다변량데이터분석

Association Rule Mining

and

Clustering

과제 5

산업경영공학부

2019170815

신건우

[Part 1: Association Rule Mining]

Dataset: MOOC Dataset (big_student_clear_third_version.csv)

해당 데이터셋은 MOOC 강좌를 수강한 수강생들에 대한 정보가 포함되어 있는 데이터 셋이다. 다음 각 Instruction에 따라 데이터를 변환하고 연관규칙분석을 수행하여 각 결과물을 제시하고 적절한 해석을 제공하시오.

[Step 1] 데이터 변환

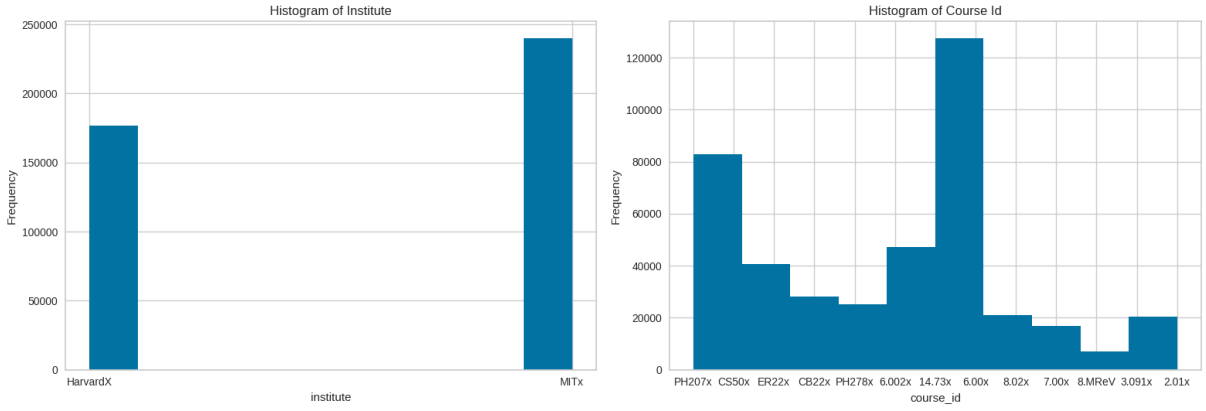
[Q1] 원 데이터는 총 416,921건의 관측치와 22개의 변수가 존재하는 데이터프레임이다. 이 중에서 아래 그림과 같이 userid_DI (사용자 아이디)를 Transaction ID로 하고, institute (강좌 제공 기관), course_id (강좌코드), final_cc_cname_DI (접속 국가), LoE_DI (학위 과정)을 하나의 string으로 결합하여 Item Name으로 사용하는 연관규칙 분석용 데이터셋을 만드시오.

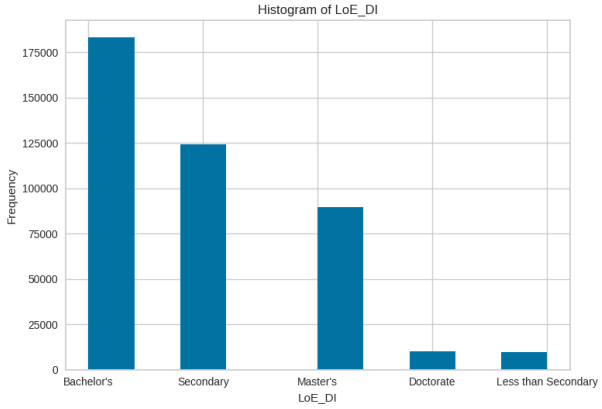
[Step 2] 데이터 불러오기 및 기초 통계량 확인

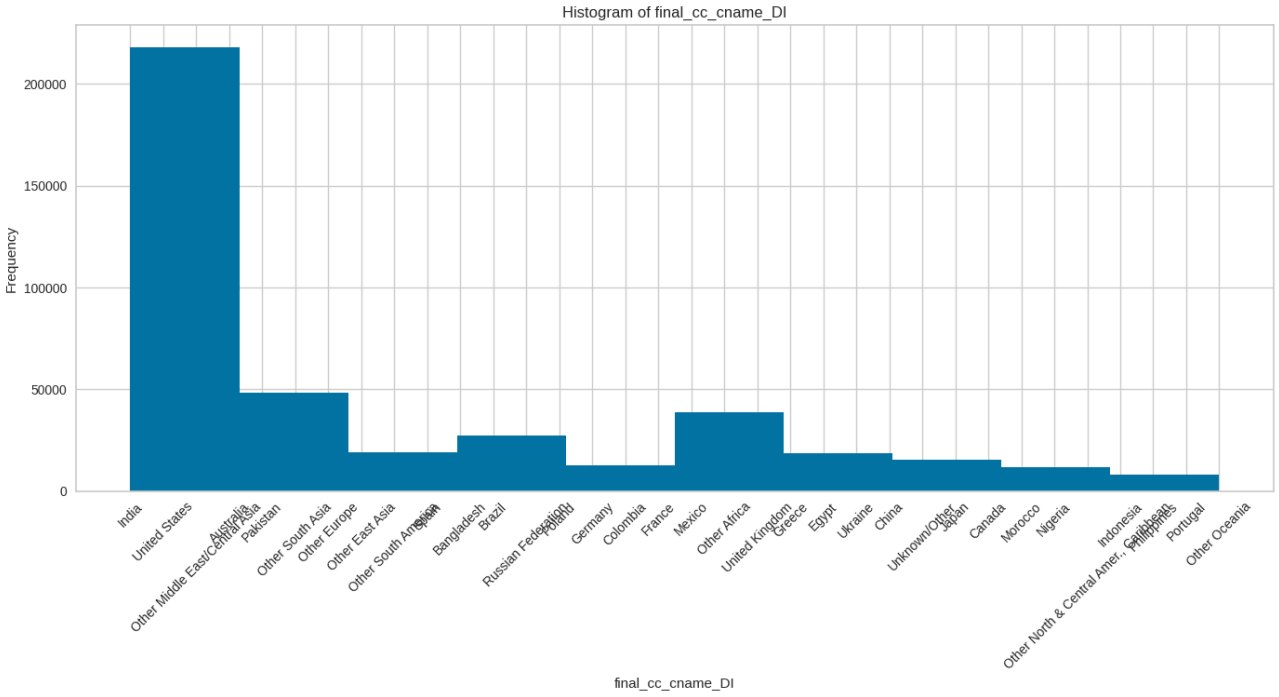
[Q2-1] [Q1]에서 생성된 데이터를 읽어들이고 해당 데이터에 대한 탐색적 데이터 분석을 수행하여 데이터의 특징을 파악해보시오.

	Transaction ID	Item name
0	MHxPC130000002	MITx_14.73x_United Kingdom_Secondary
1	MHxPC130000004	HarvardX_CS50x_India_Secondary,HarvardX_ER22x_...
2	MHxPC130000006	HarvardX_ER22x_United States_Bachelor's
3	MHxPC130000007	HarvardX_CB22x_United States_Master's
4	MHxPC130000008	MITx_6.00x_United Kingdom_Bachelor's

약 335650의 transaction ID(사용자)가 존재하며, 복수 강좌를 수강한 사람도 존재한다. 다음은 각 Item name에 사용된 institute, course_id, final_cc_cname_DI, LoE_DI의 분포를 나타낸 히스토그램이다.







특징

MITx의 수강생 수가 HarvardX보다 더 많다. MITx의 온라인 교육 프로그램이 더 인기가 있는 것으로 보인다.

특정 강의가 다른 코스에 비해 현저하게 많은 수강생을 보유하고 있다.

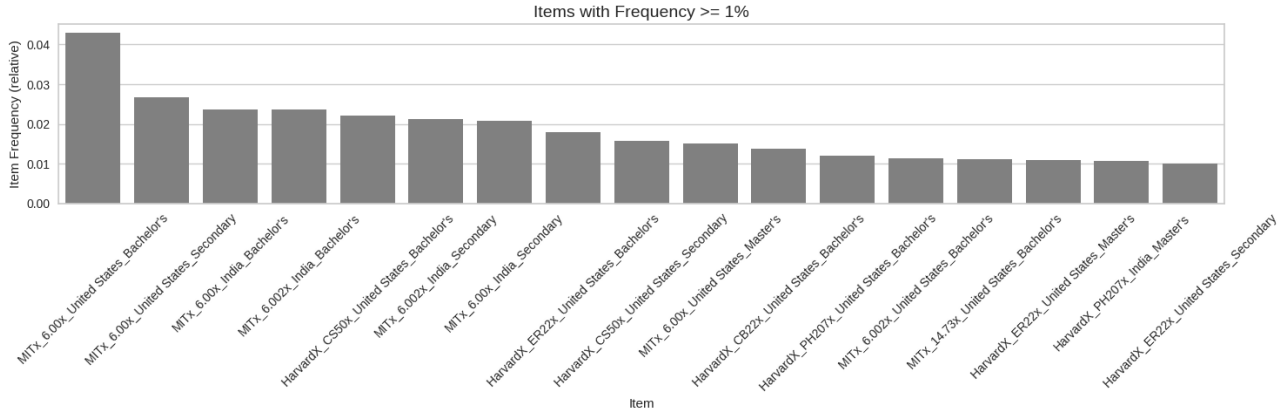
인도와 미국이 다른 국가에 비해 압도적으로 많은 수강생을 보유하고 있다.

학사 학위(Bachelor's)를 가진 수강생이 가장 많으며, 그 다음으로 고등학교 졸업(Secondary), 석사 학위(Master's)가 많다. 학력 수준이 높은 수강생들이 온라인 교육에 많이 참여하고 있음을 알 수 있다.

final_cc_cname_DI에 문자열 ‘,’가 존재하므로 각 아이템 구분자로 ‘,’가 아닌 ‘@’를 사용하여 잘못된 전처리를 피했다.

[Q2-2] 아이템 이름과 아이템 카운트를 이용하여 워드클라우드를 생성해 보시오.

[Q2-3] 최소 빈도 1% 이상 등장한 Items들의 Bar Chart를 도시하시오. 상위 5개의 Item에 대해 접속 국가는 각각 어느 국가인지 확인하시오.



상위 5개 아이템의 접속 국가는 'United States' 'United States' 'India' 'India' 'United States'이다.

앞에도 살펴봤듯, 미국과 인도의 수강생이 많음을 시사한다.

