군집화를 활용한 페어트레이딩

황정민

*** Case study 1- Clustering for Pairs Trading**

* 페어트레이딩- 80년대 중반 이래로 알고리즘 트레이더들에 의해 사용된 전략으로, 가격이 역사적으로 같이 움직인 두 자산을 찾아 스프레드를 추적하고, 스프레드가 벌어지면 공통 추세 아래로 떨어진 패자 주식을 매수하고, 승자 주식을 숏하는 전략(일종의 차익거래 전략)
* 만약 pair를 이루는 자산의 관계가 지속된다면 가격이 수렴함에 따라 이익이 발생

* 페어트레이딩의 단점
- 적절한 pair를 찾는 것이 쉽지 않다.
- 만약 두 자산 사이의 관계가 깨진다면 전략이 성립할 수 없다.

* 페어트레이딩은 크게 두 가지 단계로 이뤄진다
1) 형성 스텝: 장기 평균-회귀 관계를 가진 증권을 식별하는 단계. 이상적인 pair는 공통추세로 잘 회귀하면서 높은 분산을 갖고 있어 빈번하게 수익성 있는 거래를 허용해야 한다.
2) 트레이딩 스텝: 가격 움직임에 따라 스프레드가 발산하고 수렴함에 따라 진입과 청산 트레이딩 규칙을 발동

1. 문제 정의
- 페어트레이딩을 위해서 S&P500지수의 구성종목들 중 적절한 pair 찾기
- 2018년 이후의 가격 데이터를 바탕으로 군집화 알고리즘을 사용해 pair를 선정

2. 데이터와 패키지 가져오기

2-1) 패키지 가져오기

2-2) 데이터 가져오기

3. EDA

3-1) 데이터 살펴보기

3-2) 데이터 시각화

4. 데이터 전처리

4-1) 결측치 처리

- 결측치의 비율이 30%가 넘는 기업들의 데이터는 사용하지 않기로 함
- 나머지 데이터들의 결측치는 이전의 값들로 채워 주기로 함.

4-2) 수익률 계산 및 스케일링

- 군집화를 위해서 수익률의 평균과 분산을 사용하기로 함(단위는 1년으로 함)
- StandardScaler를 사용해 스케일링을 진행

	Returns	Volatility
ABT	0.794	-0.703
ABBV	-0.928	0.795

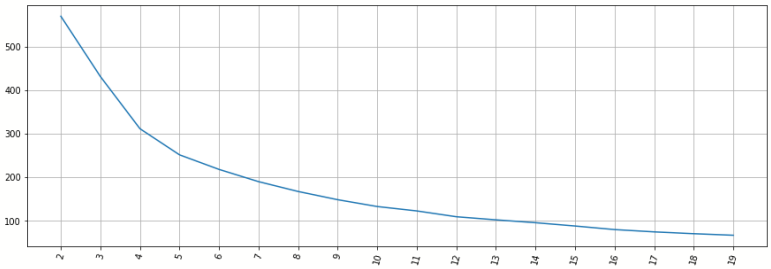
5. 모델 사용하기

5-1) K-Means Clustering

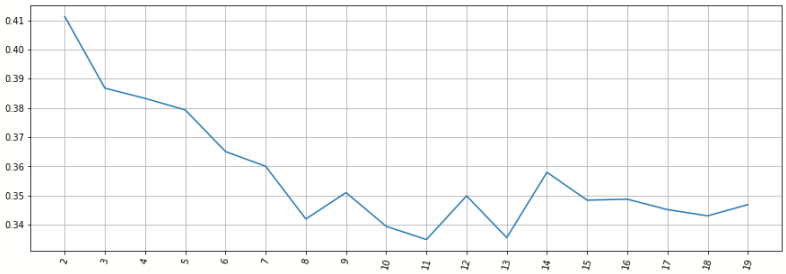
5-1-1) 적절한 군집의 개수 구하기

- SSE 값이 6 개 이후로는 많이 줄어들지 않는다.
- 따라서 군집의 개수를 6 으로 한다.

