온라인 리테일 데이터 분석

서론

ARPU(ARPPU)
코호트 분석

RFM

군집화

서론

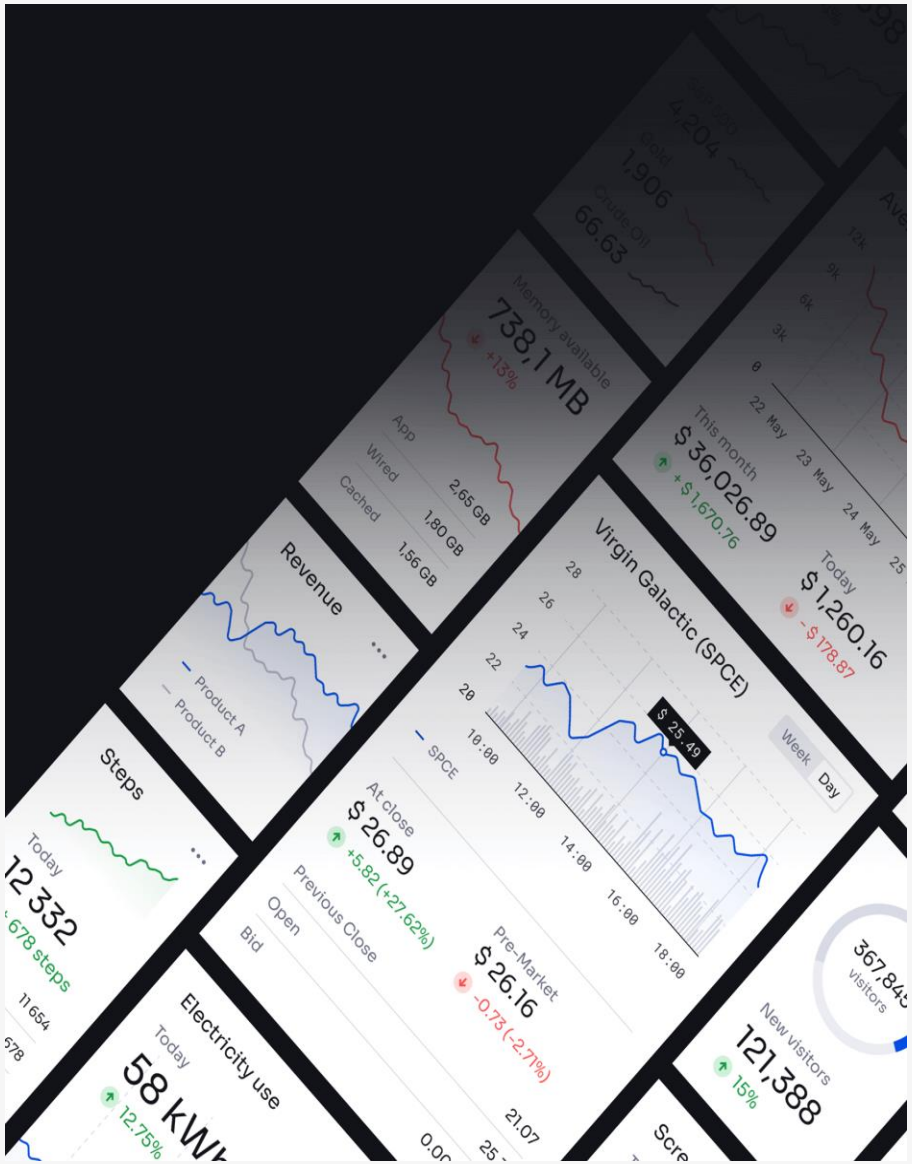
ARPU(ARPPU)
코호트 분석

RFM

군집화

서론
프로젝트 개요 및 배경

온라인 리테일 데이터를 **Cohort, RFM 분석하여 시각화**를 통해 고객 행동을 이해하고, 고객별 세그먼트를 정의한다. 또한 RFM 분석에 토대로 나누어진 **세그먼트를 통해 군집화를 실행**하는 것이 목표이다. 이를 통해 고객 유지 및 매출 증대를 위한 전략을 제시하는 것을 목표로 한다.



서론
분석 과정

데이터 구조 파악
및 전처리

코호트 분석

RFM 분석

Clustering

서론
데이터의 구조

기간 : 2010년 12월 1일 ~ 2011년 12월 9일

InvoiceNo: 송장번호. 각 고유번호 6자리 수. c로 시작하면 취소된 제품이다.

StockCode: 제품 코드. 각 고유번호 5자리 수.

Description: 제품 이름.

Quantity: 거래 당 각 제품에 대한 주문량.

InvoiceDate: 송장 날짜와 시간. (거래가 생성된 날짜와 시간.)

UnitPrice: 단가 (단위: 파운드. 현재 1GBP 당 약 1,800원)

CustomerID: 고객 번호. 각 고유번호 5자리 수

Country: 국가 이름. 주문한 고객이 거주하고 있는 국가의 이름.

서론
데이터의 구조

DataFrame Info

0 InvoiceNo 541909 non-null object

1 StockCode 541909 non-null object

2 Description 540455 non-null object

3 Quantity 541909 non-null int64

4 InvoiceDate 541909 non-null object

5 UnitPrice 541909 non-null float64

6 CustomerID 406829 non-null float64

7 Country 541909 non-null object

Null Data Check.

Description, Customer ID 에서 결측치가 존재하며, 특히 Customer ID 에서 많은 결측치가 보인다.

! 여기서 고객ID의 결측치는 **비회원일 가능성이 있다**는 것을 예상해볼 수 있다.

서론
기본 가공을 통해 알아보아야 할 점

회원 수가 많은 국가가 더 많은 비회원수를 갖는가?

비회원일 수록 주문 취소율이 높은가?

서론
회원 수가 많은 국가가 더 많은 비회원수를 갖는가?

영국 - 프랑스 - 아일랜드 ... 순으로 비회원 수가 많다.
영국에서는 많은 회원 수만큼 비회원 수도 많지만, 나머지 국가를 볼 때 **회원 수와 비회원 수가 정비례하지 않는다.** 특히 홍콩은 영국 다음으로 많은 비회원 수를 가지고 있지만 오로지 비회원으로 이루어진 국가이다.

isRegistered	False	True
Country		
United Kingdom	133600	361878
Germany	0	9495
France	66	8491
EIRE	711	7485
Spain	0	2533
Netherlands	0	2371
Belgium	0	2069
Switzerland	125	1877
Portugal	39	1480
Australia	0	1259
Norway	0	1086
Italy	0	803
Channel Islands	0	758
Finland	0	695
Cyprus	0	622
Sweden	0	462
Austria	0	401
Denmark	0	389
Japan	0	358
Poland	0	341
USA	0	291
Israel	47	250
Unspecified	202	244
Singapore	0	229
Iceland	0	182
Canada	0	151
Greece	0	146
Malta	0	127
United Arab Emirates	0	68
European Community	0	61
RSA	0	58
Lebanon	0	45
Lithuania	0	35
Brazil	0	32
Czech Republic	0	30
Bahrain	2	17
Saudi Arabia	0	10
Hong Kong	288	0

T = 회원
F = 비회원

서론
비회원일수록 주문 취소율이 높은가?