[Step 3] 규칙 생성 및 결과 해석

[Q3-1] 최소 10개 이상의 규칙이 생성될 수 있도록 support와 confidence의 값을 조정해 가면서 각 support-confidence 조합에 대해 총 몇 가지의 규칙이 생성되는지 확인하고 그 결과를 아래 표와 같은 형태로 제시하시오. 최소 한 3개 이상의 support, 3개 이상의 confidence, 총 9개 이상의 조합에 대한 규칙 생성을 수행하시오.

	Confidence 0.05	0.1	0.2
Support 0.001	51	34	10
0.002	20	16	5
0.003	6	5	0

최소 지지도가 높을수록, 규칙의 수가 적어지는 것을 확인할 수 있는데, 이는 정의상 당연한 현상이다.

[Q3-2] support = 0.001, confidence = 0.05로 지정하여 생성된 연관규칙분석들에 대해 다음 질문에 대한 답과 본인의 생각을 서술하시오.

Support가 가장 높은 규칙은 무엇인가?

Antecedents : (MITx_6.00x_United States_Bachelor's)

Consequents : (HarvardX_CS50x_United States_Bachelor's)

MITx의 6.00x 강좌와 HarvardX의 CS50x 강좌 사이에 강한 연관성이 있음을 나타내며, 두 강좌 모두 미국의 학사 학위 소지자들 사이에서 인기가 있다는 것을 보여준다. 이는 두 강좌가 서로 보완적인 내용이 있거나, 동일한 수강생 집단이 흥미를 가질 가능성이 높다는 것을 시사할 수 있다.

Confidence가 가장 높은 규칙은 무엇인가?

Antecedents : (MITx_8.02x_India_Secondary)

Consequents : (MITx_6.002x_India_Secondary)

두 강좌가 동일한 인도 고등학교 졸업자 집단에게 강한 흥미를 끌고 있음을 나타낸다. 두 강좌가 내용적으로 연관이 있거나, 학습 커리큘럼 상에서 순차적으로 듣는 것이 일반적일 가능성이 있다. 예를 들어, 8.02x 강좌를 듣고 난 후 6.002x 강좌를 수강하는 것이 자연스러운 학습 흐름일 수 있다.

Lift가 가장 높은 규칙은 무엇인가?

Antecedents : (MITx_6.002x_United States_Bachelor's)

Consequents : (MITx_8.02x_United States_Bachelor's)

높은 lift 값은 두 강좌 간의 강한 연관성을 나타내며, MITx의 6.002x 강좌를 듣는 학생들이 8.02x 강좌를 함께 듣는 경향이 강하다는 것을 보여준다. 이는 두 강좌가 학습 커리큘럼 내에서 서로 밀접하게 연결되어 있거나, 6.002x 강좌를 듣고 난 후 8.02x 강좌를 수강하는 것이 일반적이라는 것을 시사한다.

만일 하나의 규칙에 대한 효용성 지표를 Support × Confidence × Lift로 정의한다면 효용성이 가장 높은 규칙 1위~3위는 어떤 것들인가?

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	utility
(MITx_8.02x_India_Secondary)	(MITx_6.002x_India_Secondary)	0.007	0.020	0.002	0.388	19.011	0.003	1.601	0.954	0.020
(MITx_8.02x_India_Bachelor's)	(MITx_6.002x_India_Bachelor's)	0.006	0.023	0.002	0.386	16.958	0.002	1.591	0.947	0.016
(HarvardX_CS50x_India_Secondary)	(MITx_6.00x_India_Secondary)	0.009	0.020	0.003	0.294	14.386	0.002	1.387	0.939	0.011

해석

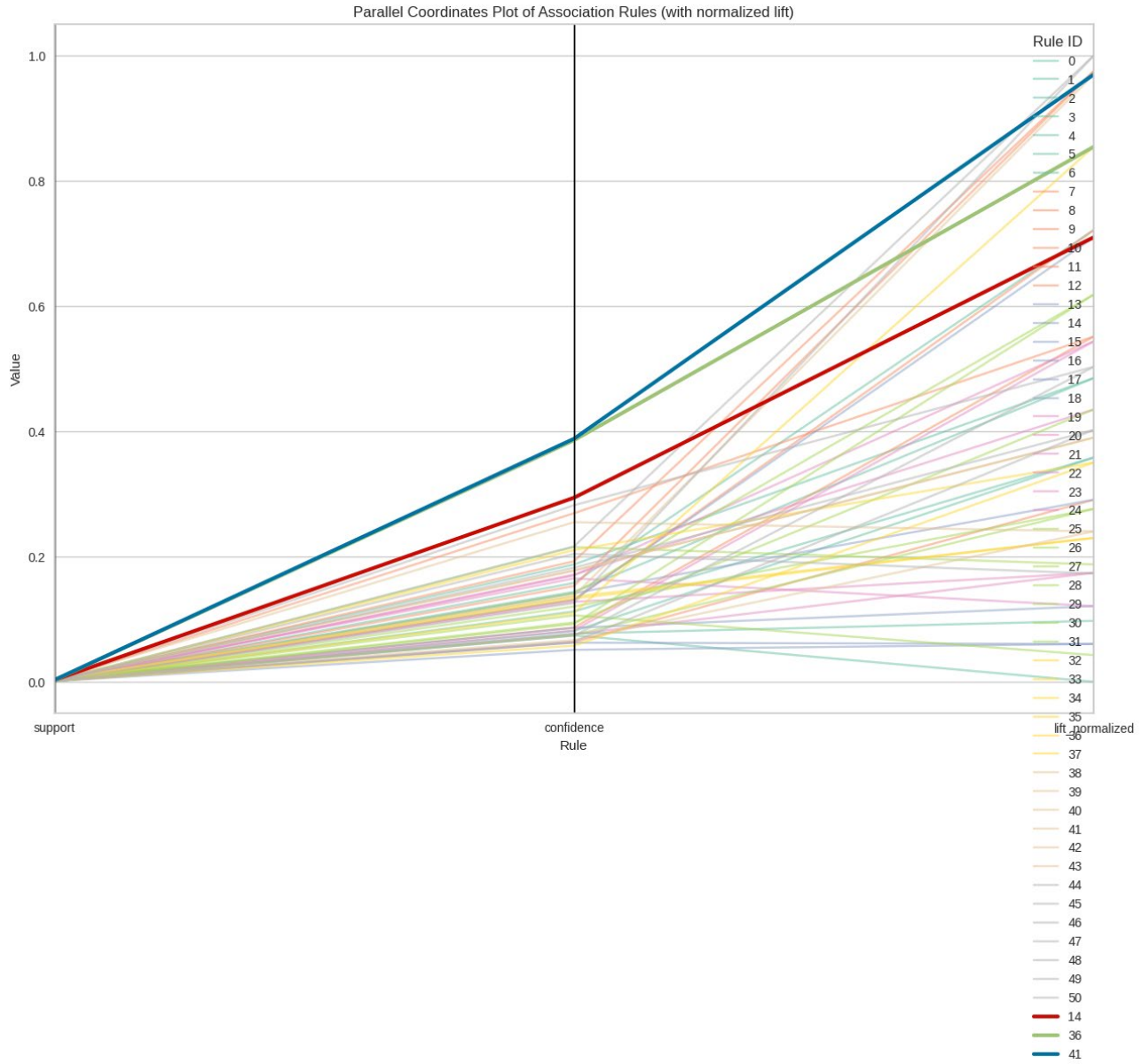
- **1위 규칙**: (MITx_8.02x_India_Secondary) → (MITx_6.002x_India_Secondary)
    - 이 규칙은 효용성 지표가 가장 높다. 이는 인도의 고등학교 졸업자가 MITx의 8.02x 강좌를 수강한

후 MITx의 6.002x 강좌를 수강할 확률이 매우 높음을 나타낸다. 두 강좌의 내용이 서로 보완적이거나 연계된 커리큘럼의 일환일 가능성이 크다.

- **2위 규칙**: (MITx_8.02x_India_Bachelor's) → (MITx_6.002x_India_Bachelor's)
    - 이 규칙은 인도의 학사 학위 소지자들 사이에서 유사한 패턴을 보인다. MITx의 8.02x 강좌를 수강한 인도의 학사 학위 소지자가 6.002x 강좌를 수강하는 경향이 강하다. 이는 위에서 언급한 **Confidence가 가장 높은 규칙** 고등학교 졸업자와 유사한 학습 경향을 보여준다.
- **3위 규칙**: (HarvardX_CS50x_India_Secondary) → (MITx_6.00x_India_Secondary)
    - 이 규칙은 인도의 고등학교 졸업자가 HarvardX의 CS50x 강좌를 수강한 후 MITx의 6.00x 강좌를 수강할 확률이 높음을 나타낸다. 이는 두 강좌가 비슷한 주제를 다루거나, 동일한 학습 집단에게 인기 있는 강좌임을 시사한다.

[Extra Question] 이 외 수업 및 실습 시간에 다루지 않은 연관규칙분석 시각화 및 해석을 시도해 보시오.

Pandas의 parallel_coordinates로 여러 메트릭을 동시에 비교할 수 있다. 규칙 간의 지원도, 신뢰도, 향상도를 한눈에 비교하여 중요한 규칙을 식별할 수 있다.



[Part 2: Clustering]

Dataset: Kaggle Clustering 데이터셋 중 1개 선택

Kaggle 사이트의 Datasets 항목에서 “clustering”을 키워드로 검색하면 총 1,438개의 데이터셋이 아래와 같이 검색됩니다(2024-05-26 기준).

링크: Kaggle Clustering Datasets

[Q1] 데이터셋 선정하기

이 중에서 군집화 후 각 군집에 대한 속성 분석이 유의미할(또는 재미있을) 것으로 판단되는 데이터셋 하나를 선정하고

본인이 해당 데이터셋을 선정한 이유를 설명하시오.

https://www.kaggle.com/datasets/dgawlik/nyse

New York Stock Exchange 데이터 셋을 선정했다.

교수님께서 수업시간에 언급한 군집화를 통한 포트폴리오 구성이 유의미한 효과를 보이는지 궁금해졌다. 관련 연구는 다음과 같다. 신승민. (2014). *횡단면적 군집분석을 이용한 주식투자 전략: KOSPI 200 주식을 중심으로* (Doctoral dissertation, 서울대학교 대학원). 해당 연구에서는 군집화에 재무지표를 사용했다. 총 11개의 재무지표를 규모, 수익성, 효율성 지표, 레버리지 지표, 유동성지표로 분류하여 군집화 결과를 해석했다.

