E-Commerce
데이터 분석

이중훈

DATA INTRODUTION

데이터 소개

영국에 기반을 둔 등록된 비점포 온라인 소매에 대해 2010년 1월 12일과 2011년 12월 9일 사이에 발생하는 모든 거래를 포함하는 다국적 데이터 셋입니다.

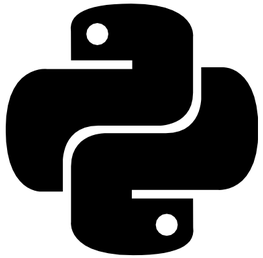
변수목록	
주문번호	InvoiceNo
상품번호	StockCode
상품명	Description
구매량	Quantity
판매일자	InvoiceDate
가격	UnitPrice
고객번호	CustomerID
판매국가	Country

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
539993	22386	JUMBO BAG PINK POLKADOT	10	2011-01-04 10:00	1.95	13313	United Kingdom
539993	21499	BLUE POLKADOT WRAP	25	2011-01-04 10:00	0.42	13313	United Kingdom
539993	21498	RED RETROSPOT WRAP	25	2011-01-04 10:00	0.42	13313	United Kingdom
539993	22379	RECYCLING BAG RETROSPOT	5	2011-01-04 10:00	2.1	13313	United Kingdom
539993	20718	RED RETROSPOT SHOPPER BAG	10	2011-01-04 10:00	1.25	13313	United Kingdom
539993	85099B	JUMBO BAG RED RETROSPOT	10	2011-01-04 10:00	1.95	13313	United Kingdom

ANALYSTIC METHOD

분석 방법

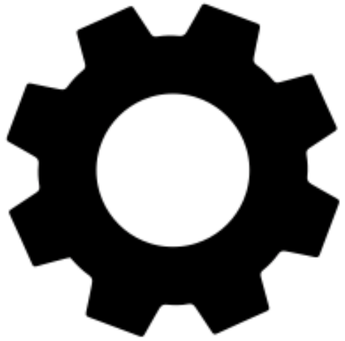
결측치 및 오류값은 파이썬으로 전처리한 후 SQL을 이용해 DB에 데이터를 적재하고 SQL문으로 분석에 필요한 지표 및 수치 데이터를 조회해 파이썬에서 데이터프레임으로 변환해 시각화 진행.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	539993	22386	JUMBO BAG PINK POLKADOT	10	2011-01-04 10:00	1.95	13313.0	United Kingdom
1	539993	21499	BLUE POLKADOT WRAP	25	2011-01-04 10:00	0.42	13313.0	United Kingdom
2	539993	21498	RED RETROSPOT WRAP	25	2011-01-04 10:00	0.42	13313.0	United Kingdom
3	539993	22379	RECYCLING BAG RETROSPOT	5	2011-01-04 10:00	2.10	13313.0	United Kingdom
4	539993	20718	RED RETROSPOT SHOPPER BAG	10	2011-01-04 10:00	1.25	13313.0	United Kingdom

DATA PREPROCESSING

데이터 전처리



01 결측치 제거 및 10개국 추출

```
United Kingdom    354345
Germany             9042
France              8342
EIRE                7238
Spain               2485
Netherlands         2363
Belgium             2031
Switzerland         1842
Portugal            1462
Australia           1185
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	539993	22386	JUMBO BAG PINK POLKADOT	10	2011-01-04 10:00	1.95	13313.0	United Kingdom
1	539993	21499	BLUE POLKADOT WRAP	25	2011-01-04 10:00	0.42	13313.0	United Kingdom
2	539993	21498	RED RETROSPOT WRAP	25	2011-01-04 10:00	0.42	13313.0	United Kingdom
3	539993	22379	RECYCLING BAG RETROSPOT	5	2011-01-04 10:00	2.10	13313.0	United Kingdom
4	539993	20718	RED RETROSPOT SHOPPER BAG	10	2011-01-04 10:00	1.25	13313.0	United Kingdom
...	...	...	...	...	...	...	...	...
