전자산업데이터분석
Final Project
- 주제 A : 재무분석 팀

TEAM 3
2131040 이윤경
2231163 진현지
2031019 황투히엔

Project 주제 A _ 재무분석 팀

“

당신은 ElectroWorld의 재무분석 팀의 팀원입니다.
ElectroWorld의 경영진은 전반적인 회사의 재무 및 수익성에 대해 분석을 하여,
더 좋은 성과를 내고자 합니다. 이번 프로젝트의 주요 핵심 포인트는
머신러닝 모델을 이용하여 이러한 수익성을 예측하고, 개선하는 것입니다.

”

데이터 전처리

1-1 데이터 불러오기

01

```
1 EW_df = pd.merge(ElectroWorld_df, city_df, how='inner', on='city')
```

```
1 EW_df.shape
```

```
(7149, 23)
```

```
1 np.random.seed(5)
2 idx = np.random.permutation(np.arange(len(EW_df['record_id'])))
```

```
1 EW_df = EW_df.iloc[idx].drop_duplicates(["record_id"])
```

```
1 EW_df.shape
```

```
(5500, 23)
```

ElectroWorld 데이터와 City 데이터의 경우 merge 후, *record_id*가 중복으로 추가 삽입된 행은 랜덤으로 한 개씩만 남기고 drop

데이터 전처리

1-1 데이터 결측치 처리

01

'ship_mode', 'customer_segment', 'product_container'의 경우 엑셀을 통해 우선적으로 결측치 처리

- '***ship_mode***' : 'product_name', 'production_container', 'order_quantity', 'city'를 기준으로 일치하는 값으로 결측치 값 대체
- '***customer_segment***' : 'customer_name와 'city'를 기준으로 일치하는 값으로 우선적으로 결측치를 채웠으며,

  'order_quantity'와 'product_sub_category' 'product_name'를 참고하여 모든 결측치 값 대체
- '***product_container***': 'product_category', 'production_sub_category', 'product_name'을 기준으로 일치하는 값으로 결측치 값 대체

데이터 전처리

1-1 데이터 결측치 처리

01

‘order_priority’ 결측치 처리

- random 함수를 사용하여 결측치를 임의로 배정

‘discount’ 결측치 처리

- ‘*product_sub_category*’ : ‘product_sub_category’별 평균값으로 결측치 처리
- ‘*discount_mean*’ : discount 전체 값의 평균으로 결측치 처리
- ‘*discount_zero*’ : '0'으로 결측치 처리
- 기존 `discount` 컬럼 삭제

`product_base_margin` 결측치 처리

- `*product_base_margin_sub_mean*` : product_sub_category별 평균값으로 결측치 처리
- `*product_base_margin_mean*` : product_base_margin_mean 전체 값의 평균으로 결측치 처리
- 기존 `product_base_margin` 컬럼 삭제

```
EW_df["order_priority"].isna()
```

```
2203     True
1268    False
3634    False
5381    False
857     False
        ...
1982    False
2121    False
5520    False
3046    False
2915    False
Name: order_priority, Length: 5500, dtype: bool
```

```
np.random.seed(3)
m = EW_df["order_priority"].isna()
EW_df.loc[m, 'order_priority'] = np.random.choice(EW_df.loc[~m, 'order_priority'], m.sum())
EW_df["order_priority"].value_counts()
```

```
low         1517
high        1408
critical    1329
medium      1246
Name: order_priority, dtype: int64
```

▲ ‘order_priority’ 결측치 처리 코드

데이터 전처리

1-2 파생변수 생성

‘new_sales’ 파생변수 생성

- 기존 ‘sales’ 값의 산정방식(개당가격*수량, 개당가격*수량*할인율 등으로 sales 값이 나오지 않음)이 계산에 따라 나오지 않아, (‘unit_price’ * ‘order_quantity’) * (1 - ‘discount_mean’) 산식으로 추가 변수 생성

‘order_date’를 통해 연월, 년, 월, 요일, 분기의 파생변수 생성

‘ship_date’를 통해 연월, 년, 월, 요일, 분기의 추가 변수 생성

‘lead_time’ 파생변수 생성

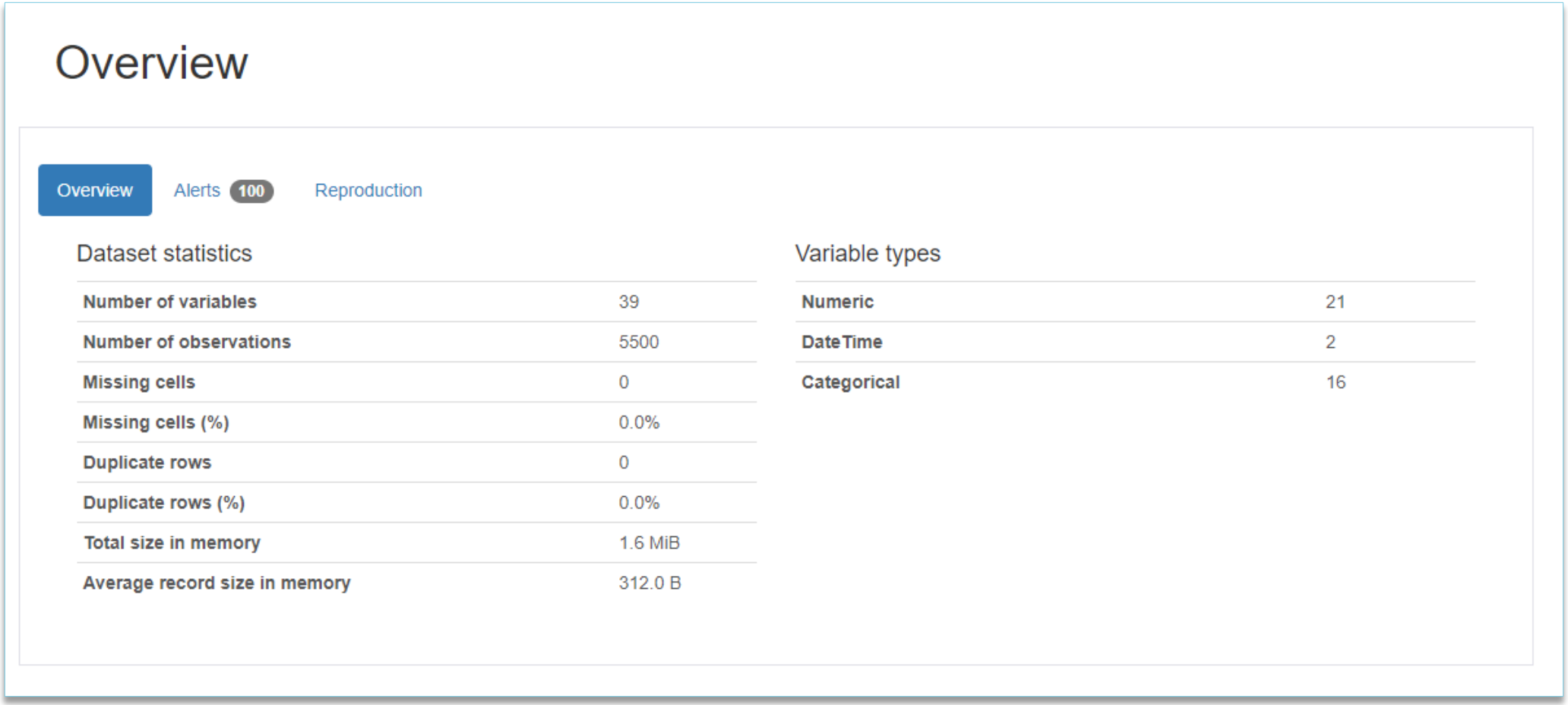
- `ship_date` - `oder_date` 하여 고객이 주문한 날짜로부터 배송되기까지 걸린 일자를 계산하여 새로운 컬럼(`lead_time`)에 할당

```
# 주문에서 부터 배송되기까지 걸린 일자 계산
EW_df['lead_time'] = EW_df['ship_date'] - EW_df['order_date']
EW_df['lead_time'] = EW_df['lead_time'].dt.days
```

```
EW_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5500 entries, 2203 to 2915
Data columns (total 38 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   record_id                       5500 non-null   int64
 1   order_id                        5500 non-null   int64
 2   order_date                      5500 non-null   datetime64[ns]
 3   order_priority                  5500 non-null   object
 4   unit_price                      5500 non-null   float64
 5   order_quantity                  5500 non-null   int64
 6   sales                           5500 non-null   float64
 7   order_status                    5500 non-null   object
 8   profit                          5500 non-null   float64
 9   ship_mode                       5500 non-null   object
 10  ship_cost                       5500 non-null   float64
 11  ship_date                       5500 non-null   datetime64[ns]
 12  customer_segment                5500 non-null   object
 13  customer_name                   5500 non-null   object
 14  city                            5500 non-null   object
 15  product_category                5500 non-null   object
 16  product_sub_category            5500 non-null   object
 17  product_name                    5500 non-null   object
 18  product_container               5500 non-null   object
 19  state_name                      5500 non-null   object
 20  region                          5500 non-null   object
 21  discount_sub_mean               5500 non-null   float64
 22  discount_mean                   5500 non-null   float64
 23  discount_zero                   5500 non-null   float64
 24  product_base_margin_sub_mean    5500 non-null   float64
 25  product_base_margin_mean        5500 non-null   float64
 26  new_sales                       5500 non-null   float64
 27  order_yearmonth                 5500 non-null   int64
 28  order_year                      5500 non-null   int64
 29  order_month                     5500 non-null   int64
 30  order_day                       5500 non-null   int64
 31  order_quarter                   5500 non-null   int64
 32  ship_yearmonth                  5500 non-null   int64
 33  ship_year                       5500 non-null   int64
 34  ship_month                      5500 non-null   int64
 35  ship_day                        5500 non-null   int64
 36  ship_quarter                    5500 non-null   int64
 37  lead_time                       5500 non-null   int64
dtypes: datetime64[ns](2), float64(10), int64(14), object(12)
memory usage: 1.6+ MB
```