해당 연구에 사용된 방식을 비슷하게나마 NYSE 거래소에 상장된 주식을 대상으로 수행해보고자 한다.

각 데이터셋은 다음과 같다.

prices.csv:

- 원본 일일 주가 데이터.
- 대부분의 데이터는 2010년부터 2016년 말까지의 기간을 포함하며, 주식 시장에 새로 상장된 회사의 경우 데이터 범위가 더 짧다.
- 해당 기간 동안 약 140건의 주식 분할이 있었으나, 이 데이터 세트에는 그러한 조정이 반영되지 않았다.

prices-split-adjusted.csv:

- prices.csv와 동일한 데이터이지만, 주식 분할에 대한 조정이 추가되었다.

securities.csv:

- 각 회사에 대한 일반적인 설명과 부문별 구분 정보.

fundamentals.csv:

- 연간 SEC 10K 보고서에서 추출한 지표들 (2012-2016).
- 대부분의 인기 있는 기본 분석 지표를 도출하는 데 충분한 데이터가 포함되어 있다.

여기서 사용할 데이터는 prices-split-adjusted.csv와 fundamentals.csv이다. 재무지표와 더불어 평균 일일 변동성과 수익률을 feature로 사용할 예정이다.

Fundamentals.csv에는 여러 feature가 존재하는데, 데이터의 수는 약 400개로 매우 적었다. 따라서 각 feature를 연구와 같이 분류하고, 변수를 선택하려고 한다.

Size (규모)

회사의 자산, 수익 및 자본 등의 규모를 나타내는 지표들로, 기업의 전체적인 크기와 재정 상태를 파악하는 데 사용된다.

- Total Assets (총자산)

- Total Current Assets (총유동자산)
- Total Revenue (총수익)
- Common Stocks (보통주)
- Capital Surplus (자본잉여금)
- Total Equity (총자본)
- Total Liabilities & Equity (총부채 및 자본)

Profitability (수익성)

회사의 수익 창출 능력을 평가하는 지표들로, 경영 효율성과 투자 수익성을 분석하는 데 사용된다.

- Gross Margin (매출총이익률)
- Operating Margin (영업이익률)
- Profit Margin (순이익률)
- Pre-Tax Margin (세전이익률)
- After Tax ROE (세후 자기자본이익률)
- Pre-Tax ROE (세전 자기자본이익률)
- Earnings Before Interest and Tax (EBIT, 이자 및 세금 차감 전 이익)
- Earnings Before Tax (세전이익)
- Net Income (순이익)
- Net Income Applicable to Common Shareholders (보통주주 귀속 순이익)

Efficiency (효율성)

회사의 자산과 자원을 얼마나 효율적으로 사용하는지를 평가하는 지표들로, 운영 효율성을 분석하는 데 사용된다.

- Accounts Receivable (매출채권)
- Inventory (재고)
- Changes in Inventories (재고변동)
- Fixed Assets (고정자산)
- Sales, General and Admin. (판매비, 일반관리비)
- Research and Development (연구개발비)

- Cost of Revenue (매출원가)
- Depreciation (감가상각비)

Leverage (레버리지)

회사의 부채 활용도를 평가하는 지표들로, 재무적 위험과 자본 구조를 분석하는 데 사용된다.

- Long-Term Debt (장기부채)
- Short-Term Debt / Current Portion of Long-Term Debt (단기부채/장기부채 현재분)
- Total Liabilities (총부채)
- Net Borrowings (차입금 순액)
- Interest Expense (이자비용)

Liquidity (유동성)

회사의 단기 채무 상환 능력을 평가하는 지표들로, 현금 유동성을 분석하는 데 사용된다.

- Cash and Cash Equivalents (현금 및 현금성 자산)
- Net Cash Flow (순현금흐름)
- Net Cash Flow-Operating (영업활동으로 인한 순현금흐름)
- Net Cash Flows-Financing (재무활동으로 인한 순현금흐름)
- Net Cash Flows-Investing (투자활동으로 인한 순현금흐름)
- Accounts Payable (매입채무)
- Short-Term Investments (단기투자)
- Net Receivables (순채권)

Others (기타)

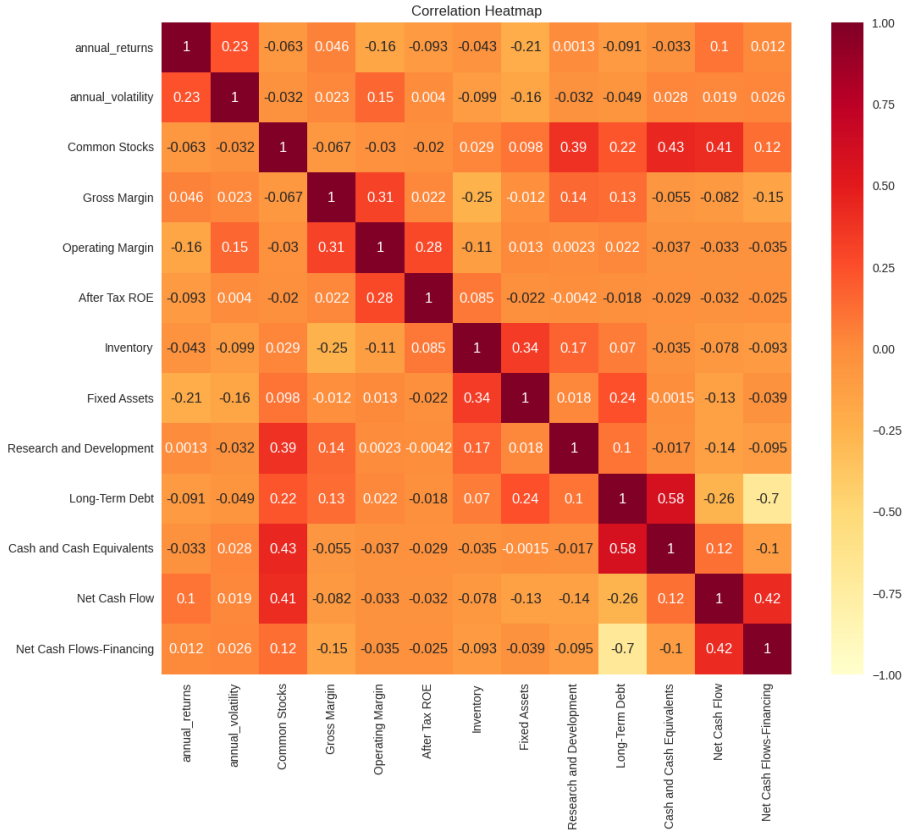
다양한 기타 재무 지표들로, 특정 항목이나 비정기적 사건에 대한 정보를 제공한다. 해당 feature는 다른 분류에 비해 큰 의미가 없다고 판단하여 제거했다.

- Add'l income/expense items (기타 수익/비용 항목)
- Capital Expenditures (자본적 지출)
- Deferred Asset Charges (이연자산비용)
- Deferred Liability Charges (이연부채비용)
- Effect of Exchange Rate (환율 효과)

- Equity Earnings/Loss Unconsolidated Subsidiary (비연결 자회사 지분수익/손실)
- Goodwill (영업권)
- Income Tax (소득세)
- Intangible Assets (무형자산)
- Investments (투자)
- Long-Term Investments (장기투자)
- Minority Interest (비지배지분)
- Misc. Stocks (기타 주식)
- Non-Recurring Items (비반복 항목)
- Other Assets (기타 자산)
- Other Current Assets (기타 유동자산)
- Other Current Liabilities (기타 유동부채)
- Other Equity (기타 자본)
- Other Financing Activities (기타 재무활동)
- Other Investing Activities (기타 투자활동)
- Other Liabilities (기타 부채)
- Other Operating Activities (기타 영업활동)
- Other Operating Items (기타 영업항목)
- Retained Earnings (이익잉여금)
- Sale and Purchase of Stock (주식 매매)
- Treasury Stock (자기주식)

전체 feature를 사용했을 때, k-means clustering의 가장 높은 silhouette index가 0.2로 좋지 않은 결과를 보였다.

해당 결과가 나온 원인이 데이터의 수(약 400개)에 비해 많은 feature(72개)를 사용했을 때 나타나는 차원의 저주때문이라고 생각하여, 각 feature를 특성 별로 분류하고 상관행렬을 계산 후, 높은 상관계수 쌍(기준 0.6)에 해당하는 feature들 중 일부를 정성적으로 제거하여 차원 축소를 수행했다. 그 결과 보다 나은 군집화 결과가 나타났었다. 선택된 feature는 13개로 다음과 같다.



Size (규모)

- **Common Stocks (보통주)**: 주식시장에서 거래되는 회사의 보통주.
- **Fixed Assets (고정자산)**: 장기간 사용될 자산.

Profitability (수익성)

- **Gross Margin (매출총이익률)**: 매출총이익을 매출액으로 나눈 값.
- **Operating Margin (영업이익률)**: 영업이익을 매출액으로 나눈 값.
- **After Tax ROE (세후 자기자본이익률)**: 세후순이익을 자기자본으로 나눈 값.

Efficiency (효율성)

- **Inventory (재고)**: 판매를 위해 보유 중인 제품 및 자재.
- **Research and Development (연구개발비)**: 연구 및 개발에 소요된 비용.

Leverage (레버리지)

- **Long-Term Debt (장기부채)**: 상환 기간이 긴 부채.

Liquidity (유동성)

- **Cash and Cash Equivalents (현금 및 현금성 자산)**: 현금과 쉽게 현금화할 수 있는 자산.

- **Net Cash Flow (순현금흐름)**: 현금 유입에서 현금 유출을 뺀 순액.
- **Net Cash Flows-Financing (재무활동으로 인한 순현금흐름)**: 재무활동에서 발생한 순현금흐름.

