**[여섯 번째 과제]**

**IT응용시스템공학과**

**1494053**

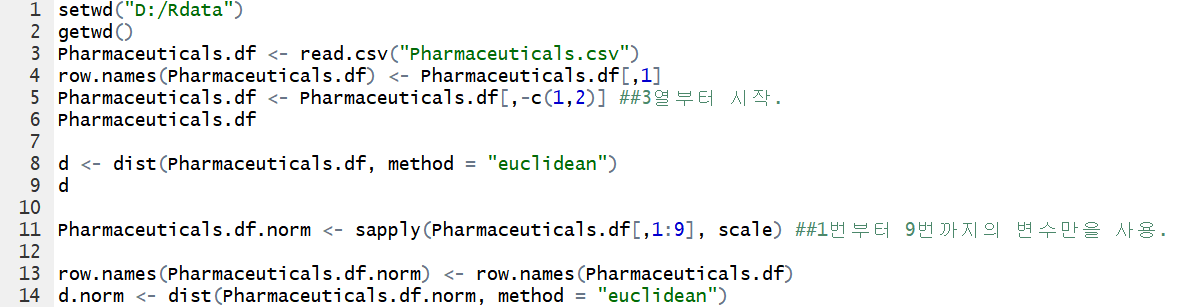
**김희택**

**15.2 한 주식 애널리스트는 제약 산업에 대해 연구 중이며, 회사에서 수집한 재무데이터를 탐색하고 이해하는 데 도움을 받고자 한다. 분석의 주된 목적은 일부 기본적인 재무변수들을 사용하여 제약 산업의 구조를 이해하는 것이다. 재무데이터는 제약 산업에 속하는 21개의 회사로부터 수집되었고, Pharmaceuticals.csv 파일에서 이용가능하다. 각 회사들에 대해서 다음과 같은 변수가 수집되었다.**

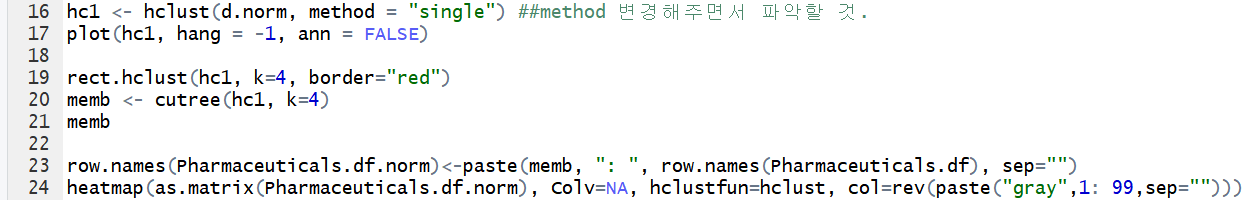
1. **시가 총액(단위: 10억 달러)**
2. **베타**
3. **주가/수익 비**
4. **자기자본 이익률**
5. **자산 수익률**
6. **자산 회전율**
7. **레버리지**
8. **추정 수익성장률**
9. **순 이윤 폭**
10. **주요 증권사들의 추천**
11. **회사의 본사 위치**
12. **회사가 상장된 주식거래소**

**주어진 데이터세트를 탐색하고 분석하기 위해 다음과 같이 군집분석을 하시오.**

**a. 21개의 회사를 군집화하기 위해 양적변수들 (1번부터 9번까지의 변수)만을 사용하시오. 군집분석을 수행할 때 고려해야 할 여러 가지 선택사항들, 예를 들어 서로 다른 변수들의 가중치, 특정 군집화 알고리즘, 군집의 수 등에 관하여 정당성을 기술하시오.**