364667	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50	0.85	12680.0	France
364668	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50	2.10	12680.0	France
364669	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50	4.15	12680.0	France
364670	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50	4.15	12680.0	France
364671	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50	4.95	12680.0	France

364672 rows × 8 columns

국가 추출

Country에 국가가 38개가 있었으며, 몇몇 국가는 데이터가 너무 적어 상위 10개 국가만 추출하여 최종 데이터셋을 구성함.

2011년도 데이터 사용

2010년 데이터는 제거한 후 2011년 데이터만 분석에 사용

결측치 제거

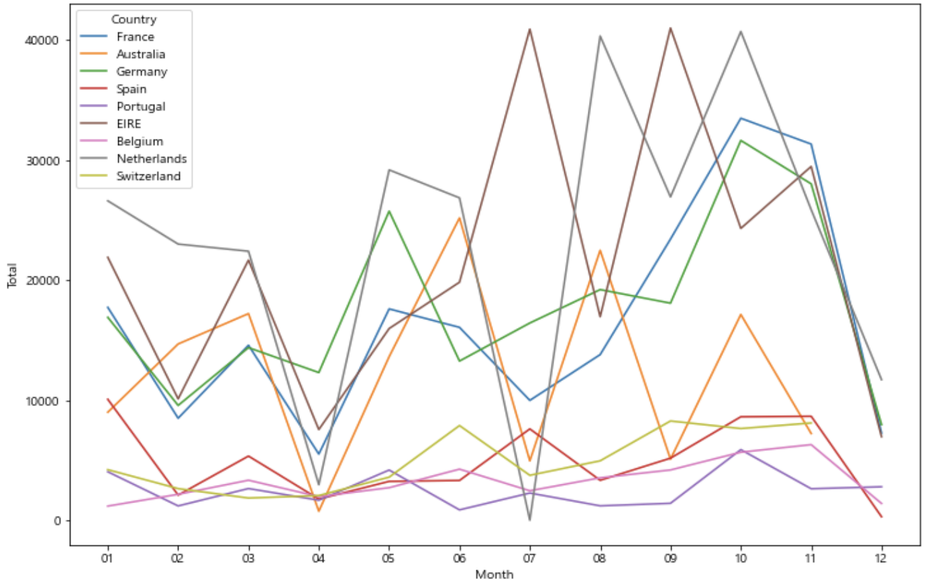
Descripiton에 1454개 결측치
CustomerID에 135080개 결측치
Quality에 음수인 데이터가 존재해 제거

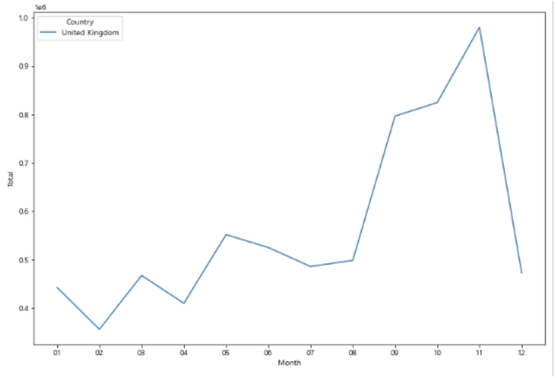
최종 데이터 : 364672개

COUNTRY ANALYSIS

국가별 분석

01 월별 매출 추이





Month	01	02	03	04	05	06	07	08	09	10	11	12
Country												
Australia	9017.71	14695.42	17223.99	771.60	13638.41	25187.77	4964.38	22489.20	5106.73	17150.53	7242.72	nan
Belgium	1200.20	2181.07	3351.98	1989.48	2732.40	4274.82	2475.57	3554.02	4208.02	5685.39	6315.76	1417.73
EIRE	21904.19	10126.52	21674.36	7570.50	15982.19	19835.99	40905.15	16967.38	40995.49	24317.92	29473.41	6978.92
France	17740.12	8515.96	14589.55	5529.61	17614.88	16078.97	10000.19	13810.96	23428.04	33485.45	31337.09	7276.92
Germany	16910.84	9581.05	14392.69	12315.54	25751.20	13274.10	16440.98	19220.77	18091.22	31638.42	28025.02	7984.17
Netherlands	26611.10	23011.91	22416.49	2976.56	29185.88	26858.09	26.02	40327.81	26937.26	40708.65	25874.01	11728.02
Portugal	4055.71	1213.90	2660.85	1687.75	4202.23	884.46	2287.85	1221.40	1433.22	5899.46	2644.90	2808.19
Spain	10086.09	2114.50	5363.15	1785.65	3257.60	3333.21	7624.92	3346.91	5189.24	8636.94	8678.96	316.21
Switzerland	4231.23	2654.92	1870.23	2076.94	3610.01	7904.15	3762.65	4969.89	8284.86	7655.19	8118.96	nan

영국의 월별 매출

연말로 갈수록 매출이 상승하다가 12월에는 감소하는 경향을 보임. 매출이 12월 9일까지밖에 존재하지 않아 그런것으로 보임.