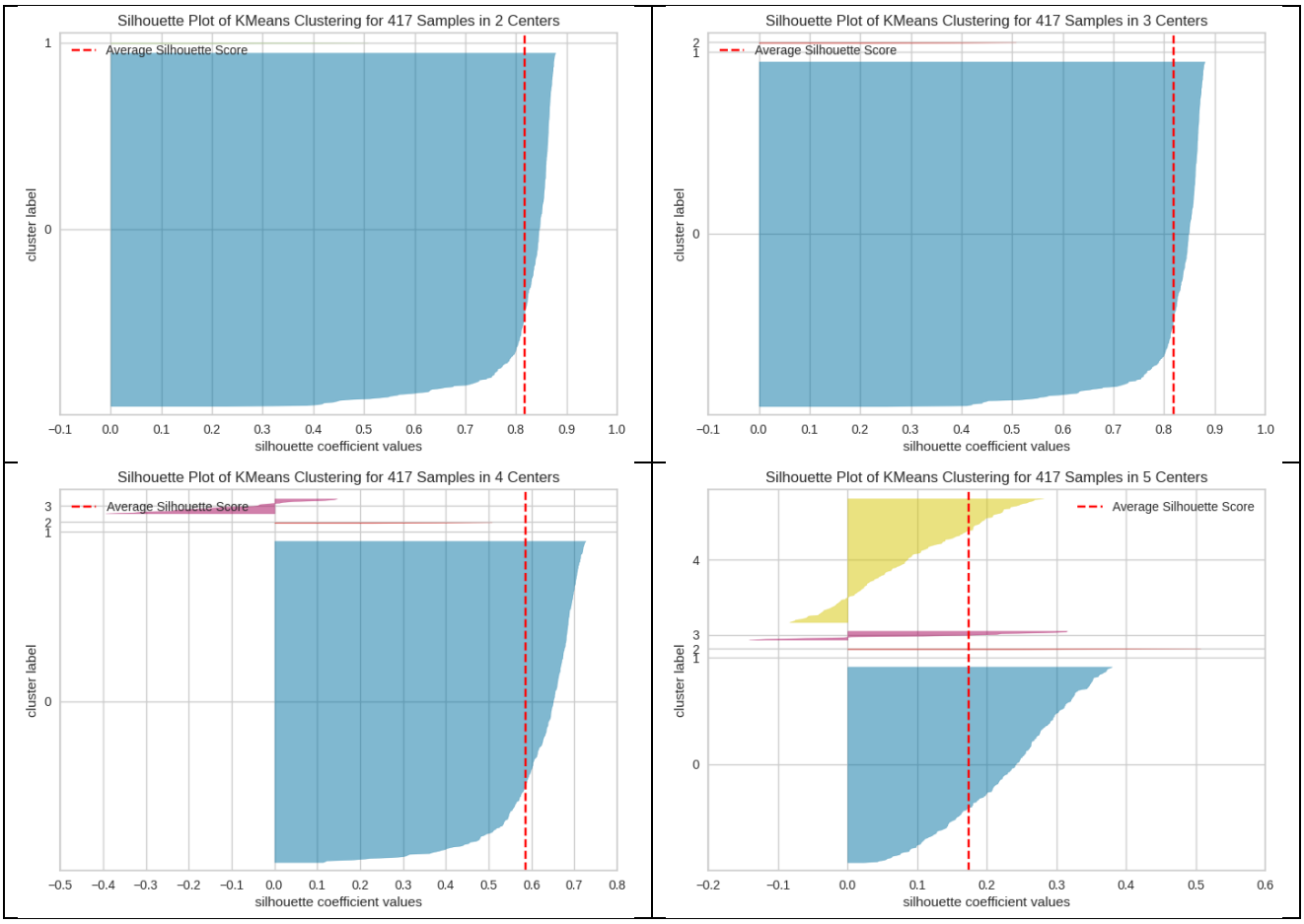
Others (기타)

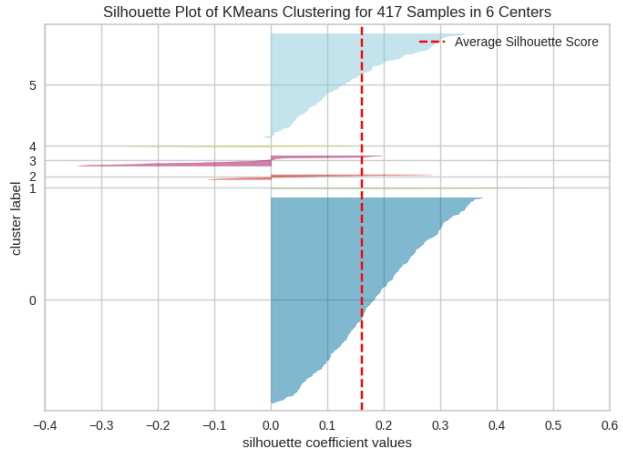
- **annual_returns (연율화 수익률)**: 주식의 연율화된 일일 수익률.
- **annual_volatility (연율화 변동성)**: 주식의 연율화된 일일 변동성.

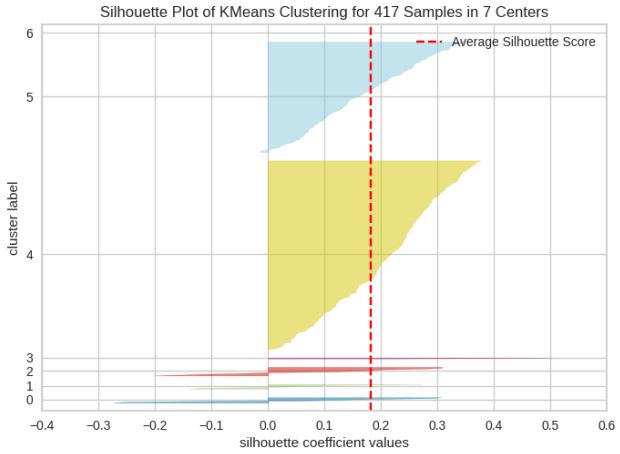
[K-Means Clustering]

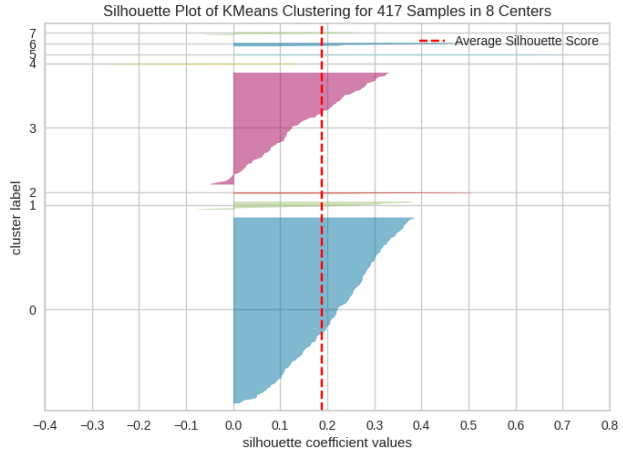
[Q2] K-Means Clustering의 군집 수를 2개부터 K(데이터의 규모에 따라서 본인이 적절히 선정할 것)개까지 증가시켜 (증가 폭 역시 적절히 설정) 가면서 군집화 타당성 지표 값들을 산출하시오. 총 소요 시간은 얼마인가? Silhouette index 기준으로 가장 최적의 군집 수는 몇 개로 판별이 되었는가?

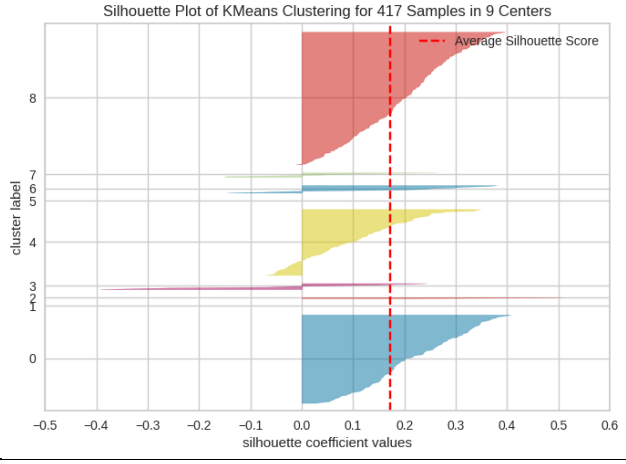
군집화 타당성 지표로 Dunn index와 Silhouette index를 사용했고, 최적의 군집 수 판별에는 silhouette index를 사용했으며, dunn index도 추가적으로 고려했다. 다음은 k별 군집별 Silhouette index를 시각화한 것이다. 각 관측치에 따라 Silhouette index가 급격히 감소하지 않고, 빨간 점선으로 표현되는 평균 Silhouette index가 높을수록 좋은 군집화 결과다.

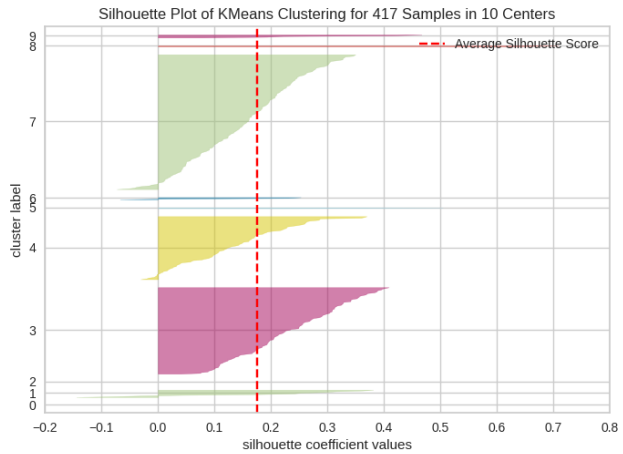



Silhouette Plot of KMeans Clustering for 417 Samples in 6 Centers
Average Silhouette Score
cluster label
silhouette coefficient values


Silhouette Plot of KMeans Clustering for 417 Samples in 7 Centers
Average Silhouette Score
cluster label
silhouette coefficient values

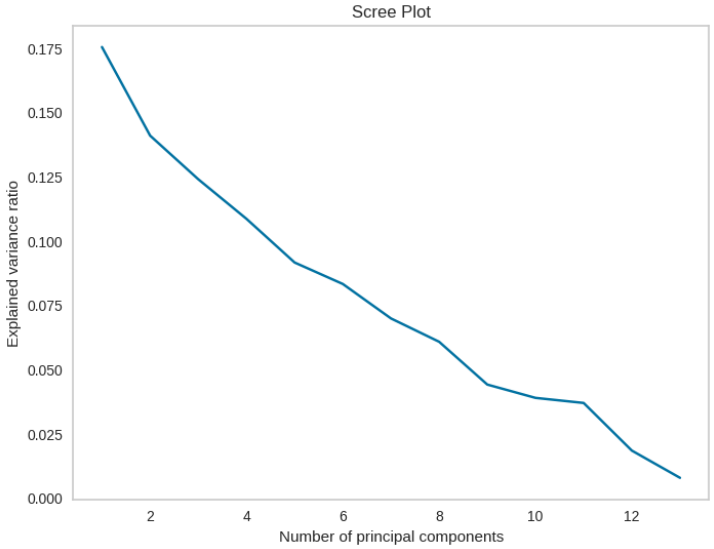

Silhouette Plot of KMeans Clustering for 417 Samples in 8 Centers
Average Silhouette Score
cluster label
silhouette coefficient values