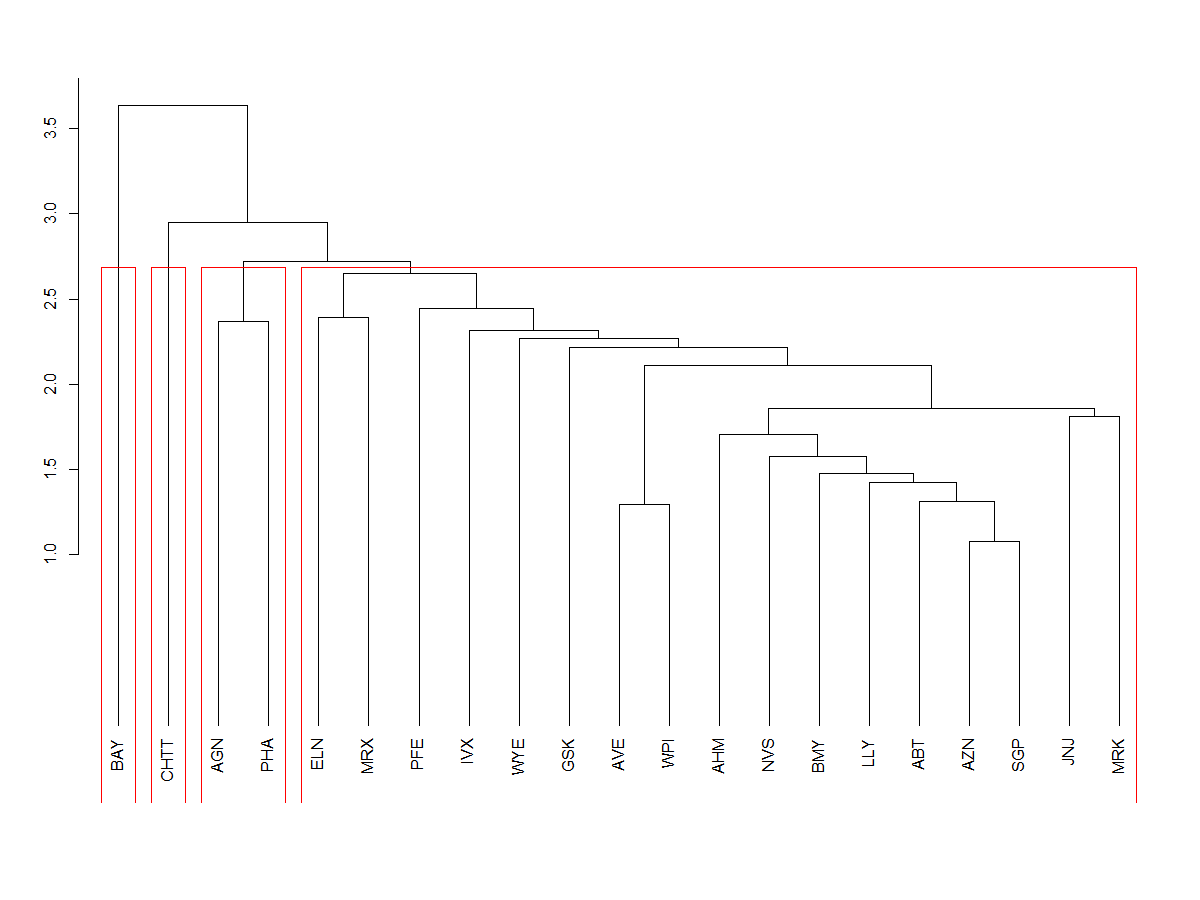


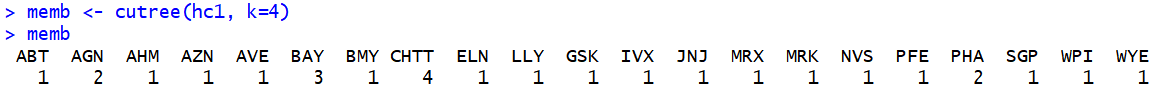
* 먼저 원시 거리 측정은 측정치의 단위에 크게 영향을 받으므로, 첫 번째 변수 ‘시가 총액’이 다른 변수들에 비해 상대적으로 단위가 크기 때문에 정규화 작업을 진행하였습니다.