(영국은 데이터수가 많아 매출 규모가 커 다른국가와 같이 보면 해석하기가 어려워 따로 출력.)

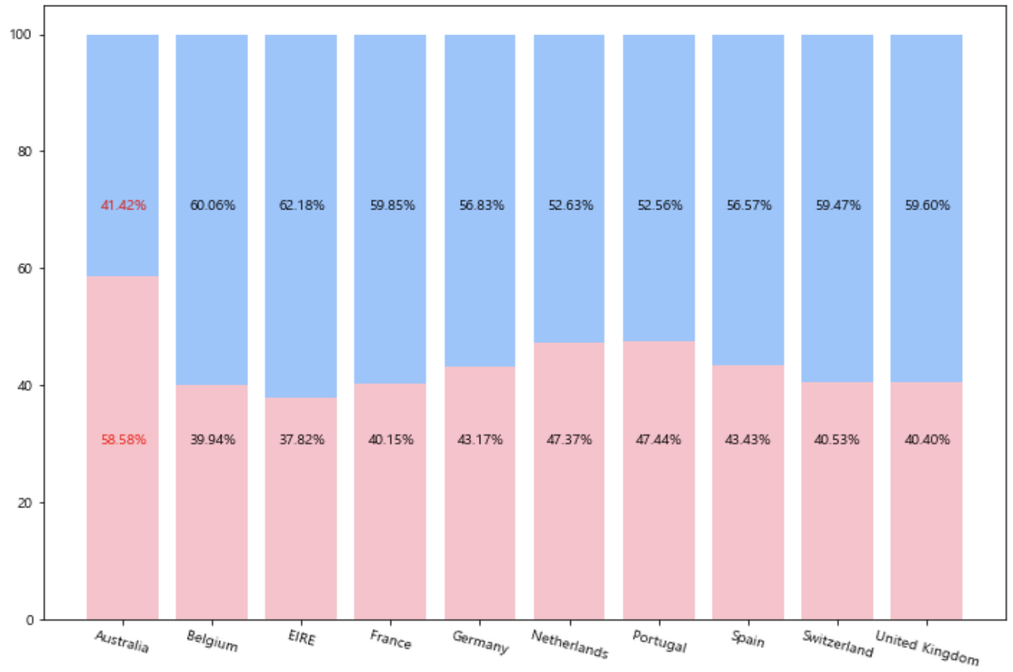
나머지 국가의 월별 매출

대체로 네덜란드가 매출이 높은 편.
포르투갈과 벨기에는 1년 내내 매출규모가 작음.

특히 네덜란드는 매출건수 대비 매출이 높은 편.
(매출 건수는 2363건으로 10개국 중 6위인데 반해 월별 총매출은 대부분 영국 다음으로 높음)

대부분의 국가가 상반기로 갈수록 매출이 상승하는 것으로 보임.