회원 주문취소율과 비회원 주문 취소율 비교

비회원일수록 제품의 신뢰성이 떨어져 주문취소가 높을 것이라 예상했으나, 정반대로 회원일 경우취소율이 높았다.
= **주문취소율은 회원, 비회원과의 관계성이 없다.**

서론

ARPU(ARPPU)
코호트 분석

RFM

군집화

ARPU(ARPPU)&코호트 분석
사전작업

Name: InvoiceDate, **dtype: object**

→ Name: InvoiceDate, **dtype: datetime64[ns]**

InvoiceDate **날짜 형식 변환**

연도, 월 요일 시간별 파생 변수 생성

InvoiceDate	InvoiceYear	InvoiceMonth	InvoiceDay	InvoiceDow
2010-12-01 08:26:00	2010	12	1	2

ARPU(ARPPU)&코호트 분석
ARPU(ARPPU)&MAU

ARPU : 가입한 서비스에 대해 **가입자 1명이 특정 기간 동안 지출한 평균 금액**

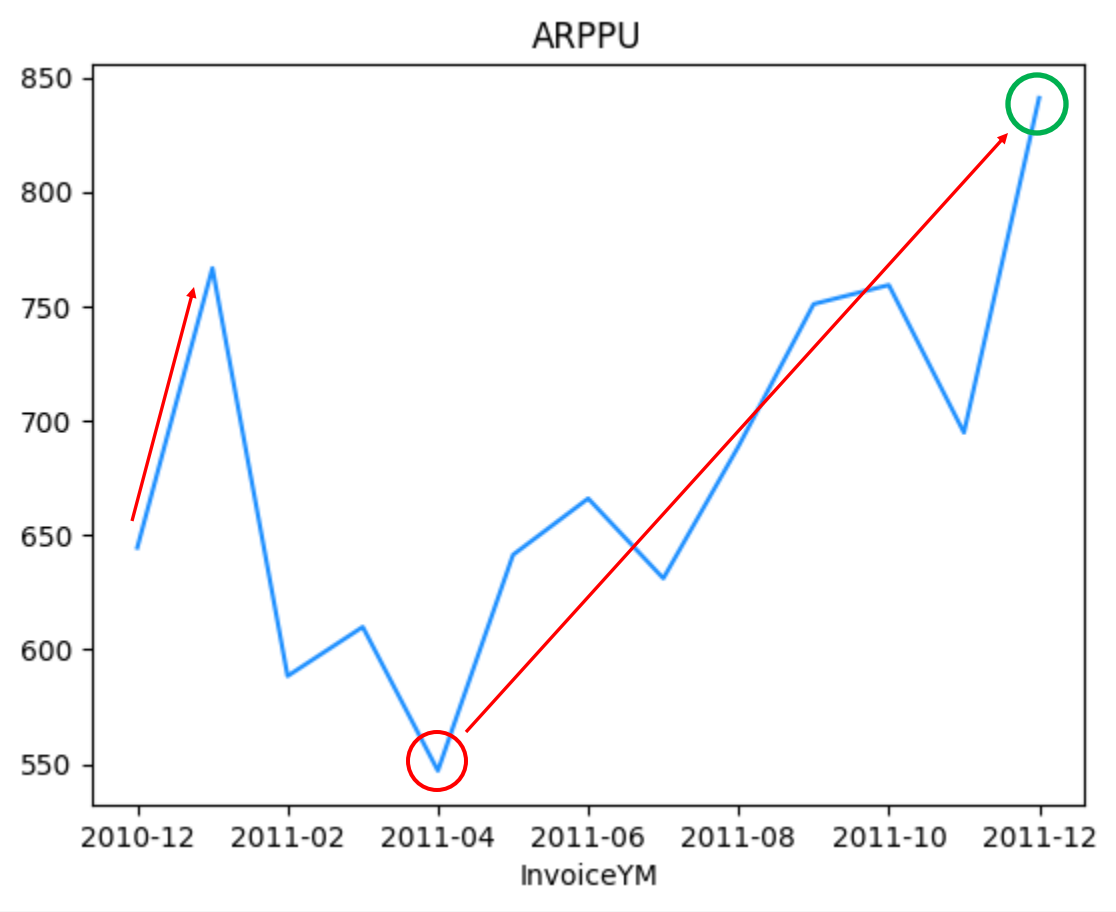
ARPPU : 유저 **1명 당 한 달에 결제하는 평균 금액**을 산정한 수치

MAU : **30일 동안** 앱 또는 웹사이트에서 **활동하는 순 유저 수**
(DAU 일일 활성 유저)

ARPU(ARPPU)&코호트 분석
ARPU(ARPPU)

2010년 12월에서 2011년 1월 까지 크게 증가
이후 2월부터 급격히 감소하고 4월 까지 지속적 하락
4월부터 성장과 하락을 반복하지만 연말로 갈 수록 ARPPU 증가

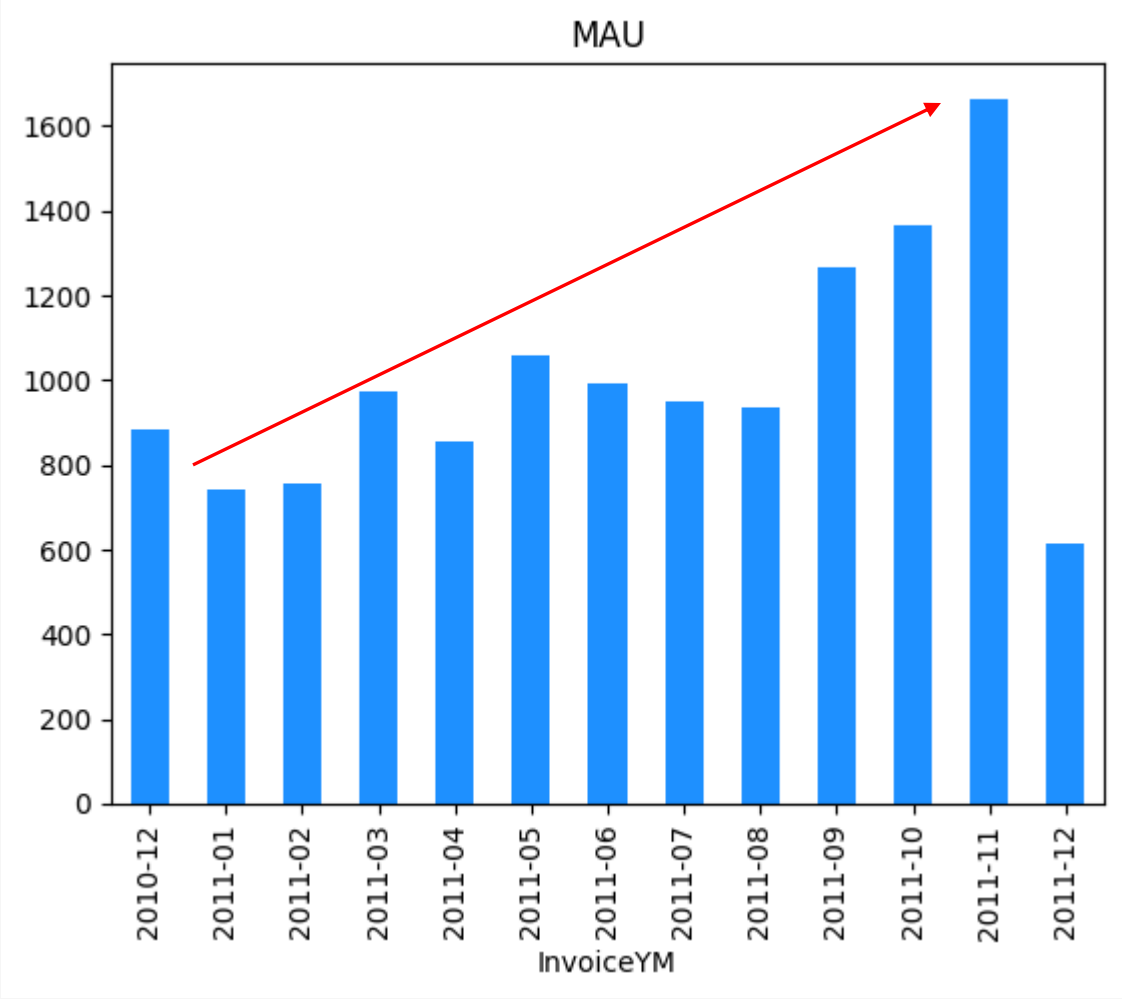
초, 중반기에 큰 하락세이고 **연말에 강한 케이스**이다.
연말에 큰 이벤트나 프로모션을 가졌을 확률이 높다.
초, 중반기에는 이벤트나 프로모션을 통해 소비를 촉진할 필요가 있다.