▲ ‘lead_time’ 파생변수 생성



전체 39개 변수 및 5,500개 행으로 이루어진 데이터

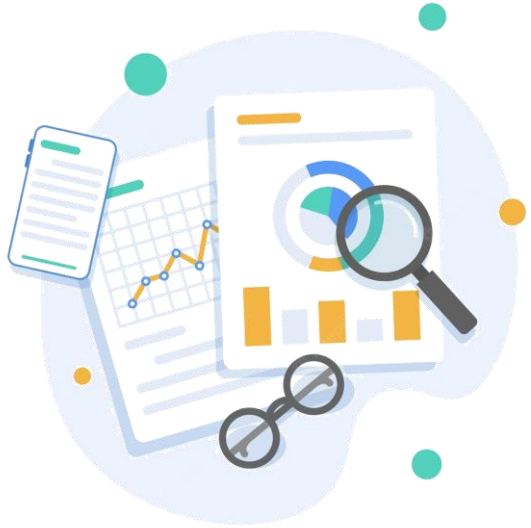
데이터 형식 : Numeric type 21개 / Categorical type 16개 / DateTime 2개

Project Prob A1

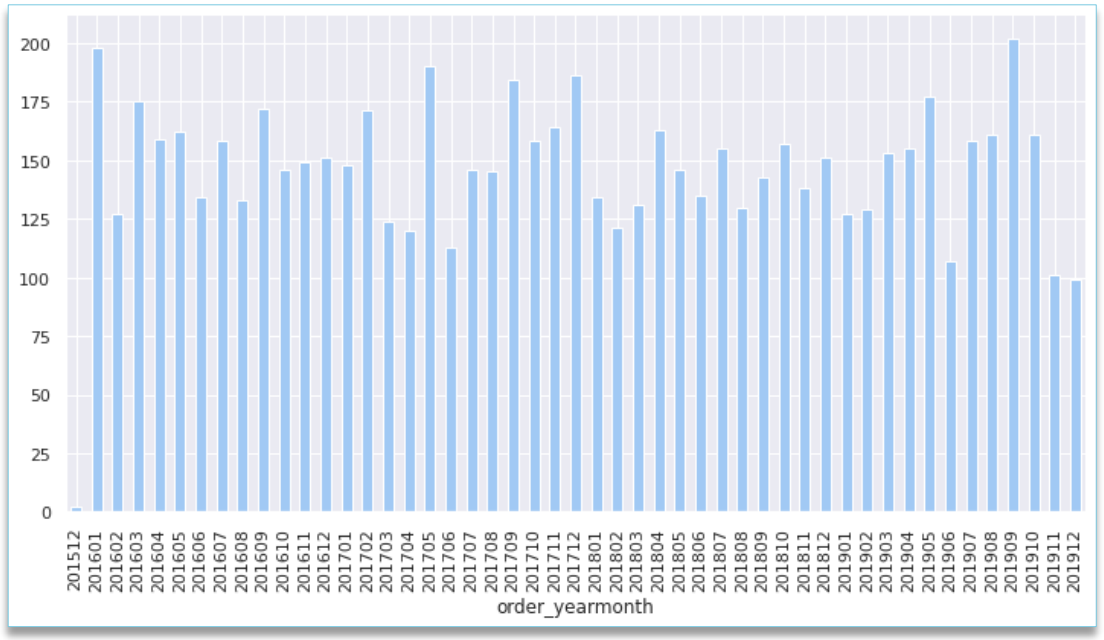
“

”