02 상/하반기 매출 분석

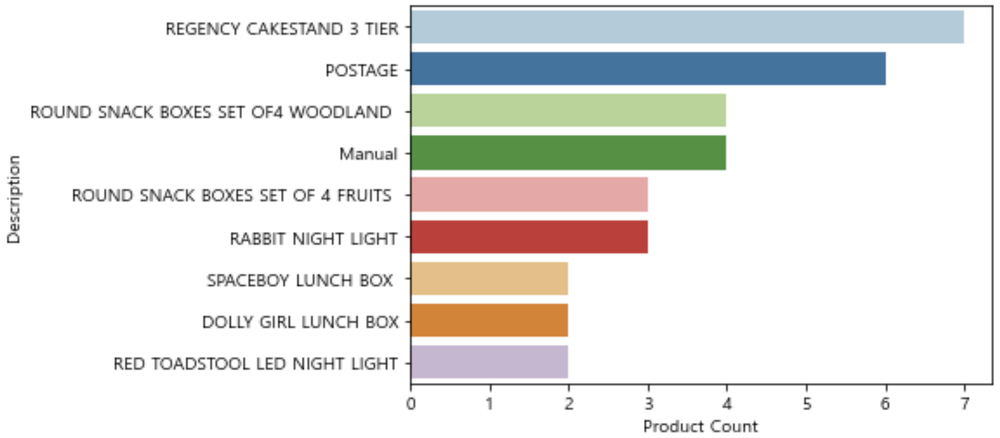


Country	H1	H2	Year
Australia	80534.900	56953.560	137488.460
Belgium	15729.950	23656.480	39386.430
EIRE	97093.750	159638.270	256732.020
France	80069.090	119338.650	199407.740
Germany	92225.420	121400.580	213626.000
Netherlands	131060.090	145601.770	276661.860
Portugal	14704.900	16295.020	30999.920
Spain	25940.200	33793.180	59733.380
Switzerland	22347.480	32791.550	55139.030
United Kingdom	2751087.721	4058641.983	6809729.704

상/하반기 매출 비중

국가별 월별 매출추이를 보면 하반기로 갈수록 매출이 높아지는 경향을 볼수있었음.
빨간색은 상반기 매출 비중 ,파란색은 하반기 매출 비중
오스트레일리아만 상반기 58.58%,하반기 42.41%로 상반기 매출비중이 더 높고,
그 외 모든 국가가 하반기 매출이 더 높은 것으로 나타남.

03 최대 매출 상품 분석



	Country	Description
0	Germany	REGENCY CAKESTAND 3 TIER
1	United Kingdom	REGENCY CAKESTAND 3 TIER
2	Australia	REGENCY CAKESTAND 3 TIER
3	Switzerland	REGENCY CAKESTAND 3 TIER
4	EIRE	REGENCY CAKESTAND 3 TIER
5	Spain	REGENCY CAKESTAND 3 TIER
6	France	REGENCY CAKESTAND 3 TIER
7	Belgium	POSTAGE
8	Switzerland	POSTAGE
9	Spain	POSTAGE
10	France	POSTAGE
11	Portugal	POSTAGE
12	Germany	POSTAGE
13	Germany	Manual
14	EIRE	Manual
15	Portugal	Manual
16	France	Manual
17	Belgium	ROUND SNACK BOXES SET OF4 WOODLAND
18	Germany	ROUND SNACK BOXES SET OF4 WOODLAND
19	Switzerland	ROUND SNACK BOXES SET OF4 WOODLAND
20	Netherlands	ROUND SNACK BOXES SET OF4 WOODLAND
21	France	RABBIT NIGHT LIGHT
22	Netherlands	RABBIT NIGHT LIGHT