Silhouette Plot of KMeans Clustering for 417 Samples in 9 Centers
Average Silhouette Score
cluster label
silhouette coefficient values

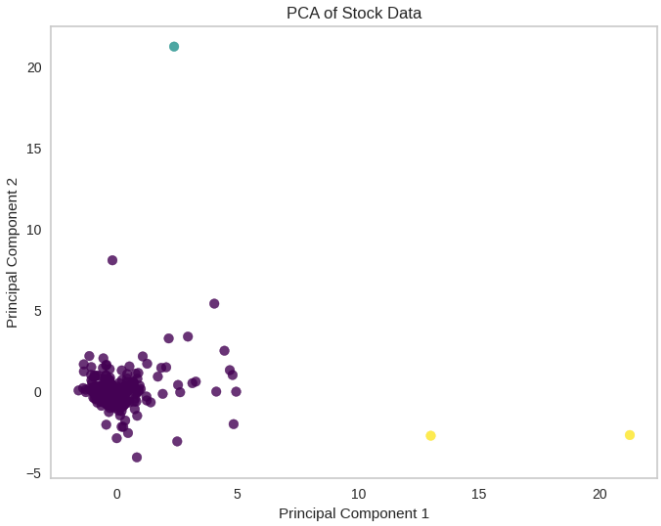

Silhouette Plot of KMeans Clustering for 417 Samples in 10 Centers
Average Silhouette Score
cluster label
silhouette coefficient values

군집화에는 약 4.022초가 소요되었고, K가 3일 때 silhouette index가 최대인 것을 확인할 수 있다. 또한 평균 점수가 0.8인 것으로 보아, 상당히 잘 군집화된 것을 알 수 있다.

군집화 결과를 2차원 평면에 표현하기 위해 PCA(주성분 분석)를 사용하고자 한다.

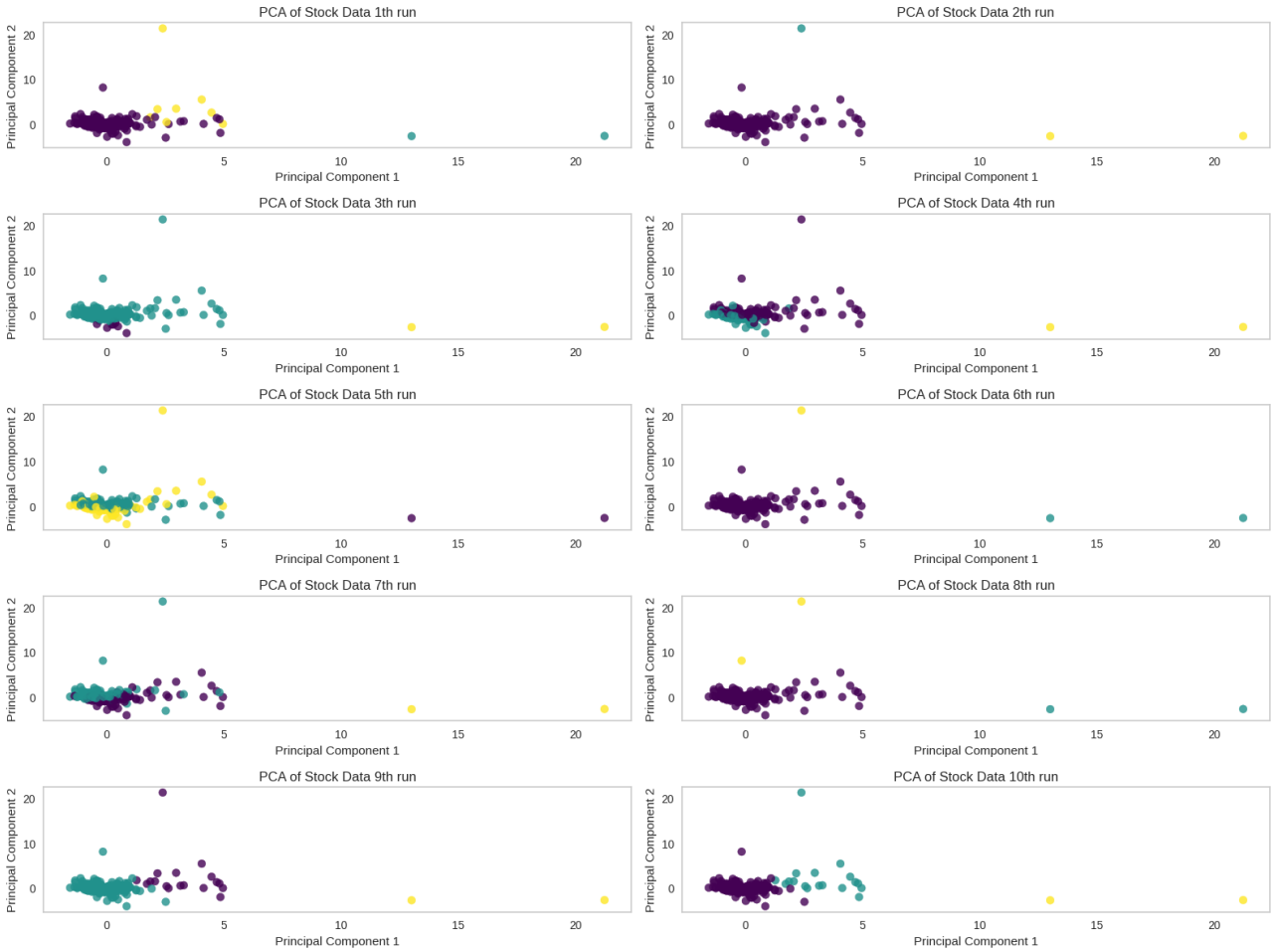


주성분이 2개를 사용하면, 전체 변동의 약 33%를 설명할 수 있는 것을 확인할 수 있다. 낮은 수치지만 단순 시각화를 위해 사용할 것이므로 크게 상관없다.



K=3일 때의 군집화 결과는 다음과 같다. 하나의 군집을 제외하면 군집 내 요소가 상당히 적은 것을 확인할 수 있다. 다음의 문제에서 initial centroid에 따른 군집화 결과를 확인해보자.

[Q3] [Q2]에서 선택된 군집의 수를 사용하여 K-Means Clustering을 10회 반복하고 회차마다 각 군집의 Centroid와 Size를 확인해보시오. 10회 반복 시 몇 가지 경우의 군집화 결과물이 도출되었으며 각 경우의 군집화는 몇번 반복되어 발생하는지 확인해보시오.



각 시행의 군집화 결과를 표로 나타내면 다음과 같다.

Run	Cluster 1	Cluster 2	Cluster 3
1	407	2	8
2	414	2	1
3	408	2	7
4	167	248	2
5	253	162	2
6	414	2	1
7	255	160	2
8	413	2	2
9	**397**	**18**	**2**
10	**397**	**18**	**2**

먼저, 개수 상으로 중복된 군집화 결과는 9번과 10번이 존재했다. 그래프로 확인했을 때에도 각 point의 군집 분류 결과가 동일했다. 따라서 반복해 등장한 군집화 결과 단 하나다. 이를 통해 해당 데이터에 k-means clustering을 적용할 때에는 시작점 초기화에 따라 민감하다는 사실을 알 수 있었다.

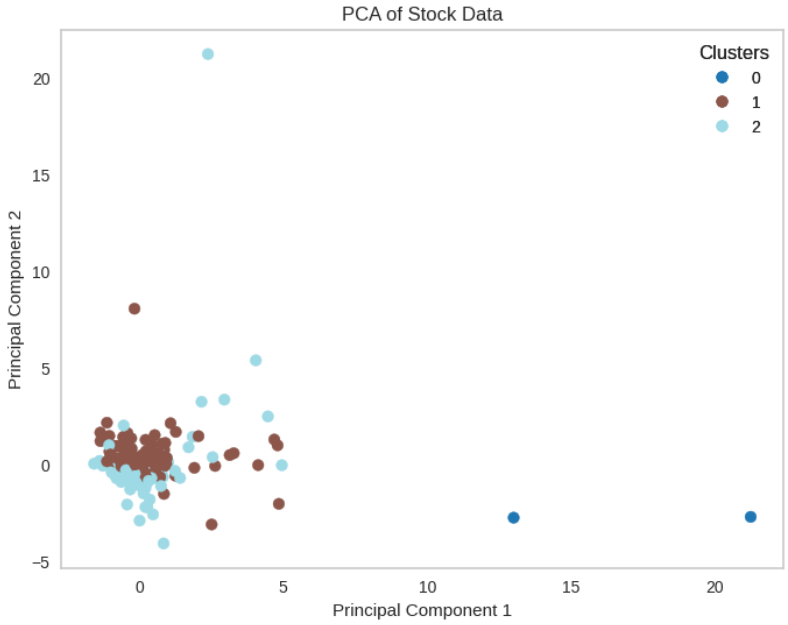
또 하나의 특징으로, 군집화 결과를 확인해 보았을 때, 군집 내 요소가 적은 결과가 빈번하게 나타났다. 해당 현상은 데이터의 관측치 부족과 잘못된 변수 선택에서 기인한다고 생각한다. 이 현상에 대해 다음과 같이 해석했다.