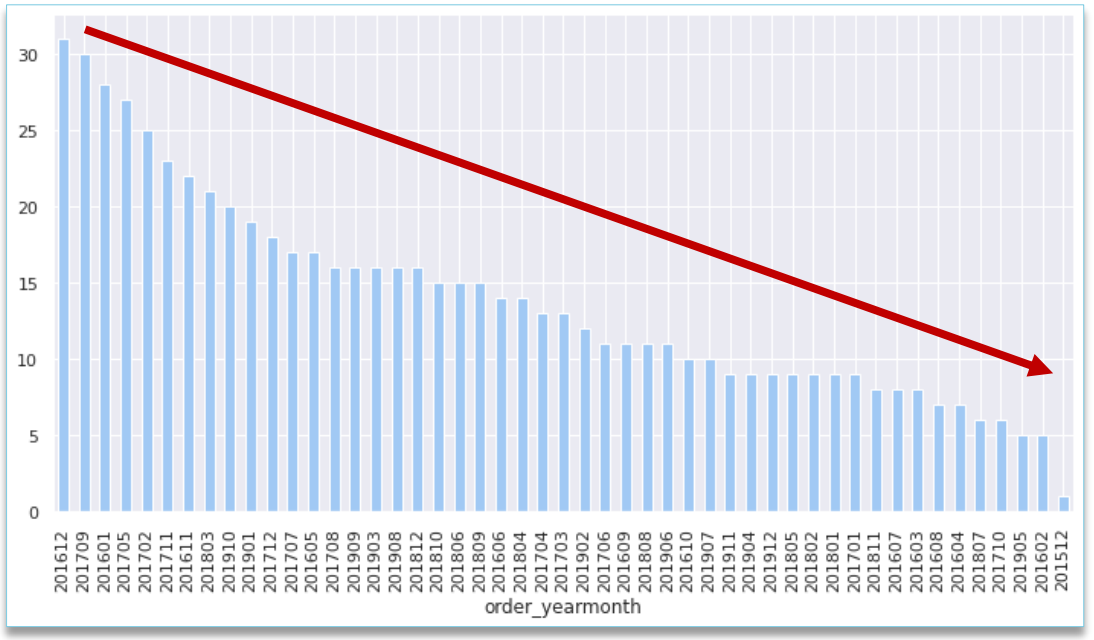
ElectroWorld 의 전반적인 비즈니스 경영 현황 점검을 수행하고자 합니다.
주어진 데이터를 이용하여 매출 및 수익성 등을 분석해보고, 개선점 or 제안점을 찾아보세요.



연월별 총 주문 건수



연월별 총 반품 건수



총 주문건수에 연월별 차이가 있지만, 반품 건수는 꾸준히 감소하는 추세를 보이고 있음

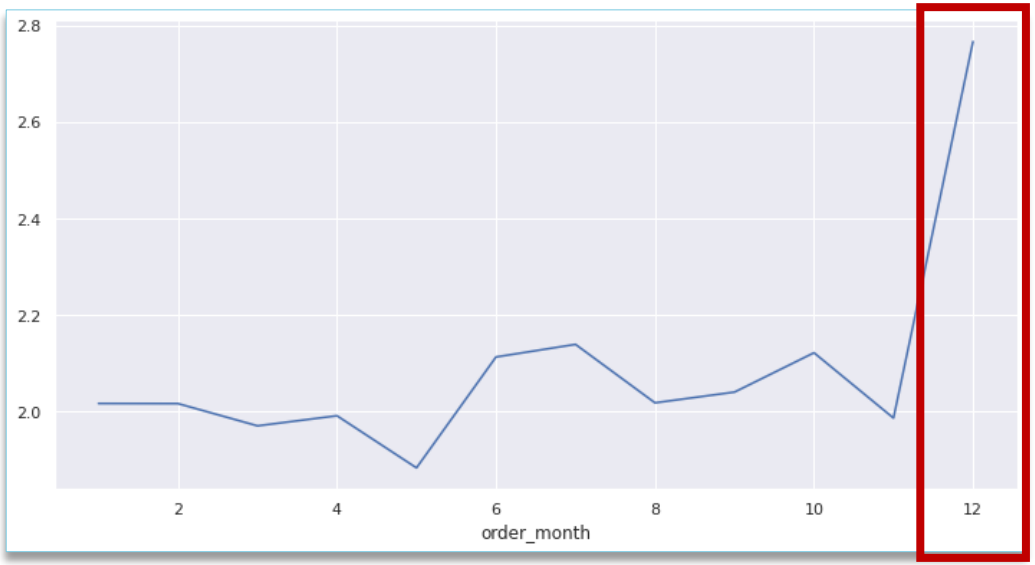
Ⅲ Project Prob A1

– 리드타임 개선을 통한 수익성 증대

03

월별 반품건수

월별 리드타임 평균



product_sub_cateory별 리드타임 max 값

```
1 EW_df.groupby(['product_sub_category', 'product_name', 'order_month']).lead_time.max().sort_values(ascending=False).iloc[:20]
```

product_sub_category	product_name	order_month	
Pens & Art Supplies	Quartet Alpha White Chalk, 12/Pack	12	92
Envelopes	#10 Self-Seal White Envelopes	12	84
Telephones and Communication	688	12	31
	282	12	28
Pens & Art Supplies	Staples SlimLine Pencil Sharpener	12	27
Telephones and Communication	Bell Sonecor JB700 Caller ID	12	24
	M3682	12	19
	6000	12	17
Chairs & Chairmats	Office Star - Contemporary Task Swivel chair with 2-way adjustable arms, Plum	12	17
Paper	Xerox 220	12	15
Computer Peripherals	Logitech Cordless Elite Duo	12	11
Paper	Xerox 1908	12	11
Labels	Avery 51	5	9
Appliances	Conquest™ 14 Commercial Heavy-Duty Upright Vacuum, Collection System, Accessory Kit	10	9
Labels	Avery 514	7	9
Storage & Organization	Letter Size Cart	4	9
Labels	Avery 51	4	9
	Self-Adhesive Address Labels for Typewriters by Universal	2	9
Envelopes	#10 Self-Seal White Envelopes	5	9
Paper	Xerox 1930	5	9