23	Australia	RABBIT NIGHT LIGHT
24	Germany	ROUND SNACK BOXES SET OF 4 FRUITS
25	Netherlands	ROUND SNACK BOXES SET OF 4 FRUITS
26	Belgium	ROUND SNACK BOXES SET OF 4 FRUITS
27	Netherlands	DOLLY GIRL LUNCH BOX
28	Belgium	DOLLY GIRL LUNCH BOX
29	Australia	RED TOADSTOOL LED NIGHT LIGHT
30	France	RED TOADSTOOL LED NIGHT LIGHT
31	Netherlands	SPACEBOY LUNCH BOX
32	Belgium	SPACEBOY LUNCH BOX
33	EIRE	3 TIER CAKE TIN RED AND CREAM
34	Spain	BLUE 3 PIECE POLKADOT CUTLERY SET
35	EIRE	CARRIAGE
36	Spain	CHILDRENS CUTLERY POLKADOT BLUE
37	Spain	CHILDRENS CUTLERY POLKADOT PINK
38	EIRE	JAM MAKING SET WITH JARS
39	United Kingdom	JUMBO BAG RED RETROSPOT
40	Portugal	JUMBO SHOPPER VINTAGE RED PAISLEY
41	Portugal	LUNCH BAG RED RETROSPOT
42	United Kingdom	MEDIUM CERAMIC TOP STORAGE JAR
43	United Kingdom	PAPER CRAFT , LITTLE BIRDIE
44	Switzerland	PLASTERS IN TIN SPACEBOY
45	Switzerland	PLASTERS IN TIN WOODLAND ANIMALS
46	Portugal	RETROSPOT TEA SET CERAMIC 11 PC
47	Australia	SET OF 3 CAKE TINS PANTRY DESIGN
48	Australia	SET OF 6 SPICE TINS PANTRY DESIGN
49	United Kingdom	WHITE HANGING HEART T-LIGHT HOLDER

국가별 최대 매출 상품 TOP 5

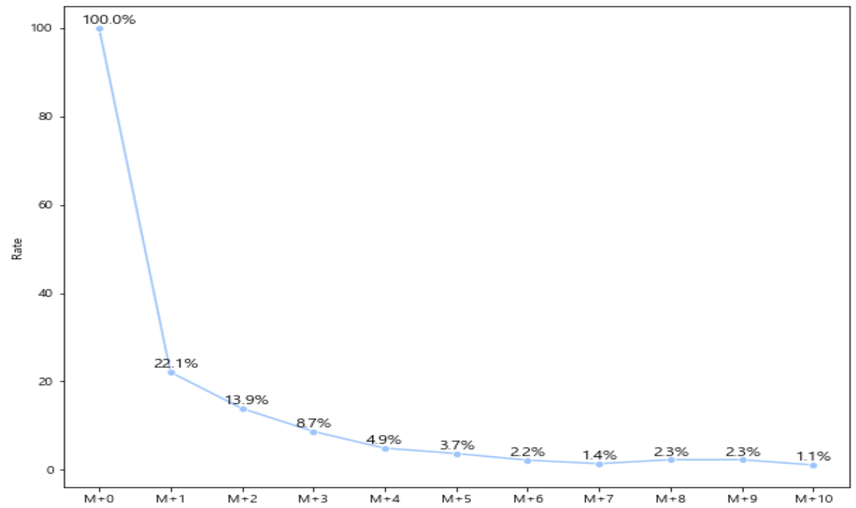
국가별 최대 매출 상품들을 추출해 보았을 때,
REGENCY CAKESTAND 3 TIER,POSTAGE 상품이 대부분의 국가에서도 가장 높은 매출을 차지하는 것을 볼 수 있음.

(그래프는 2번이상 겹친 상품만 출력)

가장 많이 겹친 상품

REGENCY CAKESTAND 3 TIER는 총 10개국 중 7개국가인 독일 영국,오스트레일리아,스위스,아일랜드,프랑스,스페인이 겹칠정도로 판매가 잘되는 것을 알 수 있음.