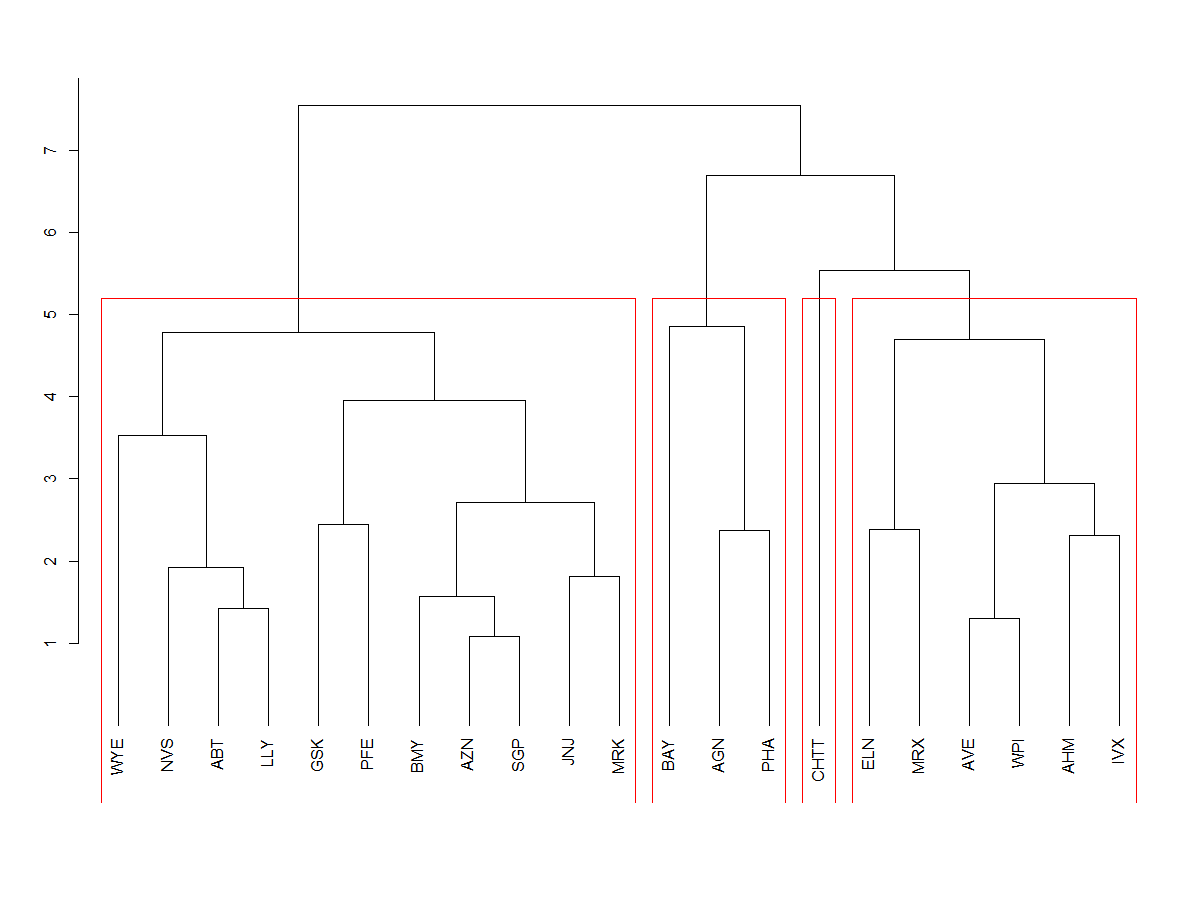
* 군집 간의 거리 측정 대표 방법인 ‘단일연결법’, ‘완전연결법’, ‘평균연결법’ 알고리즘을 이용하여 군집 4개를 기준으로 계층적 군집화 단계 응집 방법을 사용하였습니다. 또한, 의미 있고 유용한 군집을 얻었는지 시각화 판단을 할 수 있도록 ‘Heatmap’을 이용하였습니다.