*** SSE**

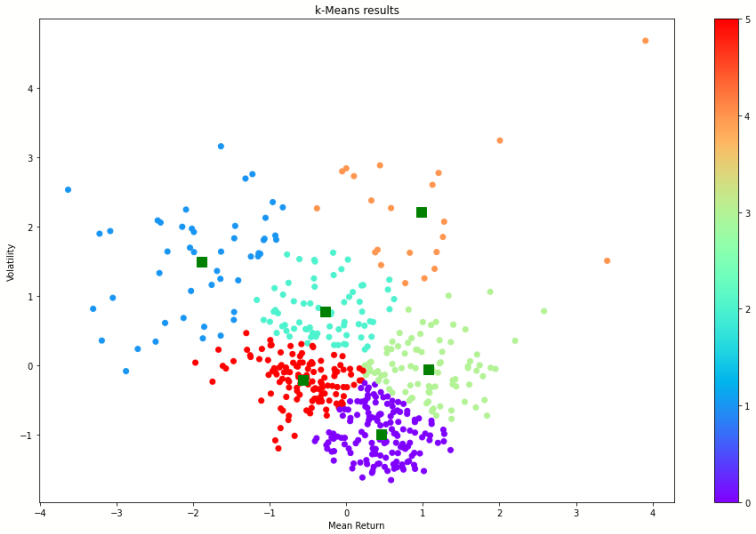


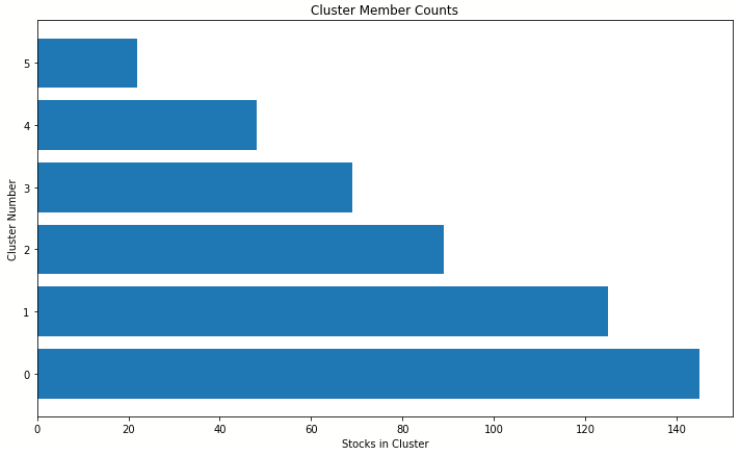
*** 실루엣계수**



5-1-2) 군집화와 시각화

- 6 개의 군집을 생성하고 그 결과를 시각화해보면 다음과 같다.
- 각 군집마다 그 안에 포함 되어있는 주식의 수가 다르다. 가장 적은 경우는 20 개 정도, 가장 많은 경우는 150 개 정도의 주식 데이터를 포함하고 있음을 알 수 있다.





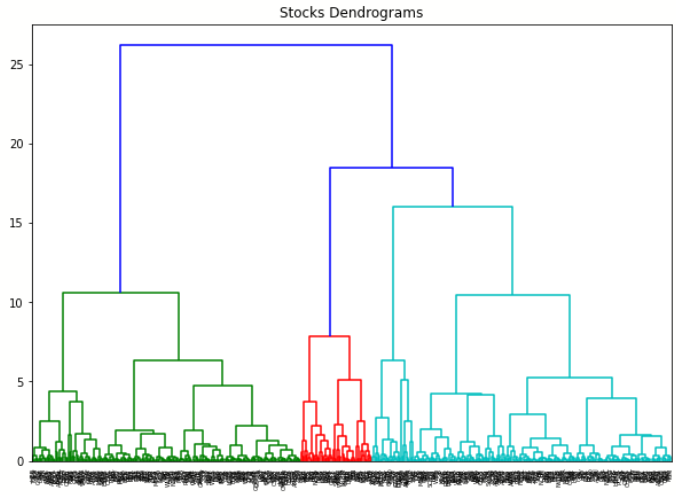
5-2) 계층적 군집화(병합군집)