04 RETENTION RATE 분석



영국의 고객 유지율

첫 구매일자가 1월인 고객 기준으로 마지막 달까지 유지된 고객의 비율은 약 1.1%
2월부터 약 78%가 빠져나갔으며 그 이후에는 비교적 천천히 줄어든 것으로 보임.

(다른 국가는 데이터의 수가 적어 Retention Rate를 분석하기에 적합하지 않아 영국의 데이터로만 분석함.)

월별 고객 유지율

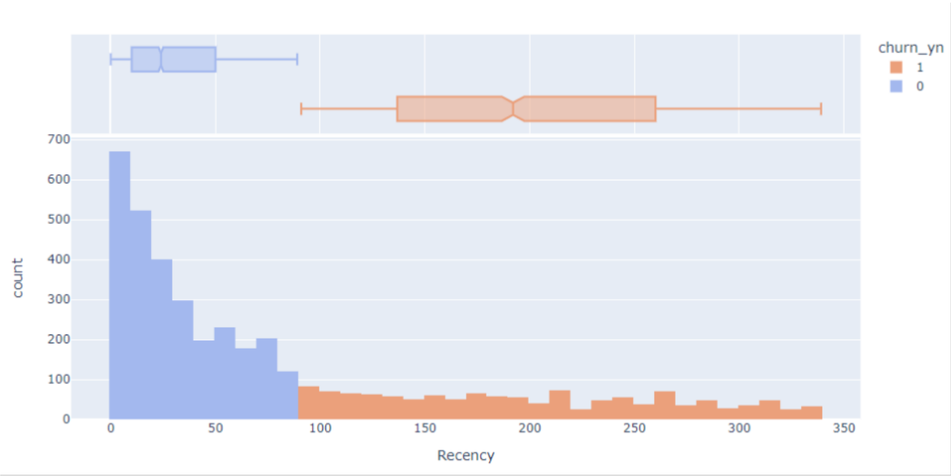
M+1을 기준으로 월별로 유지율이 가장 높은 달은 9월이며 유지율은 34%

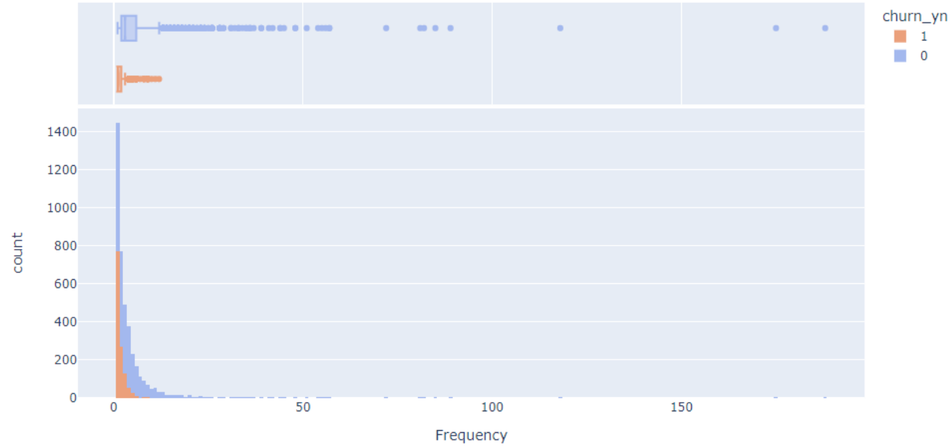
M+1을 기준으로 월별로 유지율이 가장 낮은 달은 10월이며 유지율은 20%
(데이터에 12월 9일까지의 데이터만 있으므로 11월~12월 이탈률은 제외)

CUSTOMER ANALYSIS

고객별 분석

01 고객별 지표 분석-1





Recency(최근성)

제일 마지막 구매일일 12월9일을 기준으로 Recency를 구했을 때, 이탈한 고객이 이탈하지 않은 고객보다 적었음을 확인함.

이탈하지 않은 고객은 평균 31일 동안 구매이력이 없고,이탈한 고객은 평균 199일 동안 구매이력이 없는 것으로 나타나며 약 6배 차이를 보임.