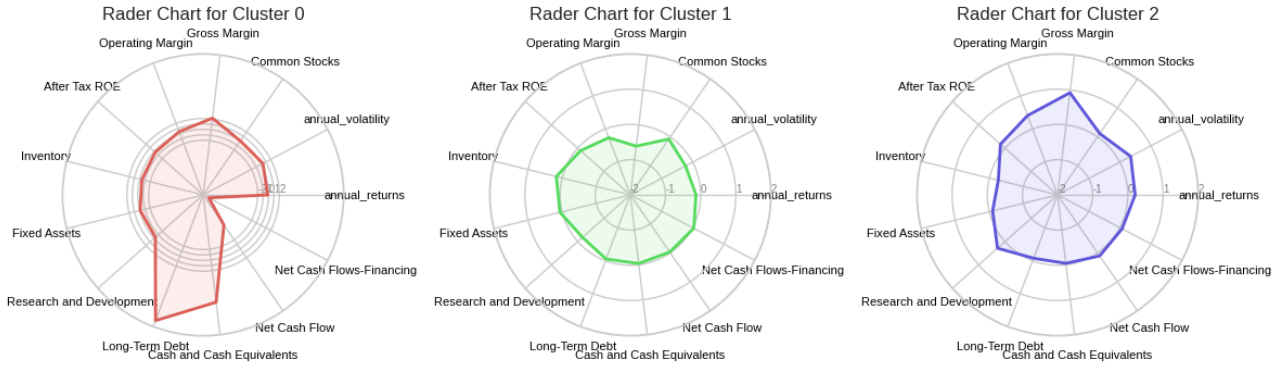
1. 군집 내 요소가 적으면 해당 군집의 특성을 일반화하기 어려울 수 있다. 이는 데이터의 패턴이나 트렌드를 파악하는 데 어려움을 초래할 수 있다. 너무 작은 군집은 실제로 유의미하지 않을 수 있으며, 이는 분석 결과의 신뢰성을 저하시킬 수 있다.

2. 군집화 알고리즘의 매개변수 설정이 부적절하거나, 데이터의 특성에 맞지 않는 알고리즘을 사용한 것일 수 있다. 예를 들어, K-means 알고리즘은 군집의 크기가 균등한 것을 가정하므로, 군집의 크기가 매우 다를 경우 적절하지 않을 수 있다.

[Q4] [Q3]에서 가장 빈번하게 발생한 군집화 결과물에 대해서 각 군집별 변수들의 평균값을 이용한 Radar Chart를 도시해보시오. Radar Chart상으로 판단할 때, 군집의 속성이 가장 상이할 것으로 예상되는 두 군집(군집 A와 군집 B로 명명)과, 가장 유사할 것으로 예상되는 두 군집(군집 X와 군집 Y로 명명)을 각각 선택하고 선택 이유를 설명하시오.

각 결과는 총 2개의 형태로 분류할 수 있다. Run 4, 5, 7과 그 나머지는 다른 형태인 것을 요소의 개수나 그래프로 간단히 확인할 수 있다. 보다 유의미한 해석을 하기 위해 Run 5를 선택하여 진행하겠다.





1. 가장 상이할 것으로 예상되는 두 군집 (군집 A와 군집 B)

- 군집 A: Cluster 0
- 군집 B: Cluster 2

선택 이유

Cluster 0에는 금융 기업 2개로 이루어져 있다.

- **JPM (JP Morgan Chase)**: 글로벌 금융 서비스 회사로서, 투자 은행, 자산 관리, 상업 은행, 개인 금융 서비스 등 다양한 금융 서비스를 제공한다.
- **C (Citigroup)**: 다국적 투자 은행 및 금융 서비스 회사로, 소매 은행, 기업 금융, 증권 서비스, 트레이딩 등 여러 금융 분야에서 활동한다.

이들 금융 기업은 대출을 통해 자산을 운용하는 특성이 있어 long term debt가 높게 나타나며, net cash flow와 net cash flows-financing이 중간 값을 보이고, 금융 자산이 더 큰 비중을 차지하므로 fixed assets은 중간 정도로 나타난다.

Cluster 2에는 주로 기술 및 헬스케어 기업이 포함되어 있습니다. 예시를 들면 다음과 같다.

- **AAPL (Apple)**: 소비자 전자제품, 소프트웨어, 온라인 서비스를 설계, 제조 및 판매하는 기술 회사다.
- **AMGN (Amgen)**: 생명공학 회사로서, 인체 약물 개발에 주력하고 있다.

기술 및 헬스케어 기업은 다음과 같은 특징을 보인다. 기술 및 헬스케어 기업은 주로 연구 개발과 혁신에 투자하므로, Long-Term Debt가 낮을 수 있다. 연구소와 제조 시설 등 고정 자산을 많이 보유하고 있을 수 있다. 기술 및 헬스케어 기업은 높은 Gross Margin, Operating Margin, After Tax ROE를 보일 수 있다. 이는 고마진 제품과 혁신적인 서비스 제공으로 인해 높은 수익성을 나타낸다고 해석할 수 있다.

각 재무지표 별로 상이한 부분을 설명하면 다음과 같다.

Gross Margin:

- Cluster 0은 매우 높은 Gross Margin (2.160270)을 보인다. 이는 기업이 제품 판매에서 상당한 이익을 창출하고 있음을 나타낸다.
- Cluster 2는 상대적으로 낮은 Gross Margin (0.919987)을 보인다. 수익성이 있지만, Cluster 0보다는 낮다.

Operating Margin:

- Cluster 0과 Cluster 2 모두 유사한 수준의 Operating Margin을 가지고 있다. 이는 운영 효율성이 비슷함을 나타낸다.

After Tax ROE:

- Cluster 0은 부정적인 After Tax ROE (-0.235174)를 보입니다. 이는 자본 대비 수익성이 낮음을 의미한다.
- Cluster 2는 긍정적인 After Tax ROE (0.170404)를 보입니다. 이는 자본 대비 수익성이 높음을 의미한다.

Long-Term Debt:

- Cluster 0은 매우 높은 Long-Term Debt (12.696327)를 가지고 있다. 이는 금리 상승 시 큰 재정적 부담을 의미한다.
- Cluster 2는 매우 낮은 Long-Term Debt (-0.075000)를 가지고 있다. 이는 금리 상승 시 재정적 부담이 적음을 의미한다.

Cash and Cash Equivalents:

- Cluster 0은 높은 Cash and Cash Equivalents (7.846413)를 보유하고 있다. 이는 높은 유동성을 의미한다.
- Cluster 2는 낮은 Cash and Cash Equivalents (-0.041095)를 보유하고 있다. 이는 낮은 유동성을 의미한다.

Net Cash Flow:

- Cluster 0는 부정적인 Net Cash Flow (-5.302990)를 보인다. 이는 운영으로 인한 현금 유출이 크다는 것을 의미한다.
- Cluster 2는 긍정적인 Net Cash Flow (0.106466)를 보인다. 이는 운영으로 인한 현금 유입이 크다는 것을 의미한다.

결론

- 수익성: Cluster 0는 매우 높은 Gross Margin을 보이지만, After Tax ROE는 부정적이다. Cluster 2는 낮은 Gross Margin을 보이지만, After Tax ROE는 긍정적이다.
- 래버리지(재정적 안정성): Cluster 2는 매우 낮은 Long-Term Debt를 가지고 있으며, 이는 금리 상승 시 재정적 부담이 적음을 의미한다.
- 유동성: Cluster 0는 금융 기업이라는 특성 상 높은 Cash and Cash Equivalents를 보유하고 있고, Net Cash Flow는 부정적이다. Cluster 2는 낮은 Cash and Cash Equivalents를 보유하고 있지만, Net Cash Flow는 긍정이다.

따라서, 현재 금리 상황을 고려할 때, Cluster 2가 재무적으로 더 안정적이고 우수하다고 판단할 수 있다. 낮은 부채 수준과 긍정적인 현금 흐름 덕분에 금리 상승에 따른 재정적 부담이 적기 때문이다.

2. 가장 유사할 것으로 예상되는 두 군집 (군집 X와 군집 Y)

- 군집 X: Cluster 1
- 군집 Y: Cluster 2

선택 이유

해당 두 군집은 Radar 도표 상 평균적으로 거의 유사한 재무 지표를 가지고 있다. 또한 Cluster 1과 Cluster 2 모두 다양한 산업에 걸쳐 기업들이 분포되어 있어, 특정 산업의 리스크에 덜 노출되어 있다.

결론

- 재무 지표: Cluster 1과 Cluster 2는 낮은 부채 수준과 유사한 현금 및 현금성 자산, 그리고 긍정적인 현금 흐름을 공유한다. 이는 두 군집이 재정적 안정성을 가지고 있음을 의미한다.
- 군집 내 기업: 두 군집 모두 헬스케어와 기술 산업의 기업들이 다수 포함되어 있어, 안정적인 수익성 및 성장 가능성을 가지고 있다. 특히, 헬스케어 산업의 기업들은 경제 상황에 덜 민감하고, 기술 산업의 기업들은 높은 수익성을 보이는 경향이 있어, 두 군집이 유사한 재정적 특성을 공유한다.

[Q5] [Q4]에서 선택된 군집 A와 군집 B에 대해 각 변수별 평균값 차이에 대한 통계적 검정을 수행하시오. 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중은 얼마인가? 또한 [Q4]에서 선택된 군집 X와 군집 Y에 대해서도 각 변수별 평균값 차이에 대한 통계적 검정을 수행하시오. 이 경우, 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중은 얼마인가?

1. 군집 A와 군집 B(Cluster 0과 Cluster 2)

유의수준 0.05에서 차이가 유의미한 변수의 비중 : 약 0.38

annual_returns

- **원인**: Cluster 2는 기술 및 헬스케어 기업이 다수 포함되어 있어, 이러한 산업들은 최근 몇 년간 높은 수익률을

기록했다. 특히, 기술 산업은 혁신과 성장으로 인해 높은 투자 수익률을 보이고, 헬스케어 산업은 안정적 수익을 유지하고 있다.

- **설명**: 기술 기업의 혁신과 성장, 헬스케어 기업의 안정적인 수익 창출이 Cluster 2의 높은 annual_returns를 설명할 수 있다.

Gross Margin

- **원인**: Cluster 0는 금융 기업이 다수 포함되어 있으며, 금융 산업은 일반적으로 높은 이익 마진을 기록한다. 반면, Cluster 2는 기술 및 헬스케어 산업으로, 이들 산업은 제품과 서비스 판매에서 높은 마진을 기록한다.
- **설명**: 금융 기업의 이자 수익과 서비스 수수료, 기술 기업의 고마진 제품 판매, 헬스케어 기업의 특허 보호된 고가 제품이 높은 Gross Margin에 기여한다.