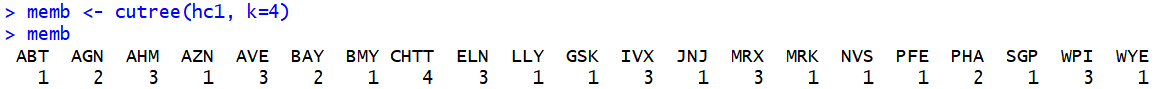
**[단일연결법]**



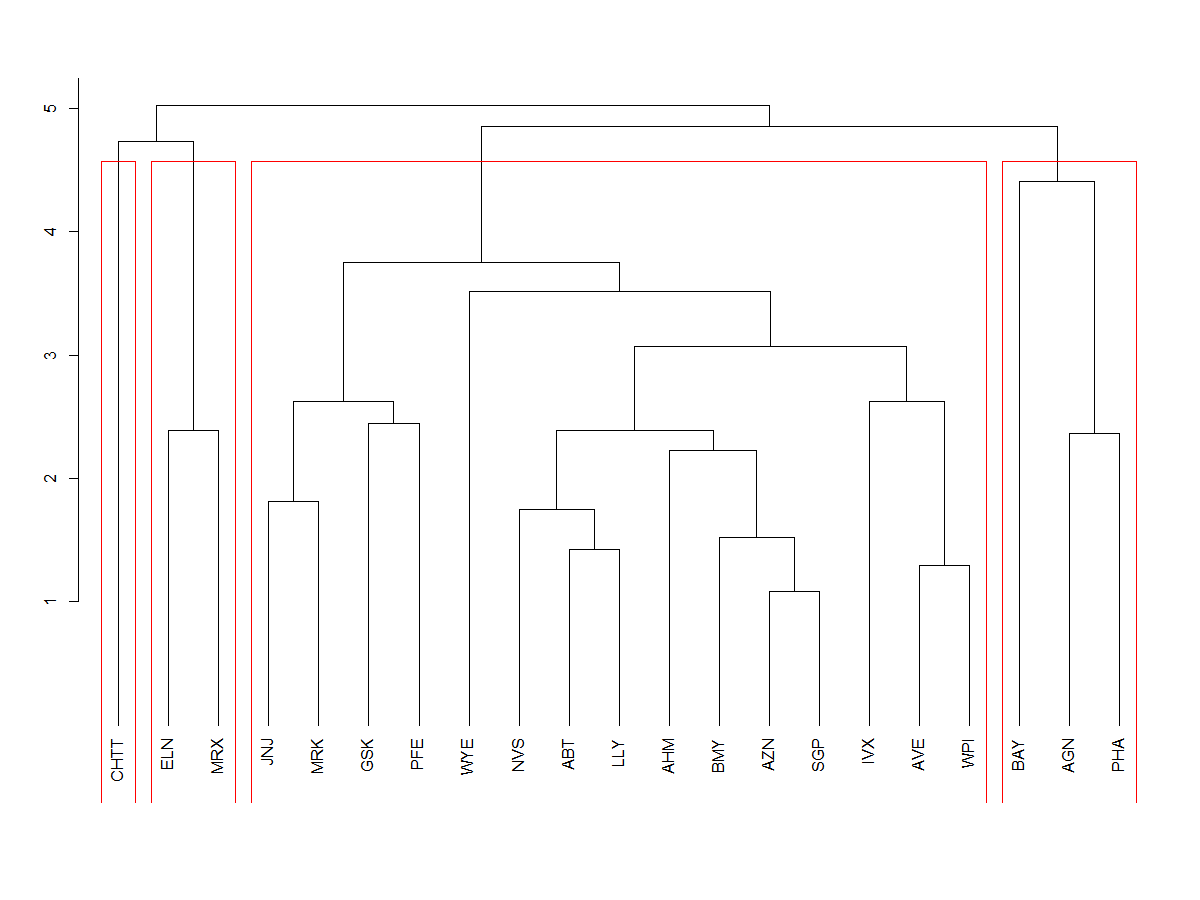


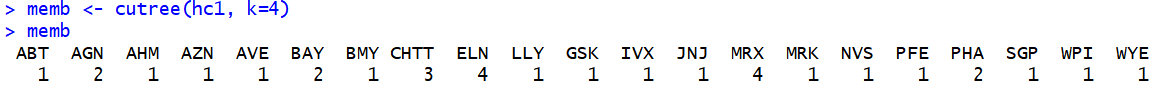
**[완전연결법]**





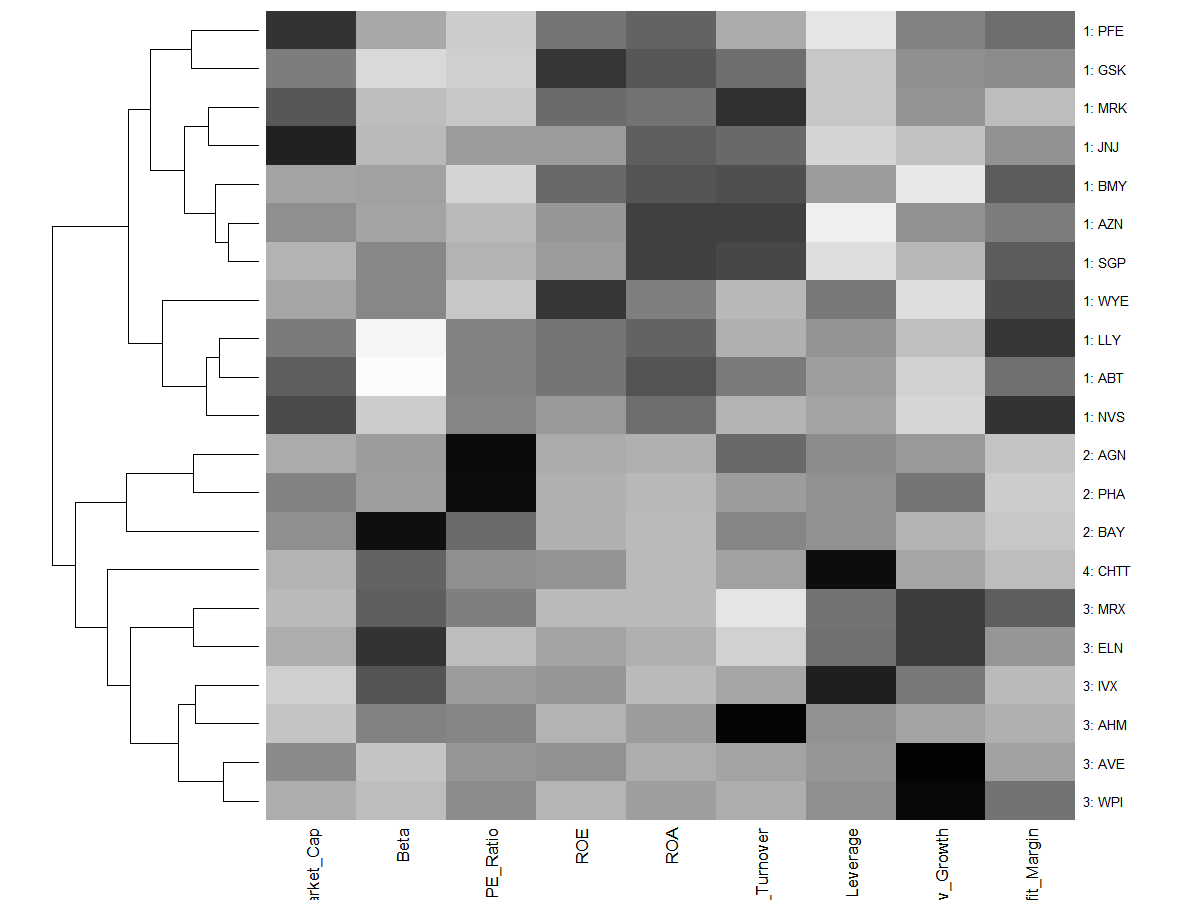
**[평균연결법]**





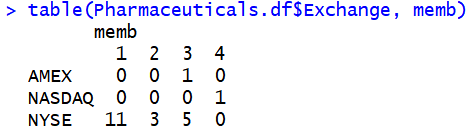
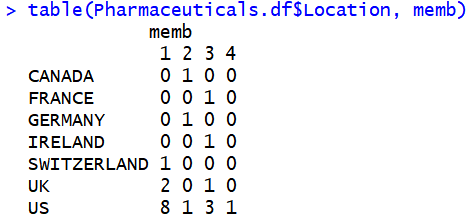
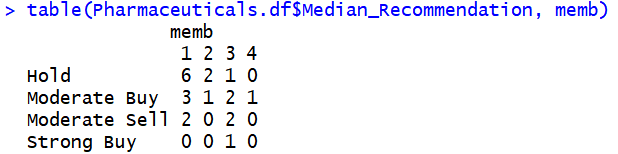
* 위의 결과는 군집 간의 거리 측정인 ‘최단 거리’, ‘최대 거리’, ‘평균 거리’를 이용한 것입니다. 또한 군집의 수는 초기에 4개로 가정하여 결과를 도출하였습니다. 위 결과를 통해 ‘계층적 방법인 응집 방법’을 사용하는 것이 군집분석을 수행하기에 타당하다고 생각합니다.