ARPU(ARPPU)&코호트 분석
MAU

ARPPU와 비슷한 추세를 가지고 있다.
2010년 12월에서 1월로 넘어갈 때 감소가 있지만 11월을 갈수록 계속 해서 상승, 최고점을 찍는다.

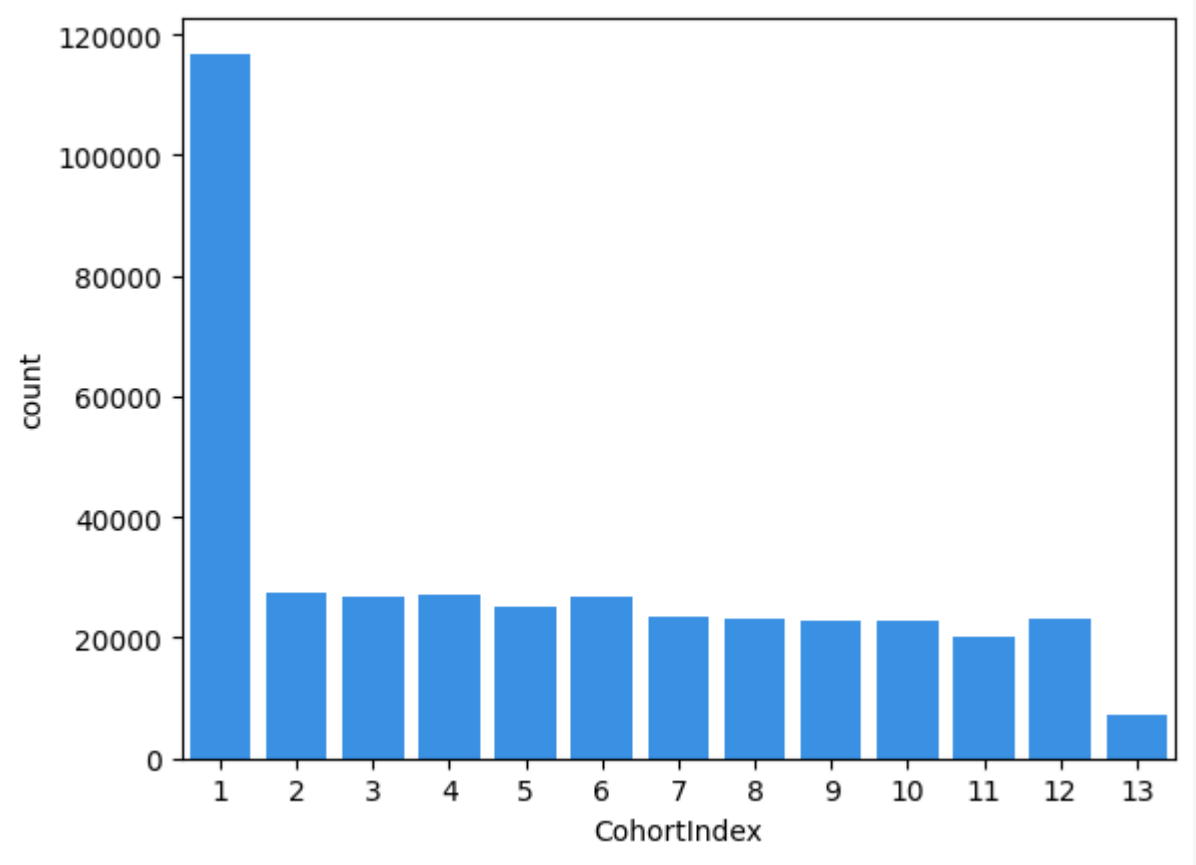
월 활성유저와 인당 월 평균 소비 금액이 비슷하게 움직인다.
매출 증대를 위해서는 **월 활성유저를 더 유입시키는 것을 목표로** 해야한다.



ARPU(ARPPU)&코호트 분석
재구매율은 얼마나 될까?

코호트 인덱스 : 처음 구매 후 얼마 뒤에 재구매가 이루어졌는가?

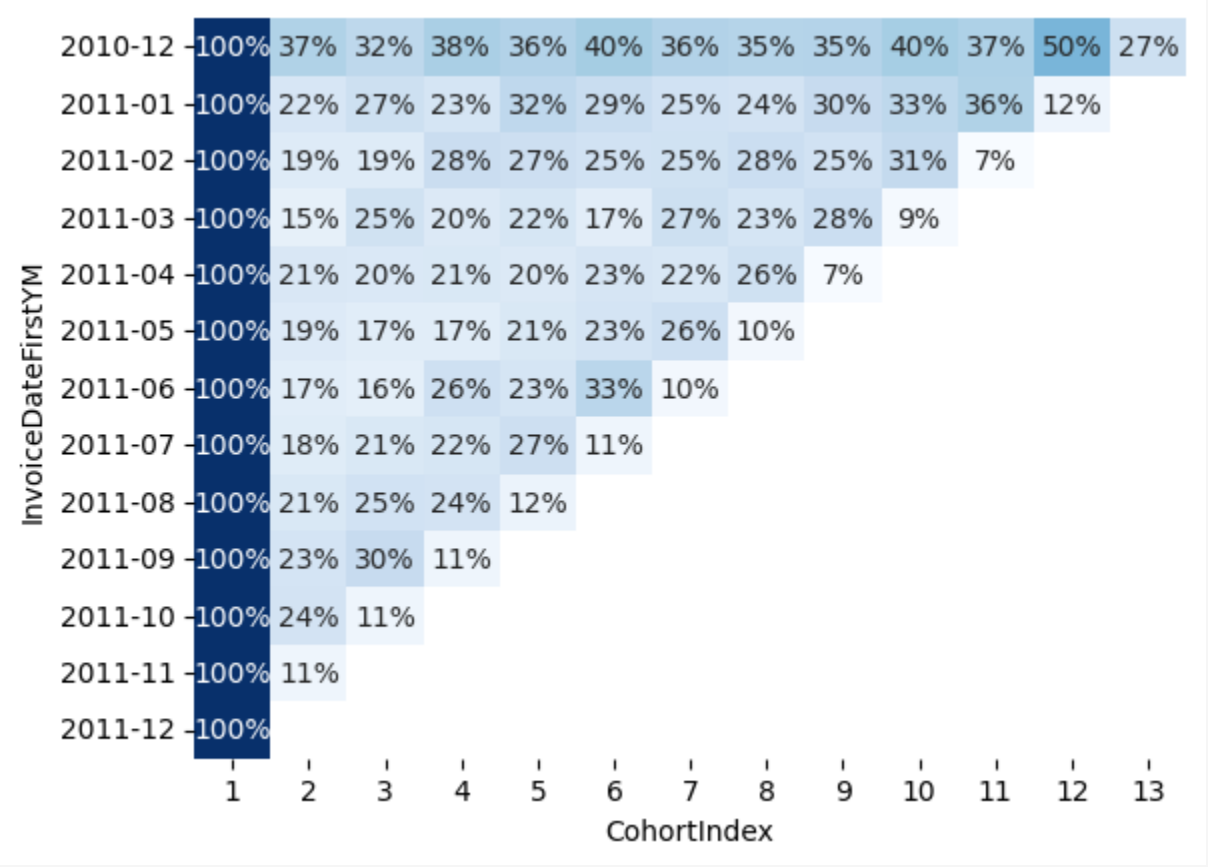
첫 번째 구매 후 재구매가 이루어지지 않는다고 볼 수 있다.
= **충성 고객의 수가 매우 적다.**
재구매를 유도할 수 있으면 전체적인 재구매율이 높아질 것이라 예상한다.



