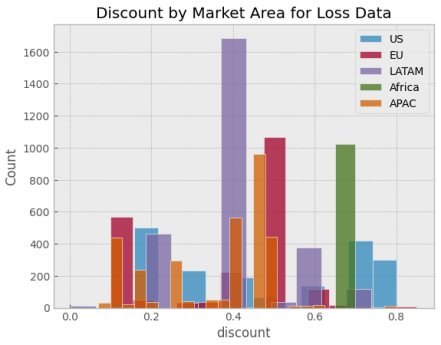




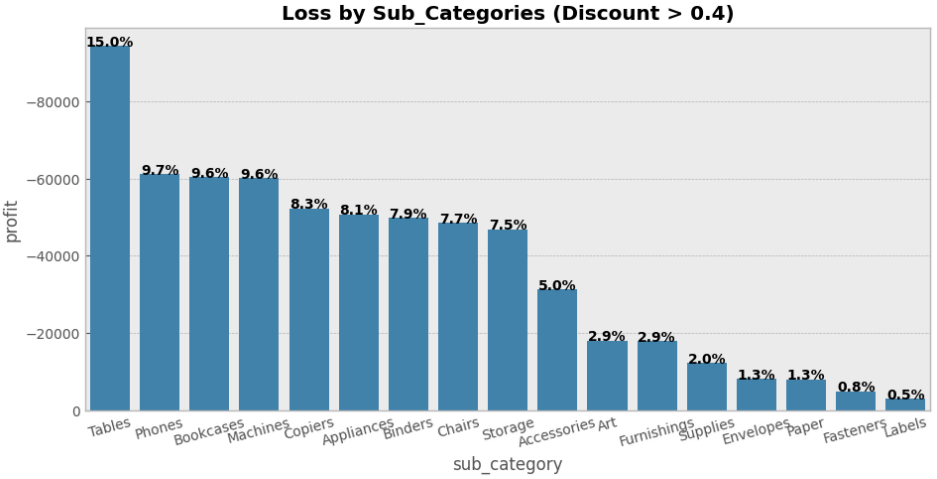
- Market Area에 따라서는 대부분 손해액의 비율이 엇비슷하다. EU에 적용된 할인률을 낮추는 쪽으로 진행하면 수익성이 더욱 증가할 것으로 예상된다.
- Market City에서 발견된 상위 3개 도시를 조사하니, 다음과 같은 Discount Value를 가지고 있었다. 이를 토대로 Istanbul, Lagos의 할인률을 0.4보다 낮게 측정하고, Texas의 높은 할인률 상품을 소폭 줄이는 방향으로 개선이 필요하다.

City	Discount	Values
Istanbul	0.6	425(total)
Lagos	0.7	333(total)
Texas, US	{0.2,0.8,0.3,0.6,etc}	{570,200,94,81,...}

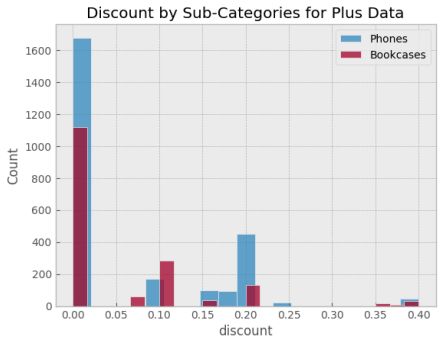


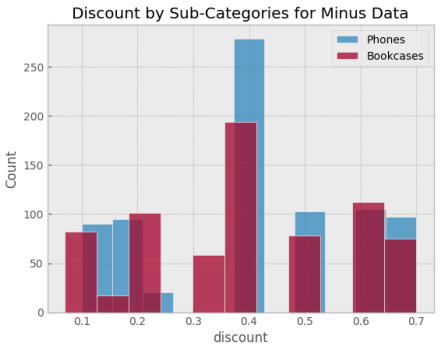


수익 / 손해가 난 Market Area의 Discount 값을 각각 plot한 그래프이다. 빨간색 EU의 그래프를 보자면, 손해가 대부분 0.4~0.6의 할인률 상품에서 발생하였다. 이러한 상품의 할인률을 **10%p~30%p** 정도 줄이면 손해가 감소할 것으로 예상된다.



- Tables은 앞서서 제품 개선을 진행 한 바, Phones, Bookcases에 대해서 자세히 살펴보자. 앞선 Sub_Category의 Distribution EDA를 통해서 수익성도 좋고, 거래가 많이 발생하는 상품이라는 것을 확인 할 수 있다. (각각 전체 거래량 중 6.5,6.7% 를 차지함)





- 앞선 그래프와 동일한 방식으로, Discount의 분포를 손해 / 수익, Sub_Category에 따라 분류하여 나타낸 그래프이다.
- 두 상품 모두 40% 이상의 할인률을 넘지는 않도록 조정해야 한다. 또한, 20% 할인 구간에 Phones의 경우 꽤 많은 양의 수익이 났기 때문에 조건에 맞는 제품들을 20%로 할인률을 낮춰 판매하는 전략이 권장된다.

Q2. Final Business Decision-making Report

💡 앞선 EDA와 자세한 분석을 토대로, 높은 Discount Value를 가지는 상품들의 거래가 기업의 손해로 이어지는 것을 확인 할 수 있었다. 특히 40%이상의 할인률을 보이는 상품에 대해서는 할인률 감소가 필요하다. 구체적으로는 EU, 유럽에서 거래되는 상품의 할인률을 평균적으로 20%p 줄이고, Istanbul, Lagos 도시의 경우에는 상품에 따른 할인률의 다양화가 필요하다. 또한, Table으로 분류되어 판매되고 있는 상품 중 "Bevis","Barricks Funiture","Lesro" 사에서 거래되는 상품의 대규모 사업 조정이 필요할 것으로 예상된다. 또한 APAC,US,LATMA 지역에서 다른 지역와 비교하여 많은 손해가 발생하고 있다. 이러한 현상에 대한 후속 조치, 개선이 필요하다.