After Tax ROE

- **원인**: Cluster 2는 높은 수익성을 가진 기술 및 헬스케어 기업이 다수 포함되어 있어, 이러한 기업들은 높은 자기자본이익률(ROE)을 기록한다. Cluster 0의 금융 기업들은 부채 비율이 높아 ROE가 낮을 수 있다.
- **설명**: 기술 및 헬스케어 기업의 높은 수익성과 낮은 부채비율이 Cluster 2의 높은 After Tax ROE에 기여한다.

Inventory

- **원인**: Cluster 0의 금융 기업들은 물리적 재고를 보유하지 않으며, Cluster 2의 기술 기업들은 효율적인 재고 관리를 통해 낮은 재고 수준을 유지한다. 반면, 일부 헬스케어 및 제조업체는 재고를 다량 보유할 수 있다.
- **설명**: 금융 기업의 재고 없음, 기술 기업의 재고 관리 효율성, 헬스케어 및 제조업체의 재고 보유 전략이 Inventory의 차이를 설명한다.

Research and Development

- **원인**: Cluster 2의 기술 및 헬스케어 기업들은 R&D에 많은 투자를 한다. 이는 기술 혁신과 신약 개발을 위한 필수적인 비용이다. 반면, Cluster 0의 금융 기업들은 상대적으로 R&D 투자가 적다.
- **설명**: 기술 및 헬스케어 기업의 R&D 집중 투자가 Cluster 2의 높은 R&D 비용에 기여한다.

2. 군집 X와 군집 Y(Cluster 1과 Cluster 2)

유의수준 0.05에서 차이가 유의미한 변수의 비중 : 약 0.54

유사하다고 판단한 것과 반대로, 차이가 유의미하게 나타나는 변수의 비중이 더 높게 나타났다. Cluster 0과 2는 그래프상 특정 재무지표에서의 차이가 매우 두드러지게 나타나고, Cluster 1과 2는 통계적으로 더 상이한 군집으로 나타났다는 차이점이 있다.

1. annual_returns

- **군집 1**: 연간 수익률이 비교적 낮음.
- **군집 2**: 연간 수익률이 높음.

- **원인**: 군집 2에는 기술 및 헬스케어 기업이 많이 포함되어 있으며, 이들 산업은 최근 몇 년간 높은 성장률과 수익률을 기록했다. 특히 기술 산업의 혁신과 헬스케어 산업의 안정적인 수익 구조가 주요 요인이다.

2. annual_volatility

- **군집 1**: 연간 변동성이 낮음.
- **군집 2**: 연간 변동성이 높음.
- **원인**: 군집 2의 기술 및 헬스케어 기업들은 시장에서 높은 변동성을 보인다. 기술 산업은 빠른 혁신과 변화로 인해 변동성이 높고, 헬스케어 산업은 규제와 신약 개발 성공 여부에 따라 변동성이 크다.

3. Gross Margin

- **군집 1**: Gross Margin이 낮음.
- **군집 2**: Gross Margin이 높음.
- **원인**: 군집 2의 기술 및 헬스케어 기업들은 고마진 제품과 서비스 판매를 통해 높은 Gross Margin을 기록한다. 특히 기술 기업의 고마진 제품과 헬스케어 기업의 특허 보호된 고가 제품이 주요 요인이다.

4. Operating Margin

- **군집 1**: Operating Margin이 낮음.
- **군집 2**: Operating Margin이 높음.
- **원인**: 군집 2의 기술 및 헬스케어 기업들은 운영 효율성이 높아 높은 Operating Margin을 기록한다. 기술 기업의 운영 효율성과 혁신 제품 판매가 주요 요인이다.

5. After Tax ROE

- **군집 1**: After Tax ROE가 낮음.
- **군집 2**: After Tax ROE가 높음.
- **원인**: 군집 2의 기술 및 헬스케어 기업들은 높은 수익성을 가지고 있어 높은 자기자본이익률(ROE)을 기록한다. 이는 기술 산업의 높은 수익성과 헬스케어 산업의 안정적인 수익 창출 때문이다.

6. Inventory

- **군집 1**: Inventory가 높음.
- **군집 2**: Inventory가 낮음.
- **원인**: 군집 2의 기술 기업들은 효율적인 재고 관리를 통해 낮은 재고 수준을 유지한다. 반면, 군집 1은 다양한 산업의 기업들로 구성되어 있어 재고 수준이 더 높을 수 있다.

7. Research and Development

- **군집 1**: R&D 비용이 낮음.

- **군집 2**: R&D 비용이 높음.
- **원인**: 군집 2의 기술 및 헬스케어 기업들은 연구개발(R&D)에 많이 투자 한다. 이는 기술 혁신과 신약 개발을 위한 필수적인 비용이다.

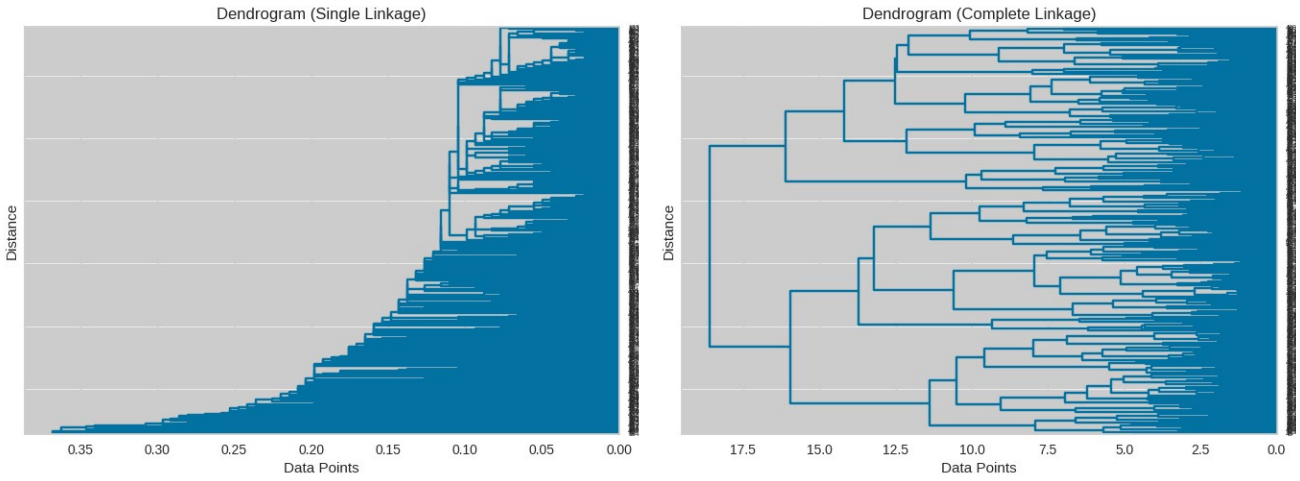
결론

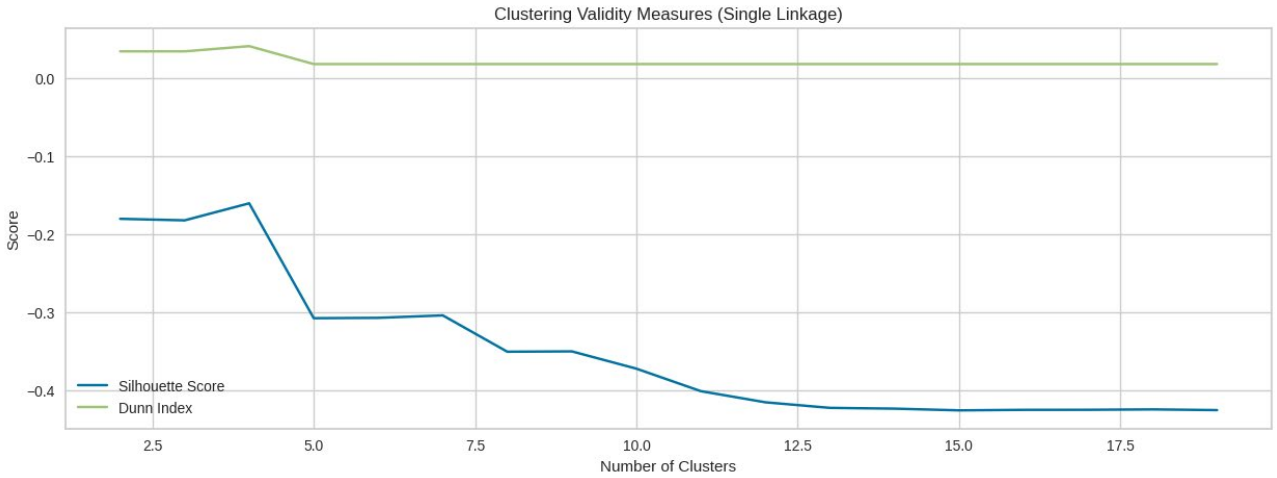
각 군집 간의 유의미한 차이가 있는 변수들은 주로 산업 특성과 관련이 있다. Cluster 2는 주로 기술 및 헬스케어 기업으로 구성되어 있으며, 이들 산업은 높은 수익성과 변동성을 보인다. 반면, Cluster 1은 다양한 산업의 기업들로 구성되어 있다. Cluster 0은 금융 기업으로 구성되어 있다. 이러한 산업 특성과 재무 지표의 차이가 각 군집 간의 유의미한 차이를 설명한다.