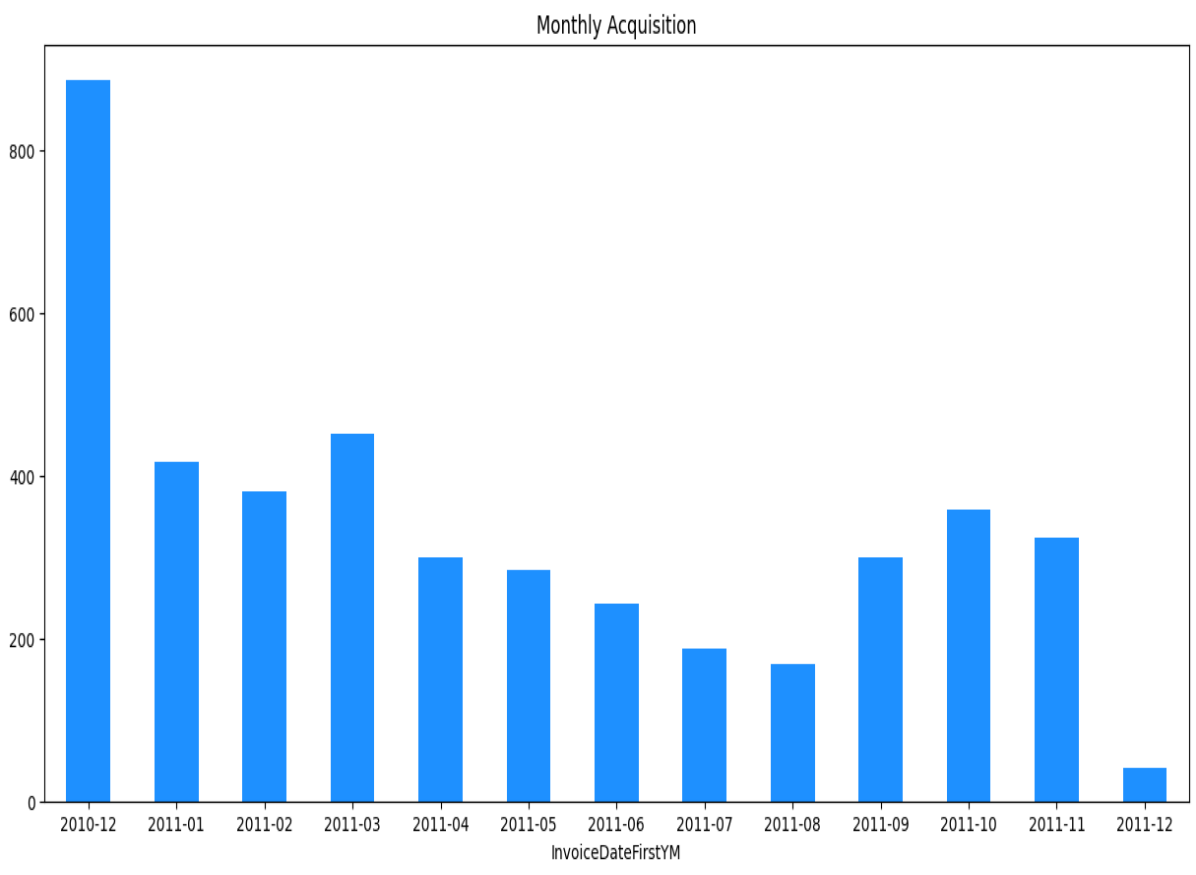
ARPU(ARPPU)&코호트 분석
잔존빈도율

특정 기간 동안 첫 구매를 한 고객들의 구매 유지율을 시각화.

지속적인 감소를 그리고 있는 패턴을 가지고 있다.
특이점은 **특정 달에서 유지율이 상승하는 패턴**이 있다.
2010년 12월 코호트에서 10월차에서 11개월 차에 23% 상승이 보이고,
2011년 6월 코호트는 5개월차에 23% 지만 6개월 차에 33%로 상승하는 것이 보인다.



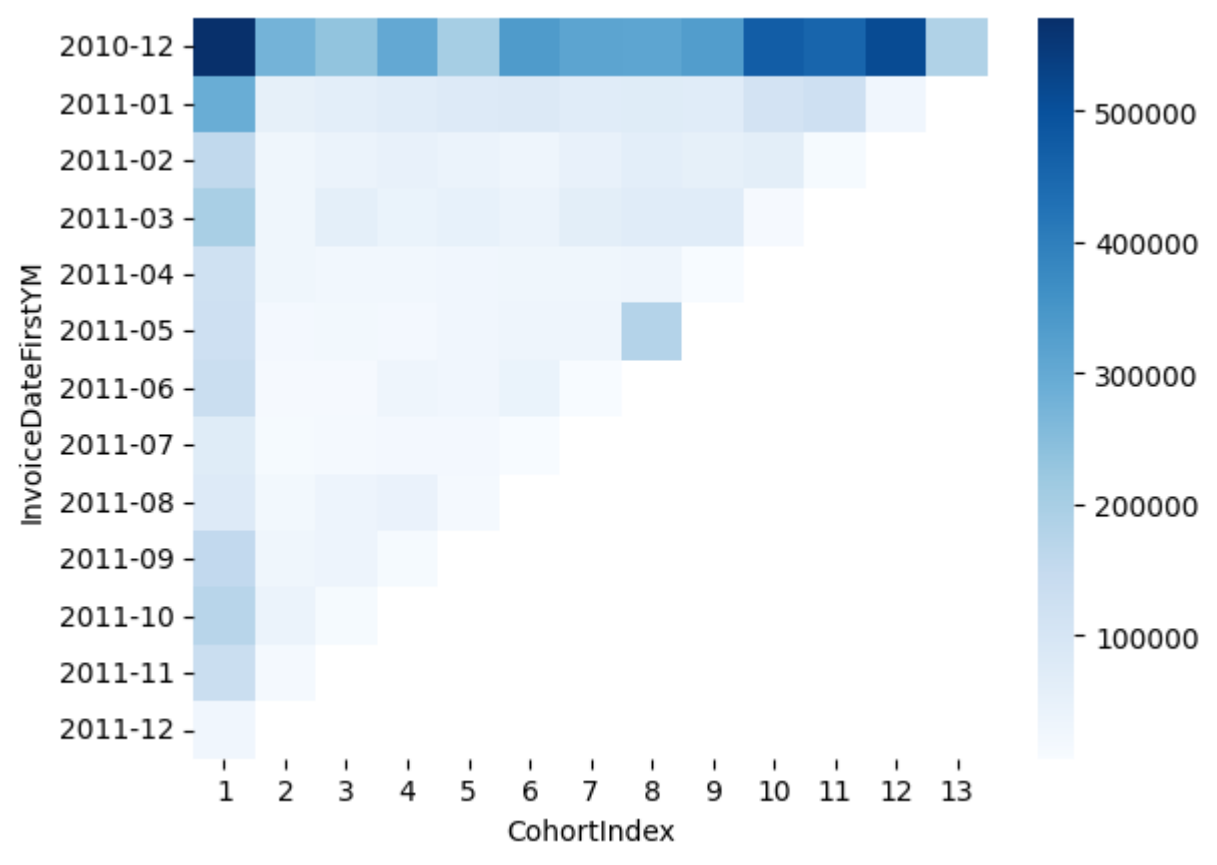
ARPU(ARPPU)&코호트 분석
Acquisition(신규 유입 고객 수)



월별 신규 유입 고객 수

2010년 12월에 매우 높은 신규 고객 획득을 기록한 후, 2011년 대부분 동안 **고객 획득이 감소하는 추세**를 보이고 있습니다. 이는 기업이 지속적인 성장을 위해 **계절적 요인과 시장 변화에 맞춘 효과적인 마케팅 전략을 개발하고, 특정 시기에 집중된 마케팅 활동이 이후의 고객 유지와 추가 고객 획득으로 이어지도록 할 필요**가 있다.