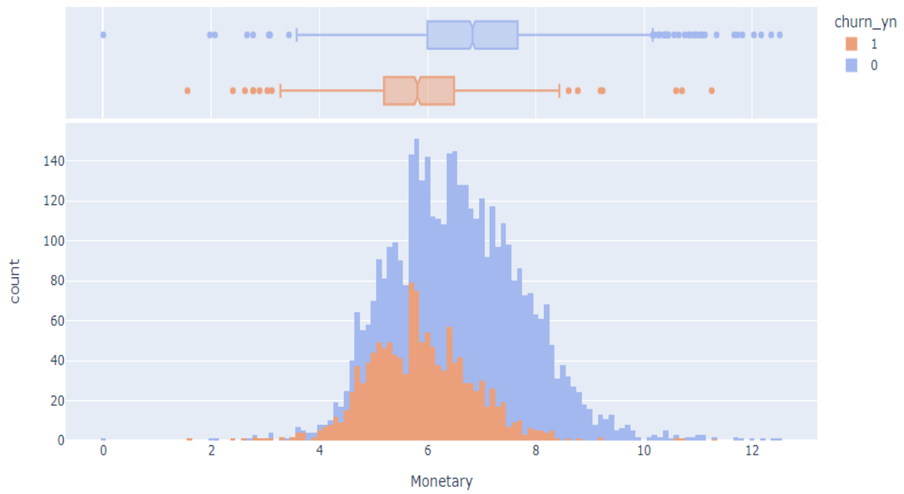
Frequency(빈도)

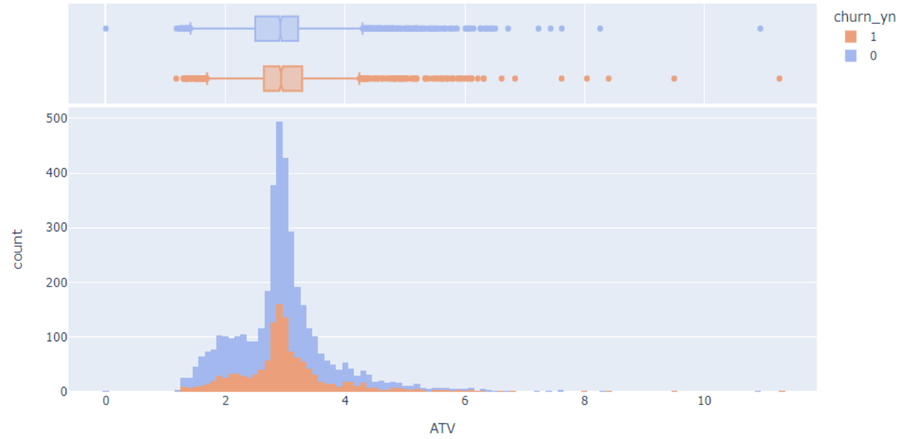
이탈한 고객은 구매빈도가 낮은쪽에 분포하여있고, 이탈하지 않은 고객은 구매빈도가 높은 쪽에 분포해 있는 것으로 보임.

이탈고객 중 가장 많이 구매한 고객은 188번,이탈하지 않은 고객 중 가장 많이 구매한 고객은 12번으로 10배 이상 차이를 보임. 또한 각각 이탈 고객은 평균 1.76번 구매했으며, 이탈하지 않은 고객은 5.12번 구매해 약 3.5배 차이를 보임.

Churn_yn : Recency를 기준으로 3달이상 지난 경우를 이탈고객으로 1,이탈고객이 아닌 경우 0

02 고객별 지표 분석-2





Monetary(구매액)

이탈한 고객은 대체로 구매금액이 낮은곳에 분포되어 있고, 이탈하지 않은 고객은 구매금액이 높은 곳에 분포함.