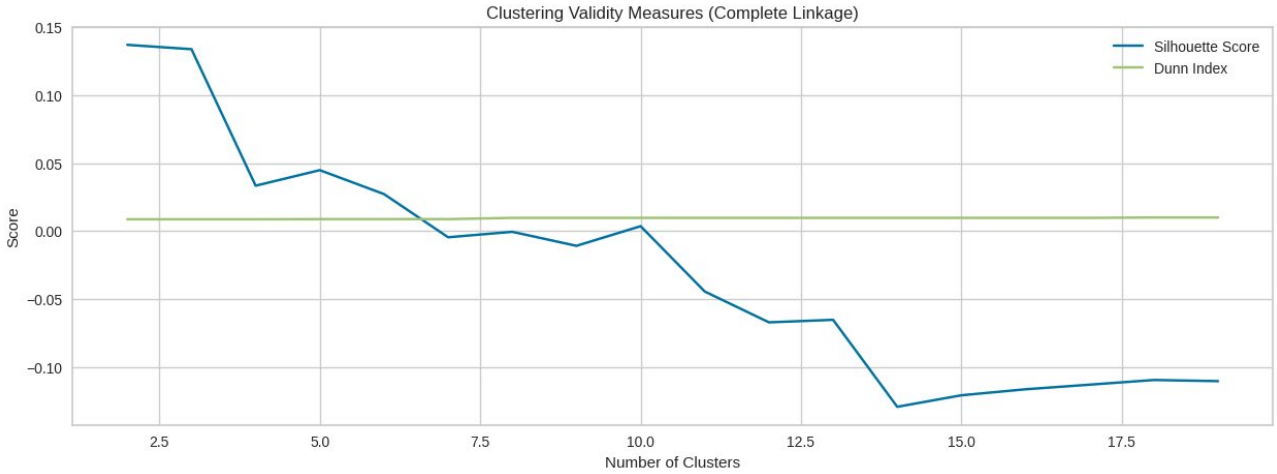
[Hierarchical Clustering]

[Q6] 두 객체 사이의 유사도를 측정하는 지표를 본인의 기준에 따라 정의하고(유클리드 거리, 상관계수 등) “single”과 “complete” 두 가지 linkage에 대해 군집 수를 2개부터 K(데이터의 규모에 따라서 본인이 적절히 선정할 것)개까지 증가시켜(증가 폭 역시 적절히 설정) 가면서 군집화 타당성 지표 값들을 산출하시오. Silhouette index 기준으로 가장 최적의 군집 수는 몇 개로 판별이 되었는가?

유사도 측정 지표로는 주식 재무지표들 간의 순위 상관관계를 반영할 수 있는 스피어만 상관계수 거리를 사용했다.



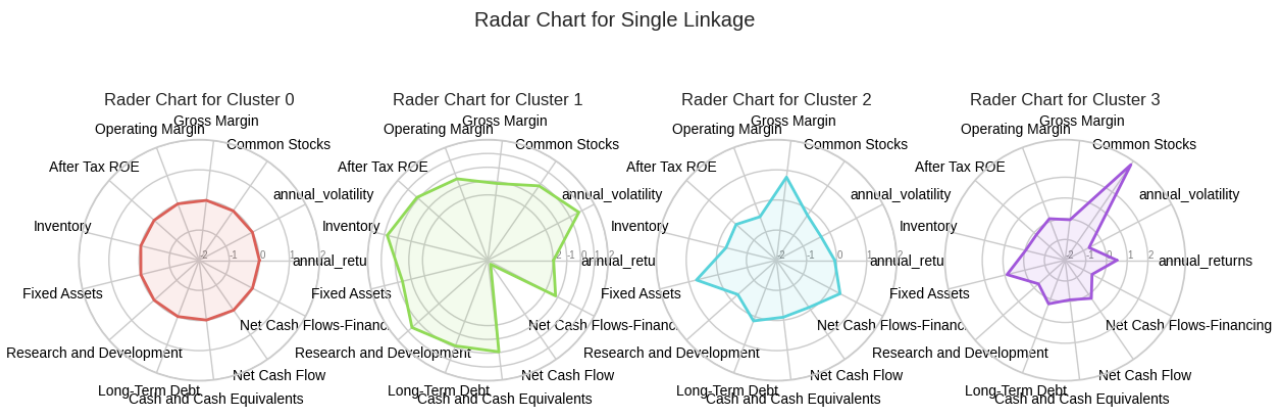


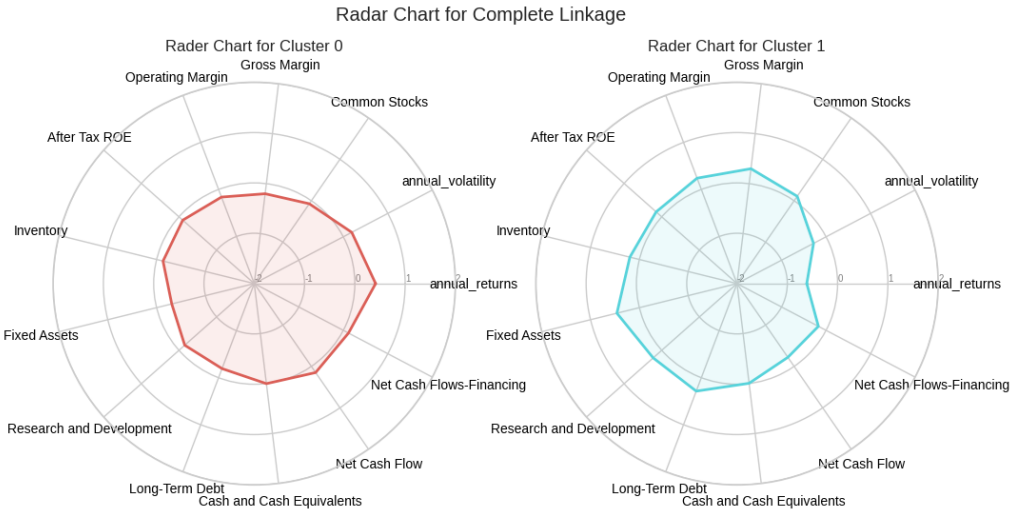


Single linkage의 경우 최적 군집 개수가 4개지만, Silhouette index가 음수로 좋지 않은 군집화로 나타났다.

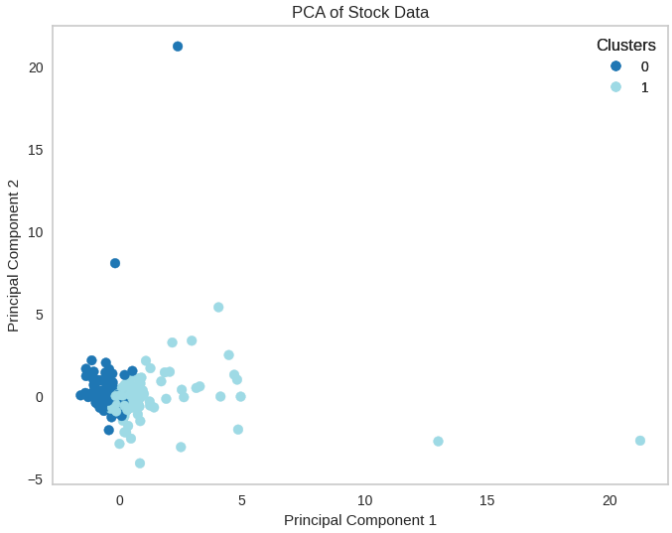
반면에 Complete linkage의 경우 최적 군집 개수가 2개이고, Silhouette index가 양수이나 작은 값으로 좋지 않은 군집화로 나타났다. 비교적 나은 군집 수는 complete linkage를 사용한 최적 군집 개수 2개로 판단했다.

[Q7] [Q6]에서 찾은 최적의 군집 수에 대해서 각 군집들의 변수값의 평균값을 이용한 Radar Chart를 도시해보시오. Radar Chart를 바탕으로 판단할 때, K-Means Clustering과 보다 유사한 결과물이 나오는 방식은 어떤 Linkage인지 본인의 생각을 바탕으로 서술해보시오.





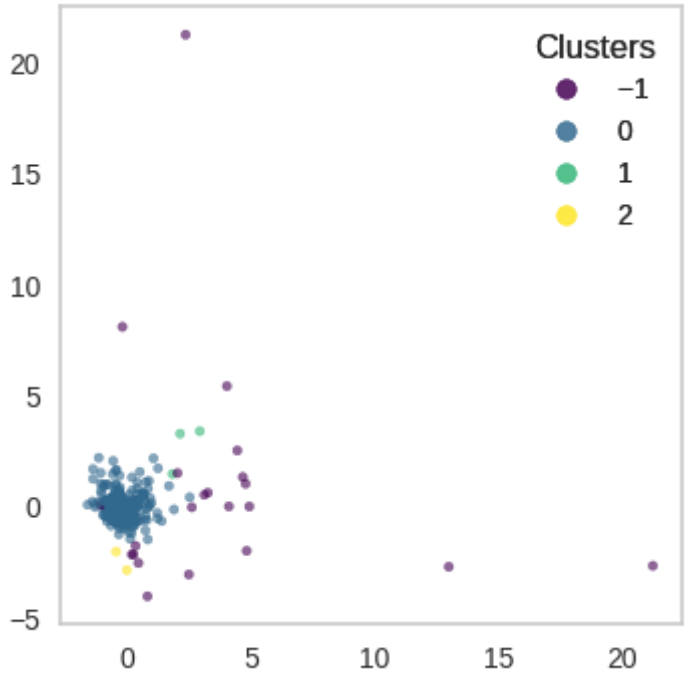
각 군집 별 요소 개수를 확인했을 때, single linkage는 하나를 제외한 나머지 군집이 모두 요소를 1개만 가지는 것으로 나타났다. 반면에 complete linkage는 cluster가 2개이나, 비교적 균등한 요소 수와 요소의 산업 분류를 확인 했을 때 K-means clustering 결과와 유사성을 보인다. 따라서 complete linkage가 K-means clustering(특히 Cluster 1과 Cluster 2)과 유사하다고 판단했다.



[DBSCAN]

[Q8] DBSCAN 알고리즘의 eps 옵션과 minPts 옵션을 조정해가면서 [Q2]에서 선정한 최적 개수의 군집이 찾아지는 eps 값과 minPts 값을 찾아보시오.

Eps가 3, min_smaples가 2일 때, [Q2]의 최적 군집 개수인 3개를 얻을 수 있었다.



[Q2]와의 유사성을 찾자면 해석한 Run 5가 아닌, 다른 형태의 결과(큰 군집 하나와 작은 군집 2개)와 유사하다.

[Q9] [Q8]에서 찾은 군집화 결과물에서 Noise로 판별된 객체의 수가 몇 개인지 확인해 보시오.

Label이 -1인 객체는 총 22개이다.

이상치로 판별된 객체들은 …

[종합]

[Q10] 이 데이터셋에 가장 적합한 군집화 알고리즘은 무엇이라고 생각하는지 본인이 생각한 근거를 이용하여 서술하시오.

PCA로 산포를 plot 해보았을 때 밀집된 클러스터와 여러 이상치가 나타난다. 이런 상황에서는 **DBSCAN**이 가장 적합한 군집화 알고리즘이라고 판단된다. 이는 밀도 기반 접근 방식을 사용하여 밀도가 높은 영역을 클러스터로 식별하고, 밀도가 낮은 영역을 이상치로 처리할 수 있기 때문이다.