ARPU(ARPPU)&코호트 분석
월별 매출 리텐션



월별 매출 리텐션
2010년 12월 첫 구매 고객들은 9 ~ 11개월 차에도 많은 구매를 해주었다.
= **충성 고객일 확률 ↑**

2011년 5월 고객들의 7개월 차의 매출이 유독 높다.
= **이벤트나 프로모션을 했을 것으로 예상**된다.

충성 고객을 어떻게 유지할 것인가?
유독 매출이 많던 달에 어떠한 이벤트, 프로모션을 진행했고, 새로운 방법은 무엇이 있을까?
(= 유저가 선호하는 이벤트나 프로모션)
에 대한 방법을 구상해야 할 필요가 있다.

서론

ARPU(ARPPU)
코호트 분석

RFM

군집화

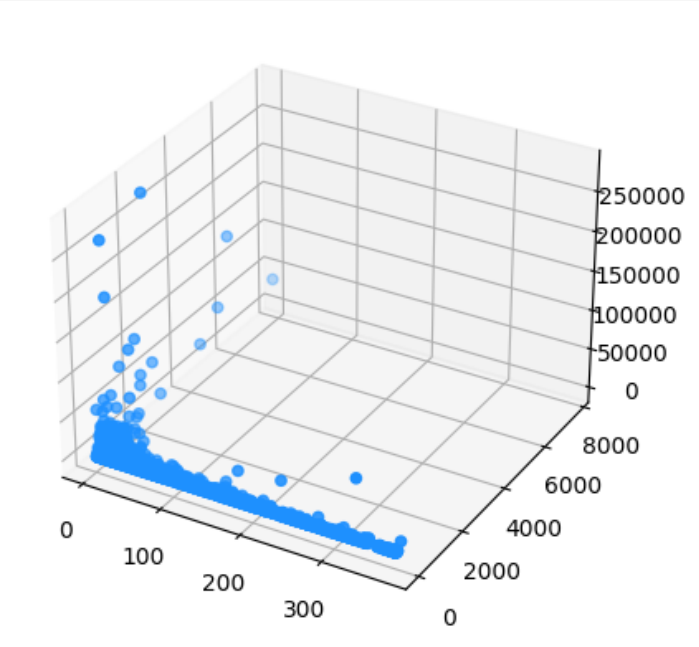
RFM
사전작업

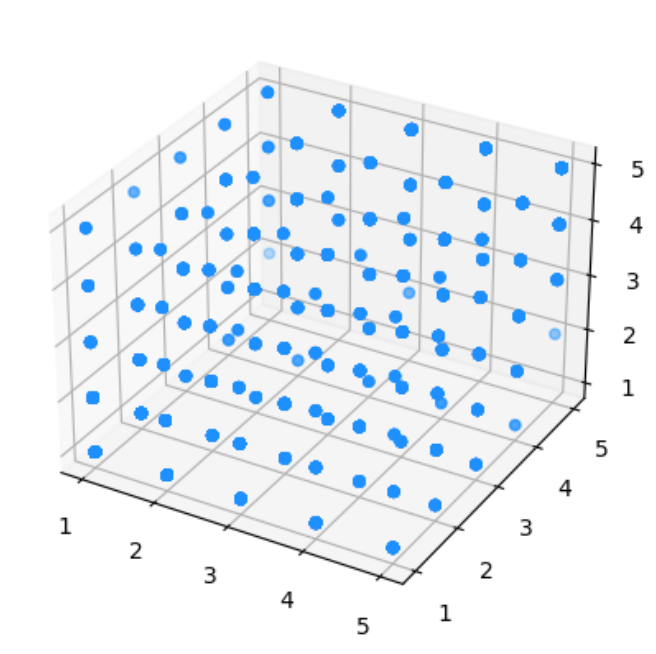
점수 부여
Recency : **최근일수록 점수를 높게** 부여
Frequency : **구매 빈도수가 높을수록 점수를 높게** 부여
Monetary : **구매 금액이 많을수록 점수를 높게** 부여

세그먼트
데이터를 특정 기준에 따라 그룹화

그룹화 된 데이터들의 **합계를 점수화**

RFM
점수 부여, 그룹화로 인한 변화

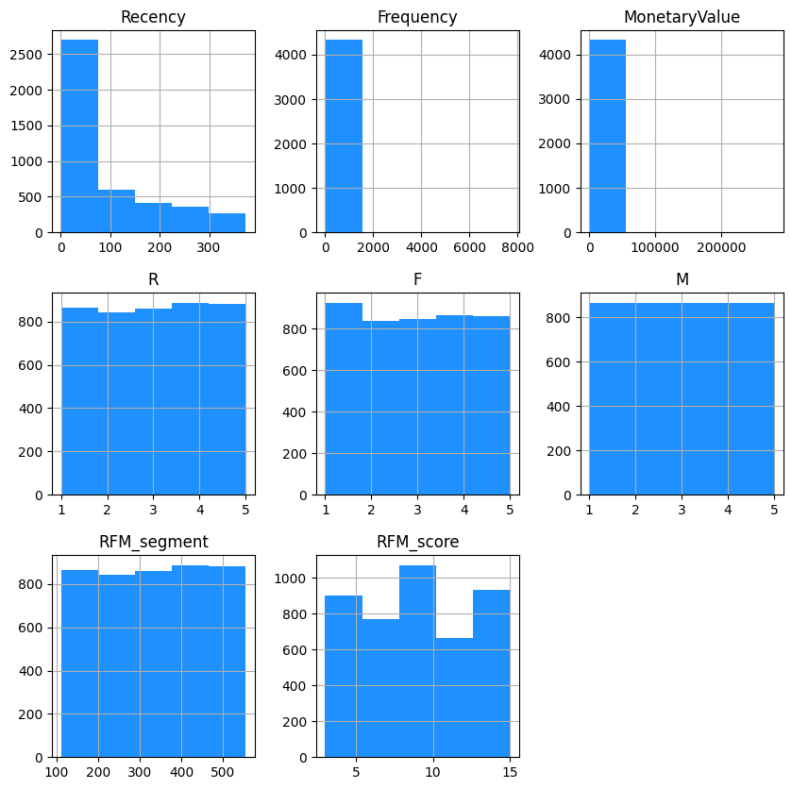




특정 기준으로 점수화 하지 않은 자료(좌)에서는 이상치 값들이 존재하였지만 점수화 한 자료(우)에서는 고루 분포되었음을 볼 수 있다.