Name: lead_time, dtype: int64

12월에서 반품건수가 가장 많으며, 리드타임이 약 2.8일로 가장 길게 나타남

또한, Telephones and communication 제품 서브 카테고리에서 리드타임이 높게 나타남

Telephones and Communication 제품의 리드타임이 반품에 영향을 끼칠 수 있어

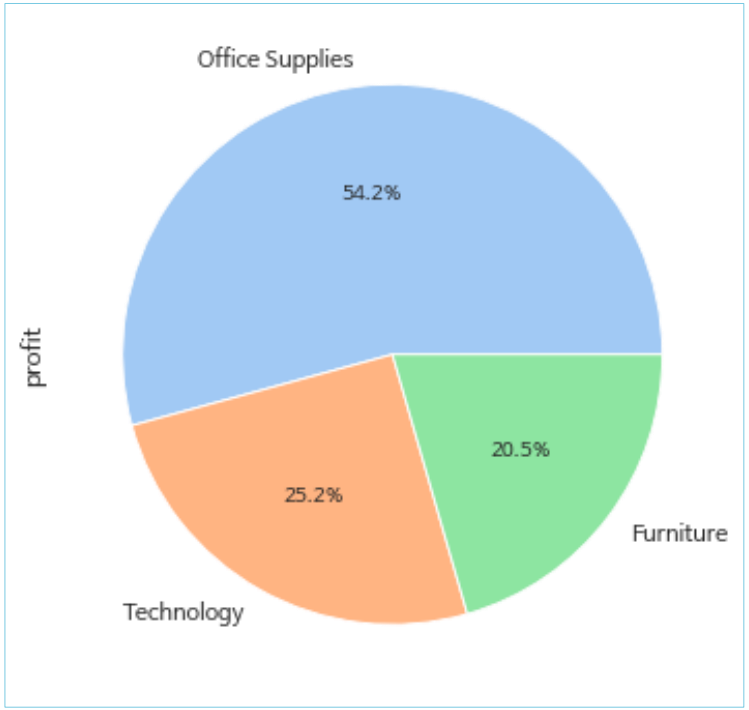
배송 방법 등을 개선해야 할 필요성이 있음

Ⅲ Project Prob A1

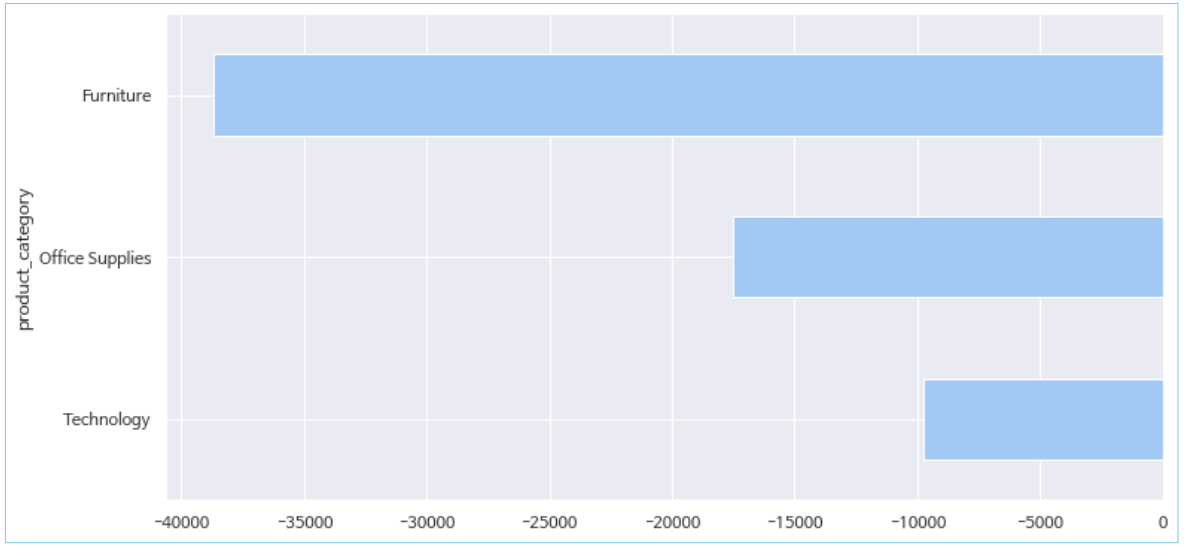
– 반품건수가 가장 높은 제품들의 개선을 통한 수익성 증대

03

product_cateory별 주문건수



product_cateory별 반품액



product_cateory별 주문건수는 가구(Furniture)가 제일 적으나 반품액은 제일 높음

Ⅲ Project Prob A1

– 반품건수가 가장 높은 제품들의 개선을 통한 수익성 증대

product_sub_cateory별 총 반품건수 및 반품액

```
1 EW_df.query("order_status == 'returned' & profit < 0").groupby(['product_category', 'product_sub_category', 'product_name'])['order_quantity'].sum().sort_values(ascending=False).iloc[:10]
```

product_category	product_sub_category	product_name	
Technology	Computer Peripherals	Imation Neon Mac Format Diskettes, 10/Pack	136
Furniture	Bookcases	Bush Mission Pointe Library	118
	Tables	Bevis 36 x 72 Conference Tables	114
Office Supplies	Pens & Art Supplies	Newell 342	111
Furniture	Bookcases	O'Sullivan 3-Shelf Heavy-Duty Bookcases	90
Technology	Telephones and Communication	Bell Sonecor JB700 Caller ID	87
Furniture	Office Furnishings	Eldon Expressions Mahogany Wood Desk Collection	87
Office Supplies	Storage & Organization	Tennsco Commercial Shelving	76
Furniture	Bookcases	O'Sullivan Elevations Bookcase, Cherry Finish	74
Office Supplies	Pens & Art Supplies	12 Colored Short Pencils	73

Name: order_quantity, dtype: int64

```
1 EW_df.query("order_status == 'returned' & profit < 0").groupby(['product_category', 'product_sub_category', 'product_name'])['profit'].sum().sort_values().iloc[:10]
```

product_category	product_sub_category	product_name	
Furniture	Bookcases	Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish	-11053.5988
		Bush Mission Pointe Library	-3068.4052
Office Supplies	Storage & Organization	Tennsco Commercial Shelving	-2324.4676
Furniture	Tables	Hon 61000 Series Interactive Training Tables	-2284.1376
Office Supplies	Storage & Organization	Tennsco Lockers, Gray	-2206.5676
Furniture	Bookcases	O'Sullivan 3-Shelf Heavy-Duty Bookcases	-1843.9664
	Tables	BoxOffice By Design Rectangular and Half-Moon Meeting Room Tables	-1679.9676
Office Supplies	Appliances	Hoover Portapower™ Portable Vacuum	-1605.7288
Furniture	Tables	Riverside Furniture Stanwyck Manor Table Series	-1536.0256
		Hon iLevel™ Computer Training Table	-1460.7688

Name: profit, dtype: float64

product_sub_cateory별 평균 리드타임 및 할인율

```
1 EW_df.groupby('product_sub_category')['lead_time'].mean().sort_values()
```

product_sub_category	
Scissors, Rulers and Trimmers	1.742268
Labels	1.958549
Computer Peripherals	1.984586
Binders and Binder Accessories	1.986231
Chairs & Chairmats	1.988327
Office Furnishings	1.996176
Storage & Organization	1.997159
Rubber Bands	2.009709
Paper	2.065217
Appliances	2.073333
Copiers and Fax	2.140351
Pens & Art Supplies	2.148058
Tables	2.158371
Office Machines	2.166667
Telephones and Communication	2.286441
Bookcases	2.359375
Envelopes	2.619632

Name: lead_time, dtype: float64

```
1 EW_df.groupby('product_sub_category')['discount_mean'].mean().sort_values()
```

product_sub_category	
Scissors, Rulers and Trimmers	0.046561
Tables	0.047458
Envelopes	0.047498
Telephones and Communication	0.048053
Computer Peripherals	0.048316
Office Furnishings	0.048474
Rubber Bands	0.048686
Copiers and Fax	0.049574
Labels	0.049643
Bookcases	0.049933
Paper	0.050196
Office Machines	0.050494
Binders and Binder Accessories	0.050651
Storage & Organization	0.050955
Chairs & Chairmats	0.050999
Pens & Art Supplies	0.051333
Appliances	0.052045

Name: discount_mean, dtype: float64

product_sub_cateory 및 product_name으로 살펴보니, 테이블(Table)과 책장(Bookcases) 제품들에서 반품건수와 반품액이 높게 나타남