이탈한 고객은 최대 77184파운드 , 이탈하지 않은 고객은 271614파운드까지 구매하였으며 평균은 각각 이탈고객 686파운드,이탈하지 않은 고객 2554파운드로 나타나며 4배 차이를 보임.

(원자료가 왼쪽으로 데이터가 많이 쏠려 분포를 잘 보기위해 로그변환해 분포도를 출력)

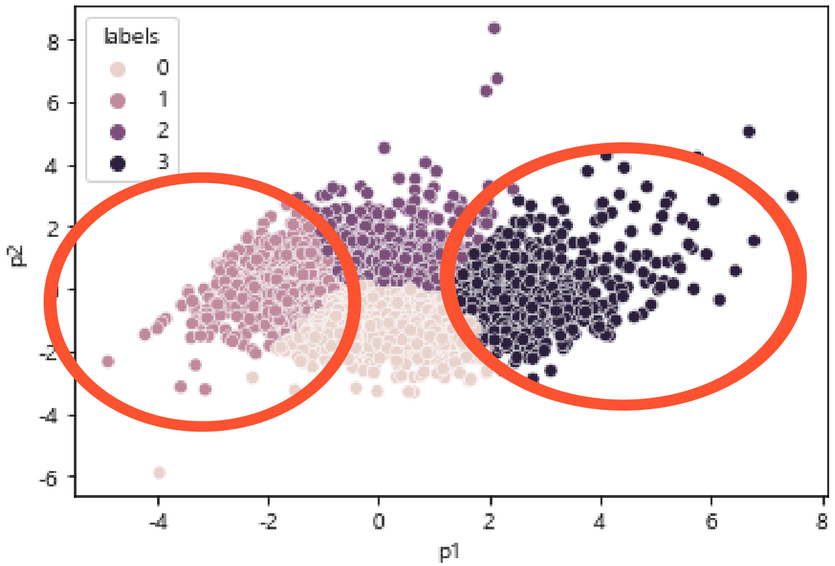
ATV(건당 구매금액)

ATV는 이탈여부에 따른 분포의 차이가 거의 없는 것으로 보임.
T-test검정 결과 t=1.18, p-value=0.24로 실제로 분포 차이가 없음을 확인.
(ATV를 제외한 다른 변수들은 검정결과 분포의 차이가 있음을 확인함.)

Churn_yn : Recency를 기준으로 3달이상 지난 경우를 이탈고객으로 1,이탈고객이 아닌 경우 0

03 고객 세그먼트

	cluster0	cluster1	cluster2	cluster3
Recency	75.334906	180.741512	23.395492	8.760417
Frequency	3.815094	1.360340	2.220287	11.522135
Monetary	1817.486050	300.744445	486.598064	6885.006276
ATV	144.664211	23.359390	14.672518	114.078278



K-means 군집화 결과

RFM과 ATV Feature들로 Standard, MinMax, Log Transform 진행 후 군집화 실행한 결과 Log 변환이 군집화가 잘 된것으로 보였음.
(2,3,4,5개 군집 결과 실루엣 계수는 낮지만 4개의 군집이 비교적 골고루 고객을 군집화 하였다고 판단함.)

0번 군집은 건당 구매금액은 높으나, 구매이력이 오래된 고객
1번 군집은 구매이력도 오래되고,자주구매하지 않으며,지출도 적은 고객
2번 군집은 구매이력이 오래되지 않았지만, 지출이 낮은 고객
3번 군집은 자주 구매하고, 총지출도 높은 고객

PCA 차원축소 시각화

2차원 평면에 군집별 시각화를 하기위해 4개변수를 2차원으로 축소함.

시각화 한 결과도 log 변환 후 시각화한 결과가 가장 군집별 경계점이 명확했으며,
1번 군집과 3번 군집이 가장 멀리 떨어진 것으로 보아 두 군집이 가장 고객별 차이가 크다고 예상됨.
집단별 수치를 보면 두 집단에 가장 큰 차이를 보이는 지표는 Recency와 Monetary 임을 알 수 있음.