= **이상치 값 처리**

RFM
R, F, M 히스토그램과 각 세그먼트의 평균과 총 지출

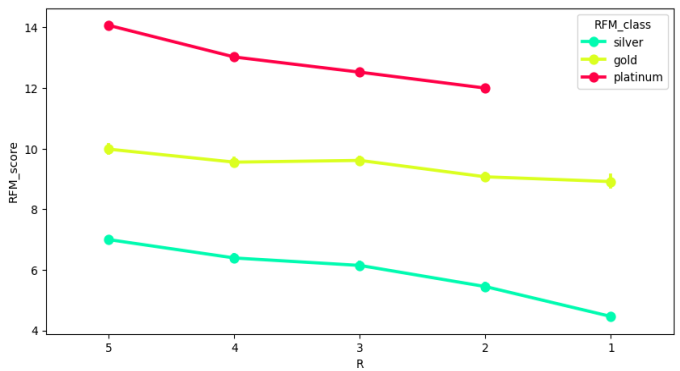


<히스토그램>

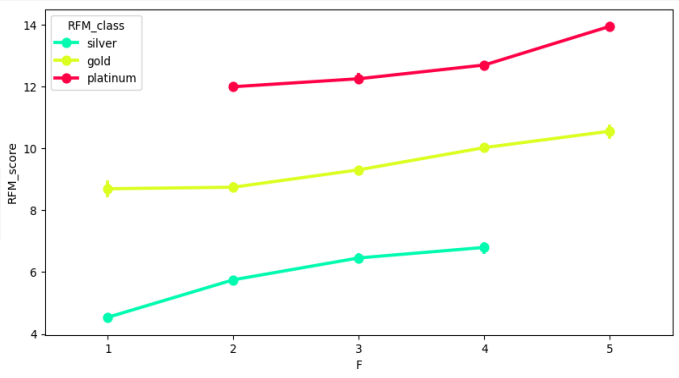
RFM_score	Recency mean	Frequency mean	MonetaryValue mean	MonetaryValue sum
3	278	7	138	37,309
4	204	11	200	52,138
5	183	16	295	108,729
6	126	20	371	142,014
7	103	26	898	345,746
8	87	36	628	227,483
9	70	46	858	309,676
10	59	62	1,123	392,016
11	45	80	1,445	487,016
12	35	108	1,794	592,058
13	23	140	3,080	973,379
14	16	230	4,797	1,467,897
15	5	439	11,596	3,583,278

<평균과 총 지출>

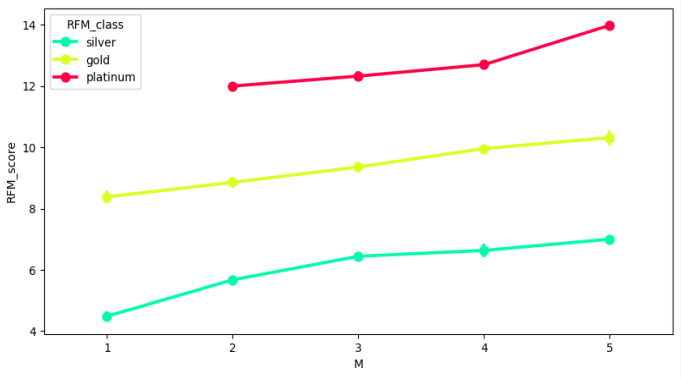
RFM
q-cut을 이용한 3단계 고객 군집화



Recency

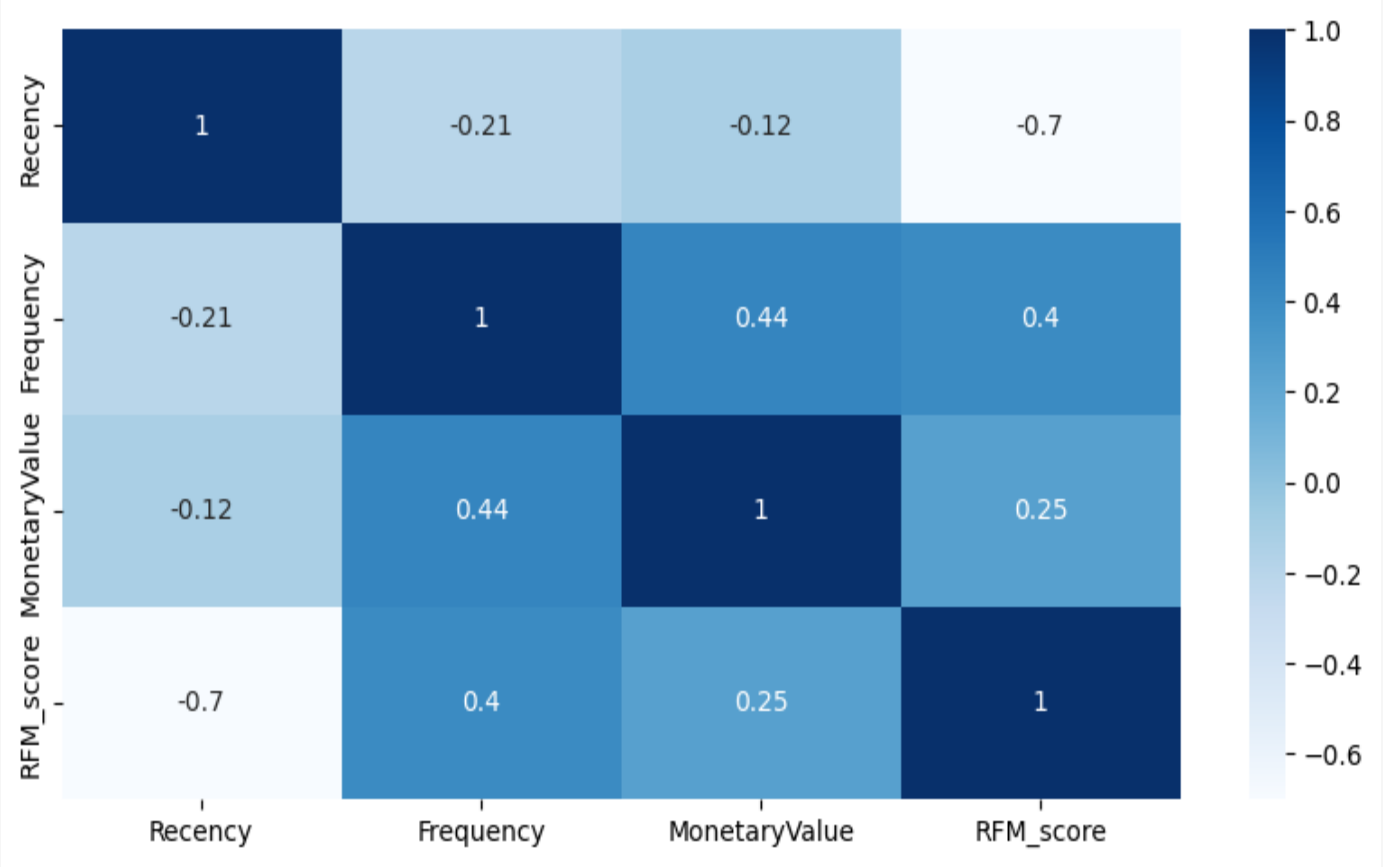


Frequency



Monetary

RFM
변수 간 상관관계 확인(선형적 관계)



Recency의 경우 최근일수록 점수가 더 높기 때문에 음의 **상관관계**를 가지고 Frequency와 Monetary는 **양의 상관관계**를 가진다.

이 시각화의 목적은 고객들이 어느 한 곳에 몰려있거나 한 눈에 고객층이 어떻게 이루어져 있는지 알기 힘들 때 **같은 간격(Segement) 로 나누어 좀 더 쉽게 볼 수 있기 위함**에 있다.