5.2.1) Dendogram

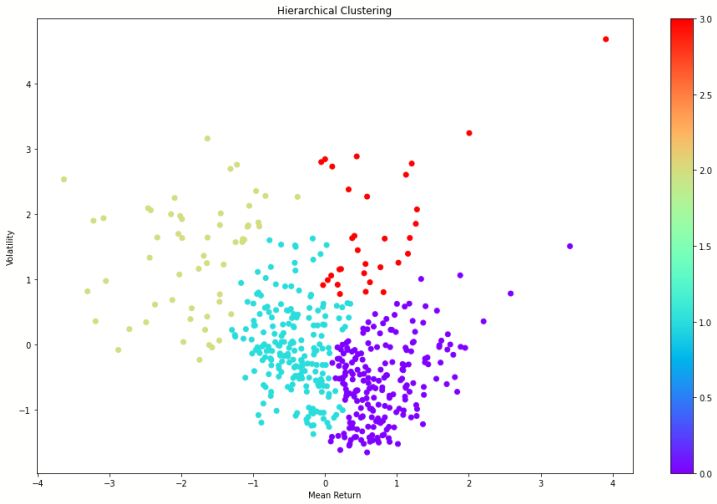
- Dendgrogram 을 활용해 시각화를 진행함(ward linkage 방법 사용)
- 계층적 군집화는 다음의 두가지 경우로 분류 될 수 있다

* Agglomerative nesting: 시작할 때 각 포인트를 하나의 클러스터로 지정, 만족할 때까지 비슷한 두 클러스터를 병합하는 과정을 반복한다.
* Divisive analysis: 하나의 클러스터로 시작해 만족할 때까지 클러스터를 분할한다.

- 13 을 threshold 로 설정하고, 군집화를 실시하면 총 4 개의 군집이 만들어진다.