**b. 군집을 형성하는 데 사용된 양적변수들과 관련하여 군집결과를 해석하시오.**



* 군집 1은 Market\_Cap (시가 총액) / ROA (자산 수익률) / Net\_Profit\_Margin (순이윤 폭)이 타 군집보다 상대적으로 높은 특징을 보입니다.
* 군집 2는 PE\_Ration (주가/수익 비)가 타 군집보다 상대적으로 높은 특징을 보입니다.
* 군집 3은 Beta (베타) / Rev\_Growth (추정 수익성장률)이 타 군집보다 상대적으로 높은 특징을 보입니다.
* 군집 4는 Leverage (레버리지)가 타 군집보다 상대적으로 높은 특징을 보입니다.

**c. 군집을 형성하는 데 사용되지 않은 수치형 변수들 (10번부터 12번까지의 변수)과 관련하여 군집에 어떤 패턴이 존재하는가?**



* 먼저 table() 함수를 이용하여 사용되지 않는 수치형 변수 (10번 ~ 12번)에 대해 2차원 분할표를 만들었습니다.
* 군집 1의 ‘주요 증권사들의 추천’은 대체로 ‘Hold’ 임을 알 수 있습니다. 또한 ‘회사의 본사 위치’는 ‘US’가 압도적이며 ‘회사가 상장된 주식거래소’도 ‘NYSE’에 전부 해당되는 패턴을 확인 할 수 있습니다.
* 군집 2의 ‘주요 증권사들의 추천’은 대체로 ‘Hold’ 임을 알 수 있습니다. 또한 ‘회사의 본사 위치’는 ‘CANADA’, ‘GERMANY’, ‘US’로 나뉘며 ‘회사가 상장된 주식거래소’는 ‘NYSE’에 전부 해당되는 패턴을 확인 할 수 있습니다.
* 군집 3의 ‘주요 증권사들의 추천’은 대체로 ‘Moderate Buy or Sell’ 임을 알 수 있습니다. 또한 ‘회사의 본사 위치’는 대부분이 ‘US’이고 ‘회사가 상장된 주식거래소’도 대부분 ‘NYSE’에 해당되는 패턴을 알 수 있습니다.
* 군집 4의 ‘주요 증권사들의 추천’은 ‘Moderate Buy’이고 ‘회사의 본사 위치’는 ‘US’이며 ‘회사가 상장된 주식거래소’는 ‘NASDAQ’인 패턴을 알 수 있습니다.

**d. 데이터세트에 있는 일부 혹은 모든 변수를 사용하여 각 군집에 적절한 이름을 부여하시오.**

* 군집 1은 시가 총액과 자산 수익률, 순이윤 폭이 높지만 ‘주요 증권사들의 추천’은 대체로 ‘Hold’입니다.

즉 아직 중립을 유지하는 것으로 보아 ‘알짜기업’이라는 이름을 부여하였습니다.

* 군집 2는 주가/수익 비가 높고 ‘주요 증권사들의 추천’은 대체로 ‘Hold’입니다.

즉 미래 가치는 높지만 아직 좀 더 지켜봐야 할 ‘비트코인’이라는 이름을 부여하였습니다.

* 군집 3은 베타와 추정 수익성장률이 높고 ‘주요 증권사들의 추천’은 대체로 ‘Moderate Buy or Sell’입니다.

즉 증권사들의 추천에 따라 적당한 매수와 매도가 필요한 ‘타이밍’이라는 이름을 부여하였습니다.

* 군집 4는 레버리지가 높고 ‘주요 증권사들의 추천’은 ‘Moderate Buy’입니다.

즉 자본 이익률이 높아 적당한 매수가 필요한 ‘스타트업’이라는 이름을 부여하였습니다.

데이터를 읽고 난 후

단위를 통일시키기 위해 표준화 작업을 한다.

그 후 거리측정을 하여 덴드로그램을 그린다.

Cutree를 이용해 네이밍을 한다.

Heatmap을 그린다.