서론

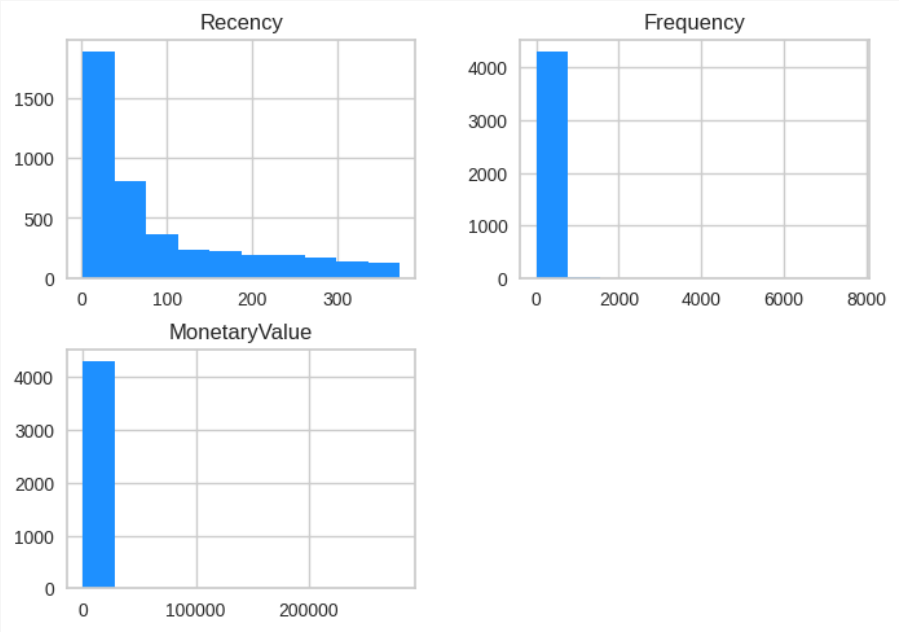
ARPU(ARPPU)
코호트 분석

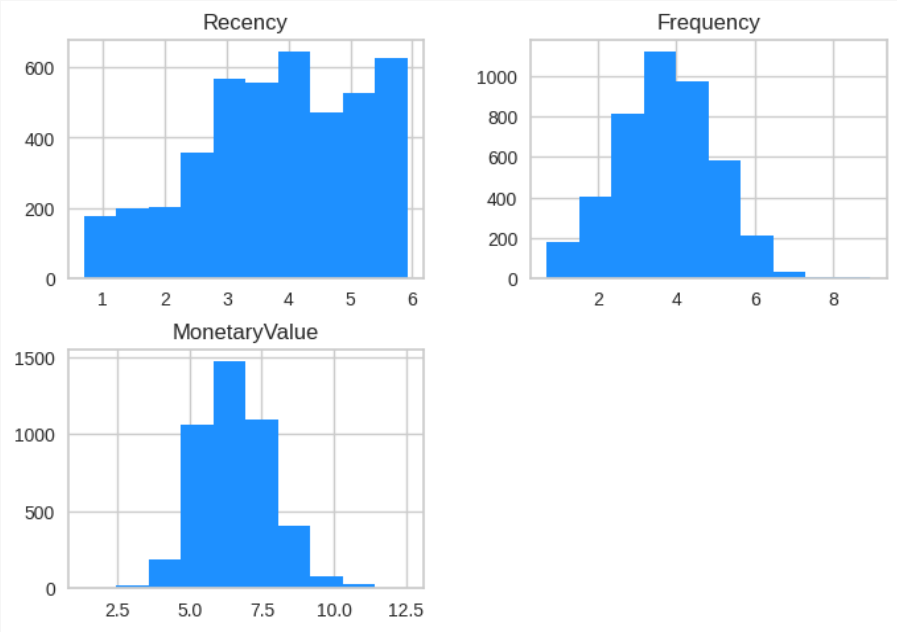
RFM

군집화

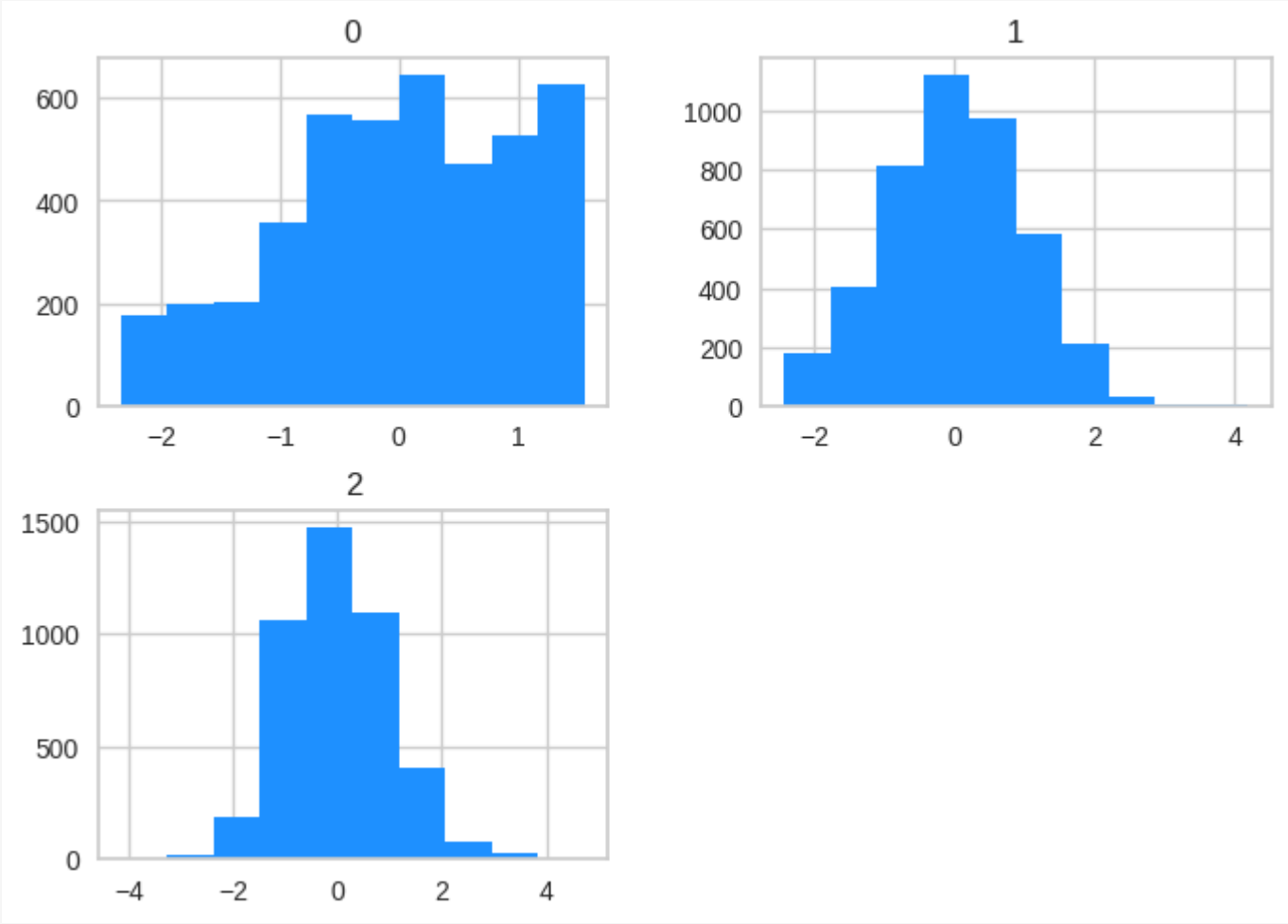
군집화
사전작업

Recency, Frequency, Monetary **이상치 처리**
로그변환 : 머신러닝 모델이 잘 이해할 수 있도록 **정규분포 형태로 변환**
작은 값에 대한 변돌을 더 잘 반영한다.(균형있게 다룬다.)