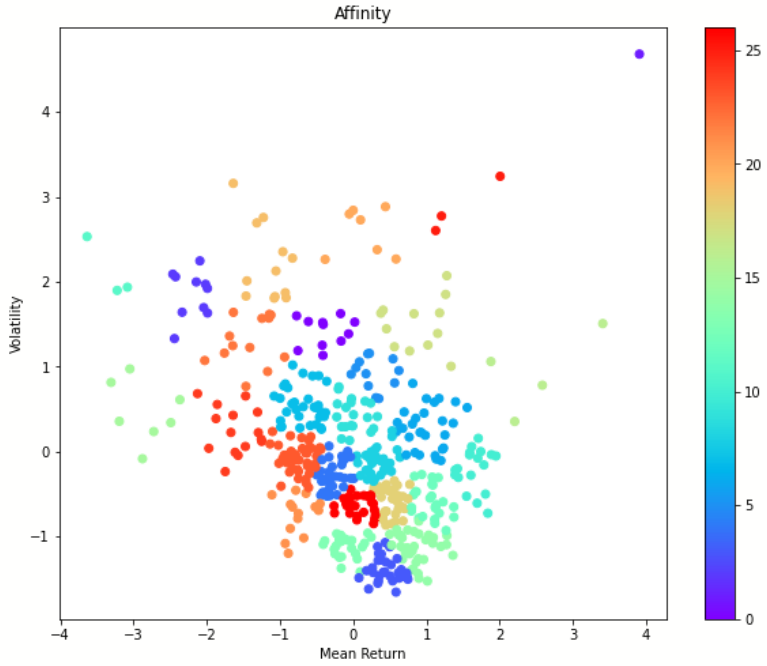


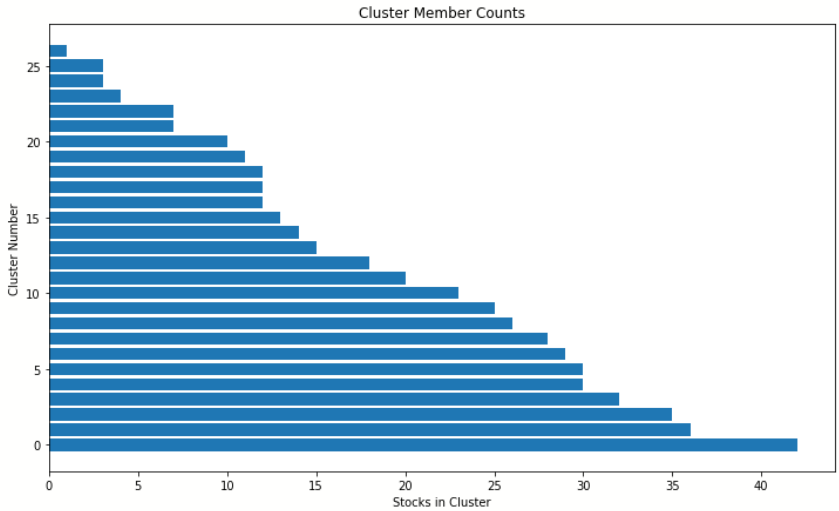
5.2.2) 군집화와 시각화



5-3) Affinity Propagation

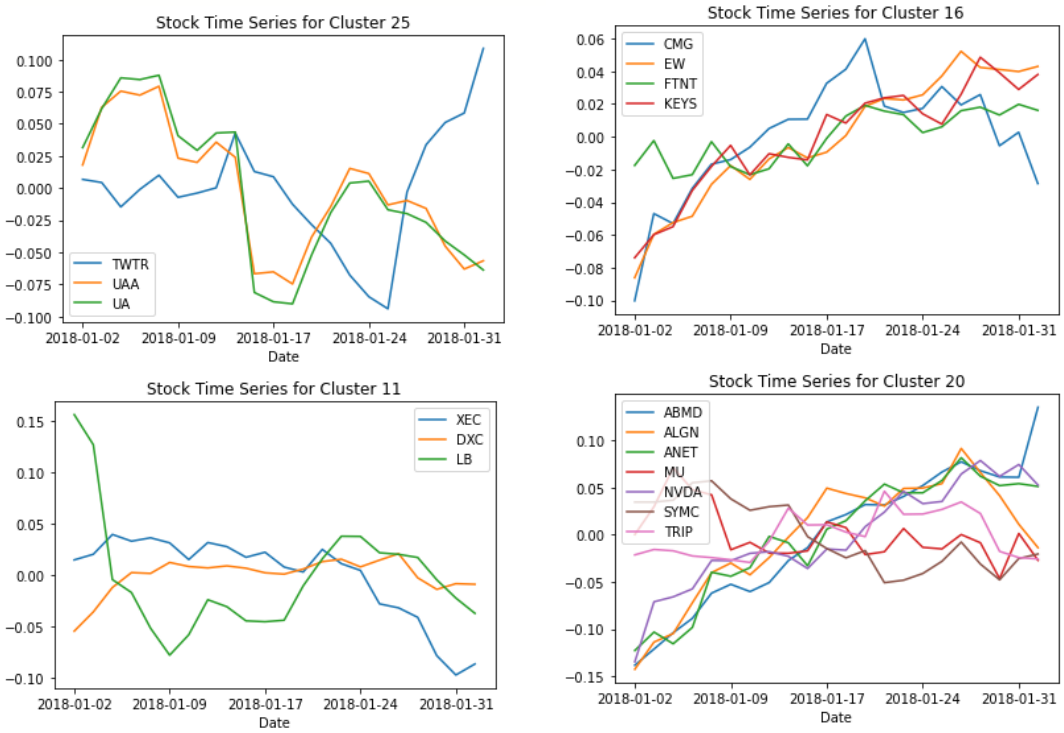
5-3-1) 군집 시각화





5-4) 군집화 평가

- 실루엣 계수를 사용해 군집화를 평가한 결과 Affinity Propagation 를 사용한 것이 가장 군집화가 잘 이뤄졌음을 알 수 있다.(교재 기준)
- 사이즈가 작은 군집들 중 일부를 시각화해보면 아래와 같다.
- 위의 차트를 보면 다른 군집에 있는 주식들 중 비슷한 움직임을 보이는 것들이 존재한다. 이러한 주식들을 활용해 pair 를 만들 수 있다.



cf)공적분이란?

비정상시계열들이 유사한 확률적 추세를 갖고 있는 경우를 말한다. 술취한 사람과 그 사람의 친구가 함께 걸어가는 경우를 생각해보면 쉽게 이해할 수 있다. 술취한 사람과 친구 모두 랜덤하게 움직일 것이다. 즉, 두 사람의 움직임 모두 어떠한 패턴도 보이지는 않을 것이다. 그렇지만 두 사람은 비슷한 움직임을 보일 것이다. 따라서 두 사람 모두 랜덤하게 움직이지만 한 사람의 움직임을 알고 있다면 다른 사람의 움직임을 어느 정도 예측할 수 있다.

한편, 수학적으로 공적분은 다음과 같은 두 가지 성질이 성립하는 경우를 말한다.