다른 제품군에 비해 비교적 리드 타임도 길고 할인율도 낮음

→ 배송 과정에서의 파손, 리드타임 등의 문제점을 살펴본다면 반품율을 줄일 수 있을 것으로 보임

Ⅲ Project Prob A1

– 반품건수가 가장 높은 제품들의 개선을 통한 수익성 증대

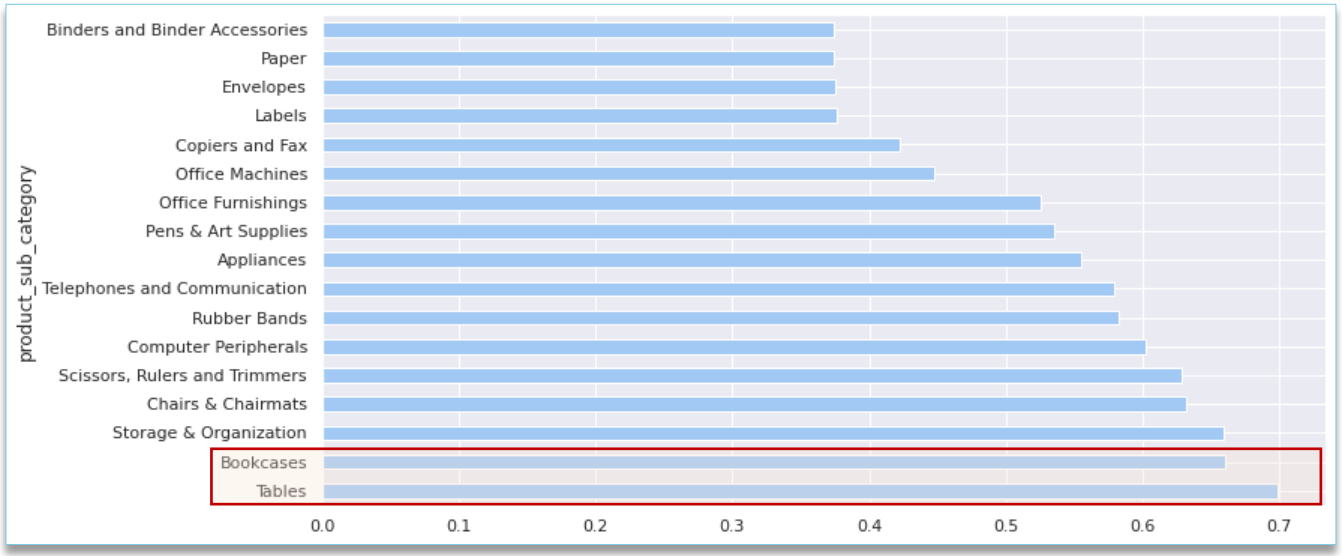
03

product_sub_cateory별 제품 당 가격

```
1 EW_df.groupby('product_sub_category')['unit_price'].mean().sort_values(ascending=False)
```

```
product_sub_category
Copiers and Fax                    681.418571
Office Machines                    643.062653
Tables                             228.720758
Bookcases                          199.138713
Chairs & Chairmats                 187.187315
Telephones and Communication        98.701462
Storage & Organization              81.374447
Appliances                          75.769638
Binders and Binder Accessories      67.794814
Computer Peripherals                41.767750
Office Furnishings                  33.663916
Scissors, Rulers and Trimmers       32.992030
Envelopes                           30.167300
Paper                               15.307849
Pens & Art Supplies                  9.845152
Labels                               4.874032
Rubber Bands                         2.851481
Name: unit_price, dtype: float64
```

product_sub_cateory별 평균 마진율



테이블과 책장에서 마진이 제일 높은 편에 속함

즉, 테이블과 책장의 경우 할인율이 전체적으로 낮게 실시하는 것을 알 수 있음

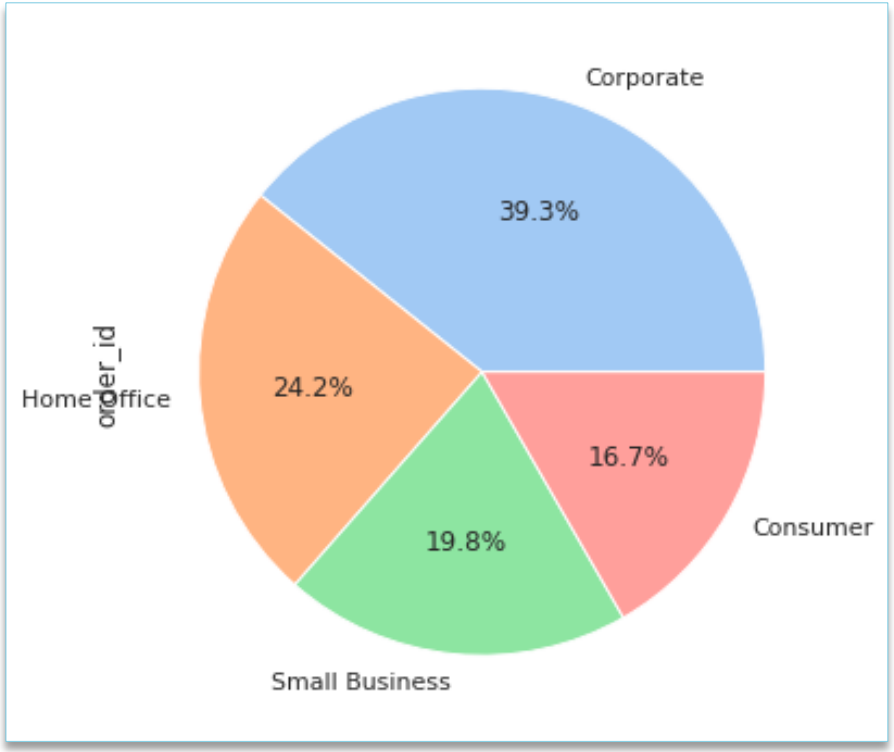
반품이 일어나는 원인(파손 등의 문제)을 찾고 해당 제품에 대해 할인 이벤트를 실시한다면 수익성이 크게 개선될 것으로 보임

Ⅲ Project Prob A1

- 고객군별 반품 현황 분석을 통한 수익성 증대

03

고객군별 반품 비율



고객군 및 sub-categpry별 반품 건수 및 반품 수량

```
1 EW_df.query("order_status == 'returned'").groupby(['customer_segment', 'product_sub_category'])['order_id'].count().sort_values(ascending=False)
```

```
customer_segment  product_sub_category
Corporate         Paper                          32
                  Telephones and Communication   23
                  Office Furnishings             21
                  Binders and Binder Accessories 20
                  Pens & Art Supplies            18
                                                 ..
Consumer          Tables                          1
                  Scissors, Rulers and Trimmers   1
Small Business    Scissors, Rulers and Trimmers   1
Consumer          Rubber Bands                    1
                  Appliances                      1
Name: order_id, Length: 65, dtype: int64
```

```
1 EW_df.query("order_status == 'returned'").groupby(['customer_segment', 'product_sub_category'])['order_quantity'].sum().sort_values(ascending=False)
```

```
customer_segment  product_sub_category
Corporate         Paper                          813
                  Office Furnishings             569
                  Telephones and Communication   552
                  Binders and Binder Accessories 542
                  Pens & Art Supplies            515
                                                 ...
Consumer          Scissors, Rulers and Trimmers   34
Corporate         Copiers and Fax                 28
Small Business    Scissors, Rulers and Trimmers   22
Consumer          Appliances                      21
Small Business    Office Machines                 14
Name: order_quantity, Length: 65, dtype: int64
```