군집화
스케일 조정(Scaling)

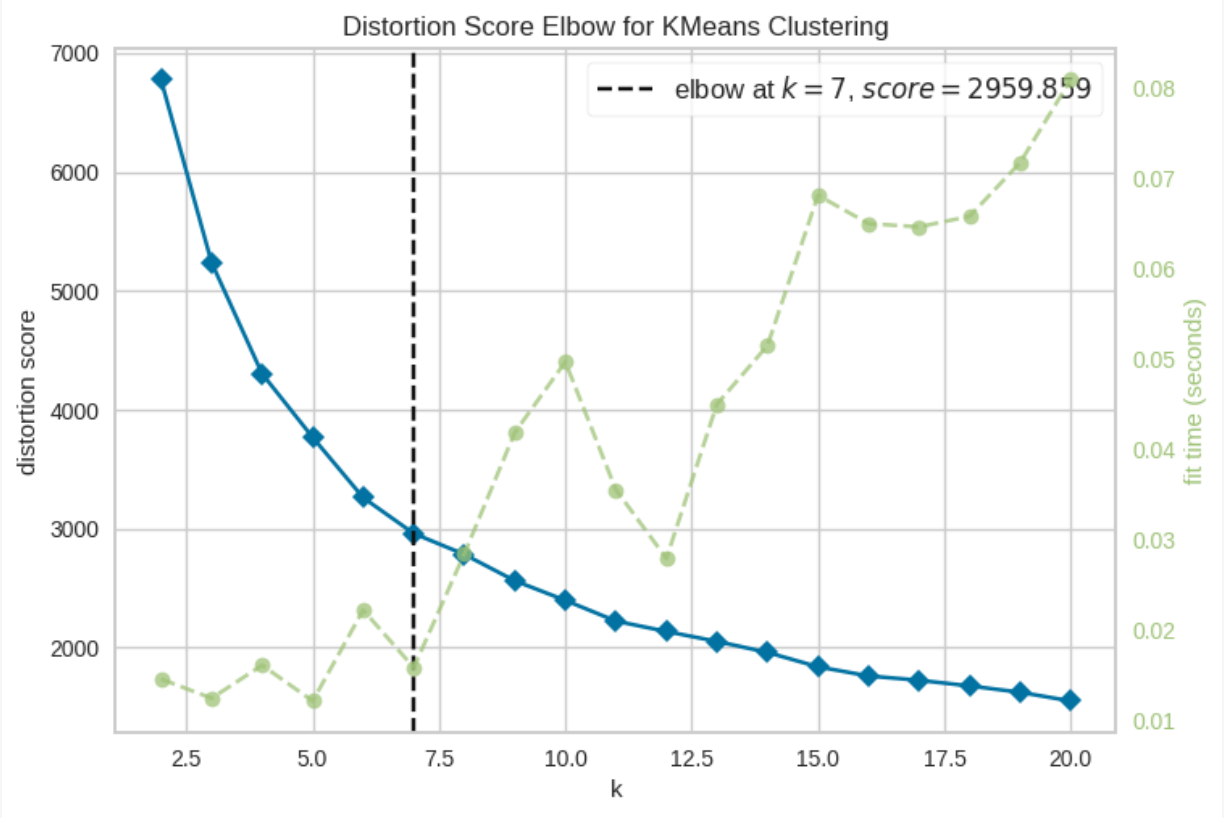


StandardScaler 를 사용 > 평균 0, 분산 1로 조정

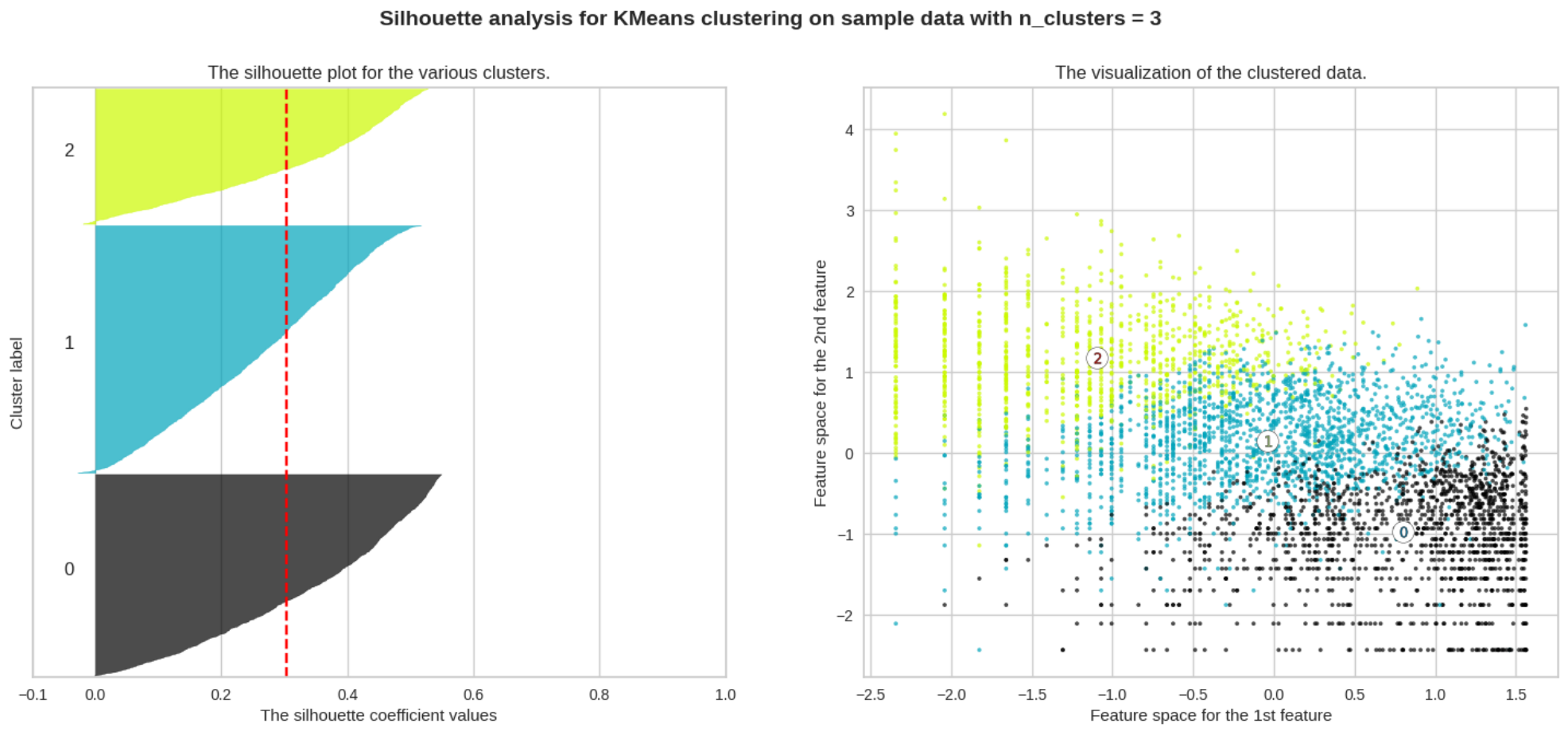
군집화
K-means, Kelbow Visualizer

K-means
: 샘플을 n개의 등분산 그룹으로 분리하여 **관성 또는 클러스터 내 제곱합 이라는 기준을 최소화** 함으로써 데이터를 클러스터링 한다.

K = 7 일 때, 즉 클러스터 수가 7개일 때 **가장 최적화로 군집화를 실행할** 수 있으나, **현 프로젝트에서는 고객 집단을 3개의 군**으로 나누었기 때문에 클러스터 수를 3개로 진행할 예정이다.



군집화
군집화 시각화


Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
4
3
2
1
0
−1
−2
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Feature space for the 1st feature

군집화
시각화 해석

실루엣 플롯(좌)
클러스터 0과 1은 비교적 실루엣 계수가 높은 반면, 클러스터 2는 실루엣 계수가 상대적으로 낮다.

평균 실루엣 계수는 약 0.4 정도로 조금 아쉽다.

군집화 시각화(우)
색상에 따라 잘 구분되어 있다. 하지만 실루엣 플롯을 보았듯이 클러스터 2에서 다른 클러스터와 겹치는 부분이 조금 보인다.

전체적으로 구분은 잘되었으나,
클러스터 2의 실루엣 계수가 낮고, 평균 실루엣 계수도 0.5를 넘기지 못하였다.
최적의 클러스터를 다시 찾고 고객 구분을 세부적으로 좀 더 나누는 것도 생각하여 좀 더 완벽한 군집화를 이룰 수 있도록 해야겠다.

Plus

온라인 리테일 전체 보기

Caggle 전체 보기

티스토리 전체 보기