비정상시계열이 정상시계열이 되기 위해 필요한 차분 횟수를 적분차수(order of integration)라고 하는데 이 때,

1. 두 개 이상의 시계열들의 적분 차수가 모두 같다.
2. 시계열들의 선형 결합의 적분 차수가 각각의 시계열 사이의 적분 차수보다 낮다.

6 pairs 고르기

6-1) 공적분과 pairs 선정하기

- 군집화 이후에는 공적분을 활용한 방법들이 사용될 수 있다. 두 시계열이 공적분인지 판단하기 위해서는 Johansen test나 Augmented Dickey-Fuller test를 활용할 수 있다.
- 이 실습에서는 파이썬에서 제공하는 coint 함수(엥글&그레인저 공적분 검정)를 사용
- 유의수준을 0.05로 설정하고, pair를 찾으면 32개의 pair를 찾을 수 있다.

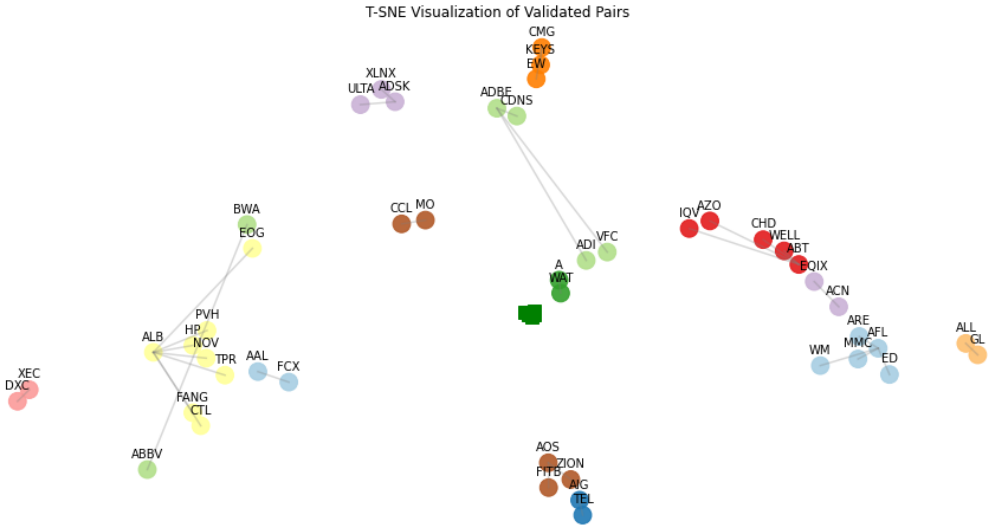
```
Number of pairs found : 32
In those pairs, there are 47 unique tickers.
```

```
[('AOS', 'FITB'),
 ('AOS', 'ZION'),
 ('AIG', 'TEL'),
 ('ABBV', 'BWA'),
 ('ACN', 'EQIX'),
 ('AFL', 'ARE'),
 ('AFL', 'ED'),
 ('AFL', 'MMC'),
 ('AFL', 'WM'),
 ('A', 'WAT'),
 ('ADBE', 'ADI'),
 ('ADBE', 'CDNS'),
 ('ADBE', 'VFC'),
 ('ABT', 'AZO'),
 ('ABT', 'CHD'),
 ('ABT', 'IQV'),
 ('ABT', 'WELL'),
 ('ALL', 'GL'),
 ('MO', 'CCL'),
 ('ALB', 'CTL'),
 ('ALB', 'FANG'),
 ('ALB', 'EOG'),
 ('ALB', 'HP'),
 ('ALB', 'NOV'),
 ('ALB', 'PVH'),
 ('ALB', 'TPR'),
 ('ADSK', 'ULTA'),
 ('ADSK', 'XLNX'),
 ('AAL', 'FCX'),
 ('CMG', 'EW'),
 ('CMG', 'KEYS'),
 ('XEC', 'DXC')]
```

6-2) pairs 시각화하기

T-SNE를 사용해 pair들을 시각화하면 다음과 같다

(교재와 다르게 나온 이유는 random_state를 다르게 설정했기 때문)



7) Conclusion

- 군집화를 활용해 pair를 찾는 방법을 살펴보았다
- pair를 찾은 이후에는 공통추세 아래로 떨어지는 주식을 매수, 공통추세보다 높은 주식을 매도하는 전략을 취한다. (backtest를 통해 전략을 검증하는 과정도 필요)