기업 고객에게서 반품건이 제일 많은 것으로 나타났으며, Paper에서 반품 건수 및 반품 수량이 제일 많음

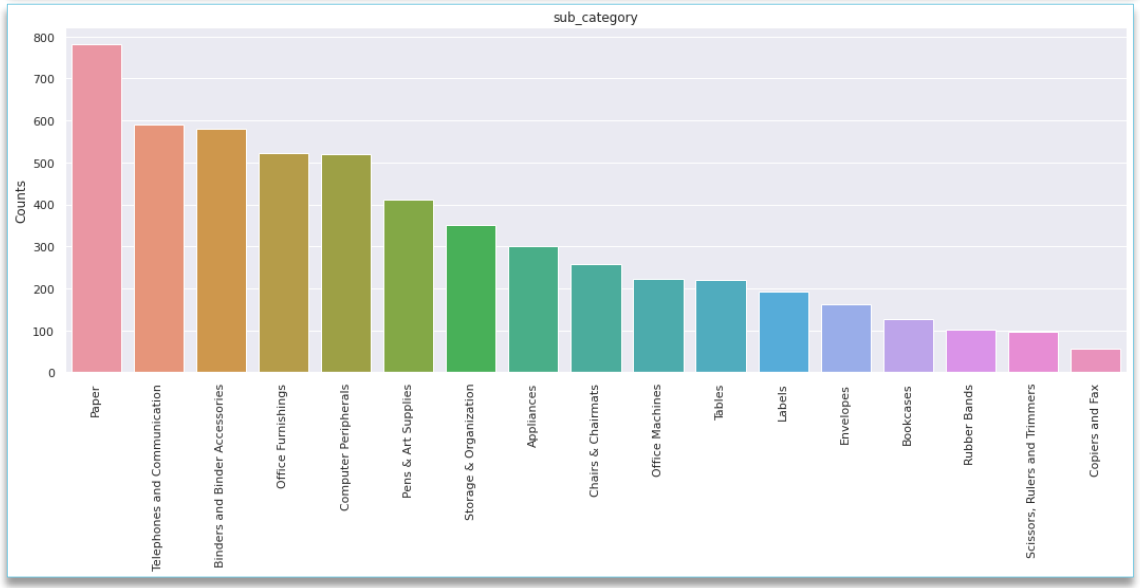
Ⅲ Project Prob A1

- 고객군별 반품건수 관련 데이터

product_sub_category별 주문 비율 및 주문 건수

```
1 EW_df.groupby(['product_category', 'product_sub_category']).record_id.count().sort_values(ascending=False).iloc[:20] /  EW_df.record_id.count()*100

product_category  product_sub_category
Office Supplies   Paper                          14.218182
Technology        Telephones and Communication   10.727273
Office Supplies   Binders and Binder Accessories 10.563636
Furniture         Office Furnishings              9.509091
Technology        Computer Peripherals            9.436364
Office Supplies   Pens & Art Supplies             7.490909
                  Storage & Organization          6.400000
                  Appliances                      5.454545
Furniture         Chairs & Chairmats              4.672727
Technology        Office Machines                 4.036364
Furniture         Tables                          4.018182
Office Supplies   Labels                          3.509091
                  Envelopes                       2.963636
Furniture         Bookcases                       2.327273
Office Supplies   Rubber Bands                    1.872727
                  Scissors, Rulers and Trimmers   1.763636
Technology        Copiers and Fax                 1.036364
Name: record_id, dtype: float64
```



product_sub_category별 반품액

```
1 EW_df.query("order_status == 'returned' & profit < 0").groupby('product_sub_category').profit.sum().sort_values()

product_sub_category
Bookcases                  -19571.640800
Tables                     -13942.961379
Storage & Organization      -9196.891200
Telephones and Communication -5349.078200
Paper                       -3493.239600
Office Furnishings          -3308.284800
Computer Peripherals        -2085.047600
Chairs & Chairmats          -1873.292800
Appliances                  -1808.812800
Binders and Binder Accessories -1617.184100
Copiers and Fax             -1312.838800
Office Machines             -1000.674300
Pens & Art Supplies          -973.797600
Envelopes                    -200.125200
Scissors, Rulers and Trimmers -127.826400
Rubber Bands                  -65.540400
Labels                         -4.966400
Name: profit, dtype: float64
```

Paper 제품군의 주문율은 약 14%이며, 반품액은 5번로 높은 것을 알 수 있음

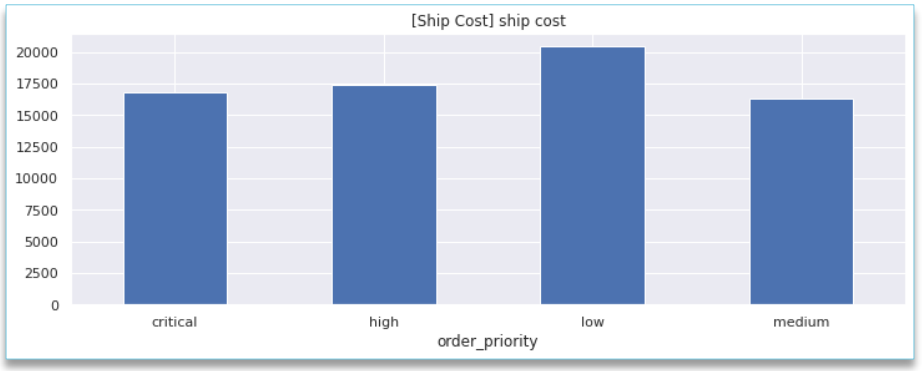
기업의 경우 Paper을 필수불가결하게 많이 쓰게 됩니다.

따라서, 기업 고객을 대상으로 Paper 제품군의 할인을 진행하여 여러 제품을 한꺼번에 구매할 수 있도록 유도 할 수 있으며, 할인 이벤트 진행을 통해 저렴하게 구매하였다는 인식을 통해 반품 비율도 줄어 들 수 있음

Ⅲ Project Prob A1

– 배송 우선순위 관련 데이터

03



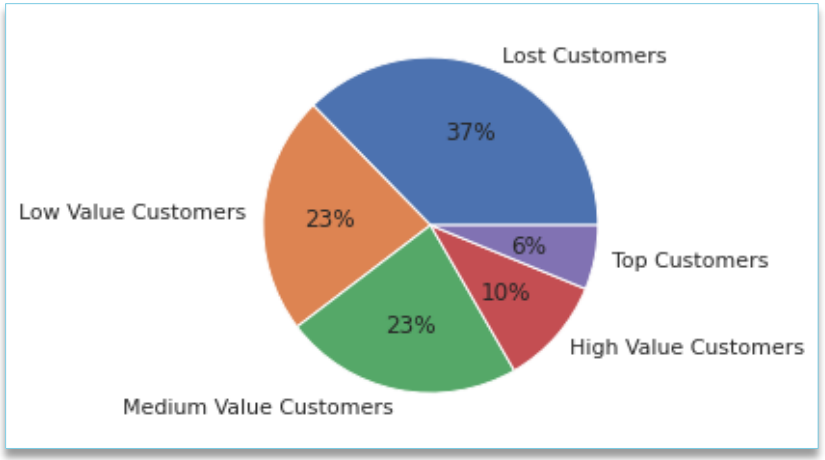
Order Priority의 critical 형이 Profit이 제일 낮게 나타남
배송비용과 배송건수를 보면 각 형식은 별차이가 없음

주문 우선순위 기준을 다시 파악하고 따라서 배송비용도 다시 산정할 필요가 있음

Ⅲ Project Prob A1

– 배송 우선순위 관련 데이터

고객수별 고객층 비율



	Customer Information	order_date	Recency	Frequency	Monetary
0	Aaron Mawkins - Bedford	2018-12-23	371	1	-317.478800
1	Aaron Mawkins - Camden	2018-12-23	371	1	-693.229195
2	Aaron Mawkins - Cranston	2018-02-22	675	1	23.121200
3	Aaron Mawkins - Fall River	2018-10-28	427	1	79.341200
4	Aaron Mawkins - Pennsauken	2017-10-02	818	1	193.121200

	CustomerSegment	Total Revenue	Total Customers
3	Medium Value Customers	335766.52	457
0	High Halue Customers	312592.94	208
2	Low Value Customers	164742.69	457
4	Top Customers	159998.74	123
1	Lost Customers	125057.92	742

RFM은 얼마나 최근에(R), 얼마나 자주(F), 얼마나 많이(M) 구매활동을 했는가에 대한 정보를 만들고 이를 바탕으로 고객의 상태를 세분화하는 모델임

Consolidated Score = 0.15*R + 0.28*F + 0.57*M

Top Customers와 High value Customers는 16% 밖에 안되지만,
큰 이익률을 가지고 있어서 고객만족도를 주의해야 함

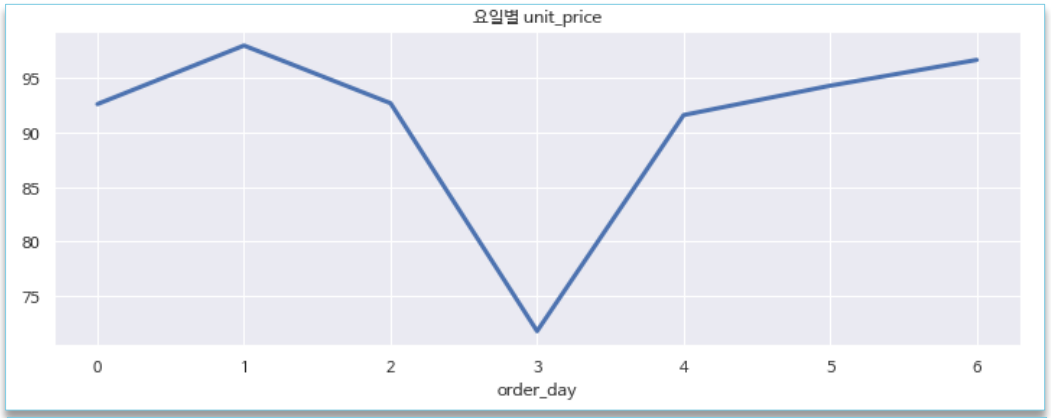
Medium Value Customers와 Low Value Customers 각 고객층은
총 이익에 30% 이상 기여하고 있어
더 높은 고객층으로 유도할 수 있는 적절한 마케팅 전략 수립 필요함

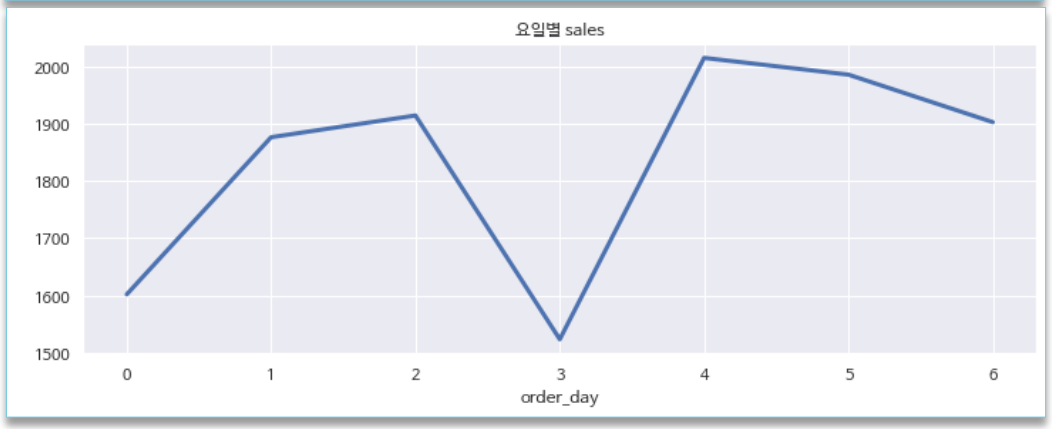
반면, Lost Customers가 37%을 차지하지만 총 수익은 11%에 불과하므로
고객 충성도를 높일 수 있는 프로모션 전략을 고민할 필요가 있음
→ 마일리지 제도 활용 추천

Ⅲ Project Prob A1

- 구매 혜택을 통한 수익성 증대

03





목요일에 고객들이 구매하는 상품의 단가도 낮고 주문 수량도 낮은 것을 확인할 수 있음

→ 이익과 매출의 감소로 이어짐

→ 따라서 목요일에 구매에 대한 혜택을 주면 보다 많은 소비자들이 구매하여 수익성을 증대할 수 있음

월요일에는 고객들이 구매하는 상품의 단가는 높은데 구매 수량이 현저히 낮음

→ 월요일 고객들이 더 많은 수량을 구매할 수 있도록 하는 마케팅 전략 수립 필요

Project Prob A2

“ 제공된 ElectroWorld 데이터를 이용하여 profit을 예측할 수 있는 머신러닝 모델을 만들어보세요.
그리고 학습된 머신러닝 모델로, 주어진 “test_set.csv” 에 대해 profit을 예측해보세요.
예측한 결과를 “ans_profit.csv” 형식으로 작성하여 첨부 제출하세요.
(머신러닝 모델의 성능은 정확할수록 좋습니다.) ”

TEAM 3가 사용한 머신러닝 모델 기법

Linear Regression
Ridge
Lasso
Decision Tree Regression
Random Forest Regression
XGBoost
LightGBM

기본 데이터로 먼저 적합 시킨 뒤 데이터의 이상치를 제거한 후 다시 적합

* profit 5500이상, -5000이하인 데이터 삭제

Ⅳ Project Prob A2

- Machine Learning Model

04

Hyperparameter
random_state : 43
train : test = 8 : 2
cv = 5,7,10
n_estimators = 10,50,100,150,200,250,300,500,1000,1500,2000,2300,2400,2500 등 여러가지 시도
max_depth = 3,4,5,6,7,8,9,10,14,15
max_features = 3,4,5,6,7,8,9,10
min_samples_split = 8,9,10,11,12
eta = 0.01,0.05,0.1,0.15,0.2
learning_rate = 0.01,0.1
gamma = 0.5,1,2

Ⅳ Project Prob A2

- Machine Learning Model

04

순서	모델	데이터	하이퍼파라미터	MAE	RMSE
1	Linear Regression	이상치 제거	x	266.9690649	260265.8169
2	Ridge	이상치 제거	x	266.6151548	259802.2114
3	Lasso	이상치 제거	x	261.6325096	256813.1668
4	Decision Tree	이상치 제거	x	153.9656367	227175.6319
5	Random Forest	이상치 제거	n_estimators = 50	10.51458204	404.2082675
6	XGBoost	이상치 제거	[Grid Search] cv = 10 max_depth = 4 eta = 0.01 gamma = 0.5 learning_rate = 0.1 n_estimators = 3000	9.710626922	400.6099706
7	LightGBM	이상치 제거	[Grid Search] cv = 5 max_depth = 5 learning_rate = 0.06 n_estimators = 4000	10.1308911	431.668326

시도해보았던 모델 중 XGBoost의 MAE가 제일 높아서 이 모델 채택함
이후, 이를 학습시켜 test data set을 채움

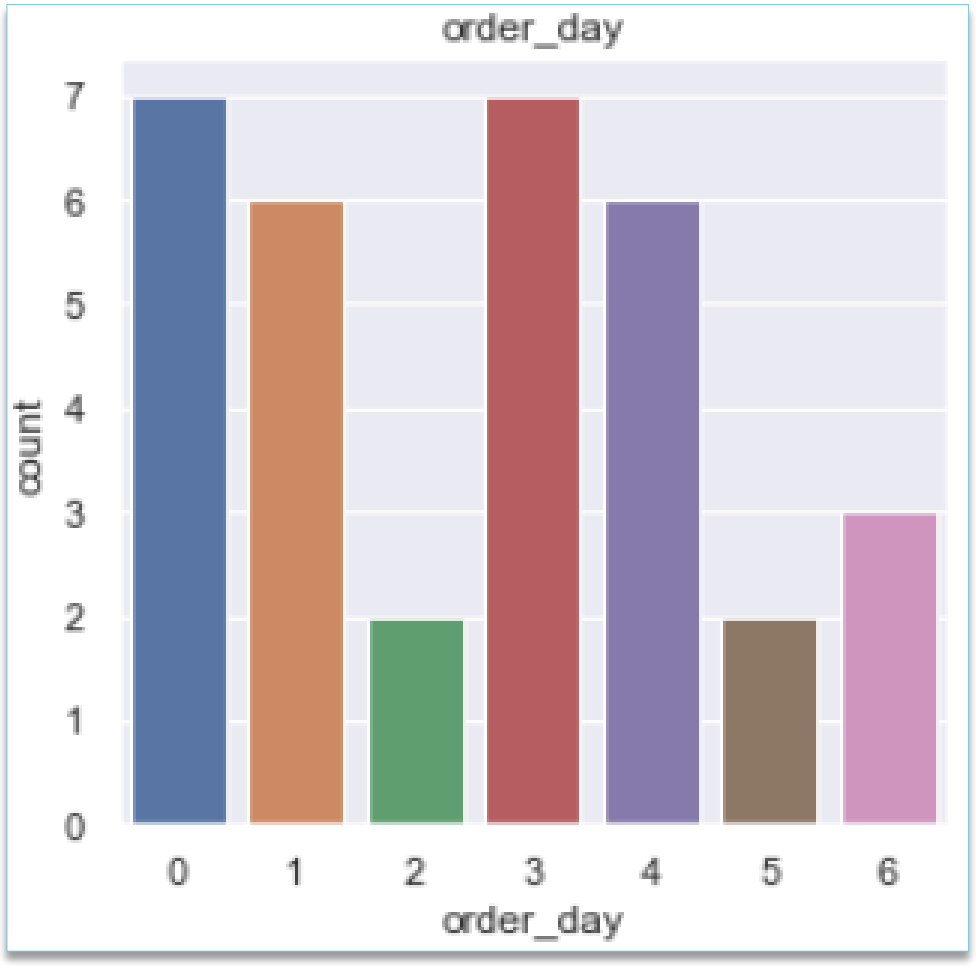
Project Prob A3

“test_set.csv”에 있는 주문들 중 머신러닝 모델에서 profit이 낮게 예측되는 주문에 대해 분석해보세요.
이들 주문에 대해 어떠한 조치를 취하면 profit을 개선할 수 있는지 방안을 찾아보세요.
그리고 이러한 방안을 적용한 후, profit을 ML모델로 다시 한 번 예측하여 효과가 있는지 살펴보세요.

V Project Prob A3

– Machine Learning Model

05



A2 예측 결과(test data) EDA

수익성이 낮은 상품들에 대해 분석
→ 수익성이 – 인 항목만 추출

A1에서 목요일에 고객들의 적은 주문수량이 이익과 매출의 감소로 이어지는 것을 확인
BUT, 오히려 수익성이 낮은 물품 중 반품된 건은 월요일과 목요일에 많음

V Project Prob A3

– Machine Learning Model

05

	record_id	product_category
46	5546	Office Supplies
52	5552	Technology
68	5568	Furniture
100	5600	Office Supplies
231	5731	Office Supplies
263	5763	Technology
282	5782	Office Supplies

A2 예측 결과(test data) EDA

월요일, 목요일에 반품된 15건 중 Office Supplies가 차지하는 비중이 가장 큼
→ 이에 대한 개선 필요함을 느낌

A1에서 세웠던 수익성 증가 방안이 효과가 있는지 보기 위해 할인율을 높여 다시 확인
→ discount를 2배 올려 다시 예측

V Project Prob A3

- Machine Learning Model

05

record_id	기존 discount의 profit	새로운 discount의 profit	개선여부
5546	-13.575133	49.971897	개선됨
5552	-6.490261	884.6463	개선됨
5568	-201.901242	-746.8735	개선되지 않음
5600	-698.814275	4847.7915	개선됨
5731	-86.697383	29.57025	개선됨
5763	-28.605297	68.21261	개선됨
5782	-108.618651	-243.26382	개선되지 않음

7개 중 2개를 제외한 나머지는 모두 개선되는 효과가 있는 것으로 나타남

심지어 개선 효과가 -수익성에서 +수익성으로 바뀜

할인율을 더 높이면 수익성을 더 개선할 수 있을 것으로 보임

Appendix

정리

EletroWorld Data 분석 결론 및 제언

- **대체적으로 반품 건(returned)에 대해 살펴보고 수익성을 개선할 수 있는 방안에 집중하여 데이터를 분석하였음**
- **주요 개선점은 할인과 배송으로 볼 수 있음**
  - Prob1을 통해 살펴본 결과, Electro Word회사는 배송 과정에서의 파손, 배송비 등을 집중적으로 살펴볼 필요가 있으며, 주문 날짜로부터 출고날짜의 term을 줄인다면 반품율을 줄여 수익성 개선이 가능 할 것으로 보임
  - 일부 상품군은 마진율이 높고 할인율이 전체적으로 낮음을 확인할 수 있었음
  - Prob3을 통해 확인했듯이, 할인율을 높이면 수익성 개선에 긍정적인 영향을 주기 때문에 이를 참고하여 할인 이벤트, 쿠폰 등의 프로모션을 진행할 필요가 있음
- **기업 고객을 대상으로 일부 상품에 대한 마일리지 제도, 할인쿠폰 등의 프로모션을 진행하여 구매를 유도하여 대량구매로 이어지게 할 수 있을 것으로 보이며, 나아가 충성 고객으로 까지 만들면 수익성이 극대화 될 수 있을 것으로 보임**
- **추가로, test data에는 상품 발송일자(ship_date) 컬럼이 없어 테스트 해보지 못했지만, 리드타임을 앞당긴다면 profit 개선에 도움이 되는지 살펴보면 좋을 것 같음**

프로젝트 담당 업무	
Data Preprocessing 및 EDA	이윤경, 진현지
Project Prob 1	이윤경, 진현지, 황투히엔
Project Prob 2	이윤경, 진현지
Project Prob 3	진현지
PPT	이윤경
발표	진현지

“

이상 3조의 발표를 들어 주셔서
감사합니다.

”