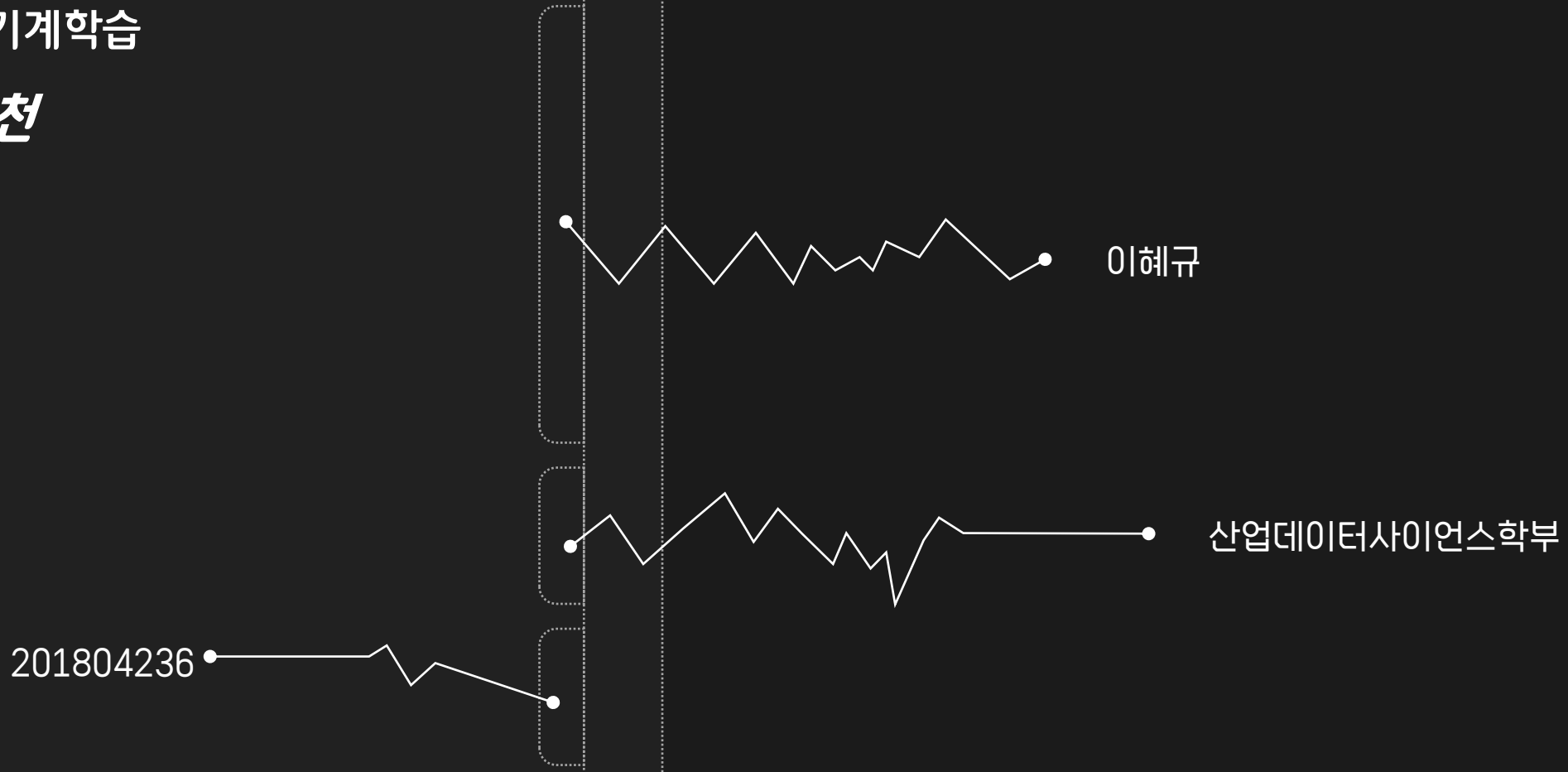
데이터분석과 기계학습

주식종목 추천



INDEX

01

주식 종목 추천

데이터분석과 기계학습

201804236 이혜규

주제

정의

데이터

데이터 이해 및 전처리

데이터 분석

추천 알고리즘 이용

결론

분석 결과 및 결론

마무리

아쉬운 점, 느낀점

주식 종목 추천

- 코스피 200에 등재된 주식 종목을 이용하여 과거 데이터 기반, 유사한 변동을 가진 주식 종목 추천

출처 : 한국거래소(KRX)

- 한국을 대표하는 지수
- 1994년6월15일 주가지수선물 및 옵션의 기초지수로 활용하기 위하여 도입
- 유가증권시장에 상장된 주 중 시장대표성, 산업대표성, 유동성 등 기준으로 선정된 200종목

데이터 출처

- https://github.com/choosunsick/Korea_Stocks
- 기간 : 2000-01-04 ~ 2021-04-16
- 마지막 수집일인 2021년 04월 16일 기준의 코스피 200 데이터

	Date	Open	High	Low	Close	Volume	Adj_Close	name
0	2005-09-29	43517	45553	43466	0	49472	36129	000070
1	2005-09-30	45808	47436	45604	0	72960	37515	000070
2	2005-10-04	47742	47742	46011	0	25256	37393	000070
3	2005-10-05	46927	47131	44841	0	36053	35925	000070
4	2005-10-06	44383	44993	43110	0	38066	34987	000070

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 832352 entries, 0 to 832351
Data columns (total 8 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   date       832352 non-null  object
 1   open       832352 non-null  int64
 2   high       832352 non-null  int64
 3   low        832352 non-null  int64
 4   close      832352 non-null  int64
 5   volume     832352 non-null  int64
 6   adj_close  832352 non-null  int64
 7   name       832352 non-null  object
dtypes: int64(6), object(2)
memory usage: 50.8+ MB
```

데이터 전처리

- 날짜와 관련된 데이터 추가
- 저가와 고과와 관련된 데이터 추가

	name	adj_close	open	high	low	close	differ	avg	volume	date	year	month	day	dayofweek
0	70	74031	84000	84000	79000	82300	1700	2500.0	71250	2011-12-05	2011	12	5	0
1	70	72322	81000	82000	72900	80400	600	4550.0	75973	2011-12-06	2011	12	6	1
2	70	75381	80500	85800	78300	83800	-3300	3750.0	70873	2011-12-07	2011	12	7	2
3	70	71063	83000	83800	78600	79000	4000	2600.0	77794	2011-12-08	2011	12	8	3
4	70	68184	77500	77500	75200	75800	1700	1150.0	54451	2011-12-09	2011	12	9	4

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 807311 entries, 0 to 807310
Data columns (total 14 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   name       807311 non-null  int32
 1   adj_close  807311 non-null  int32
 2   open       807311 non-null  int32
 3   high       807311 non-null  int32
 4   low        807311 non-null  int32
 5   close      807311 non-null  int32
 6   differ     807311 non-null  int32
 7   avg        807311 non-null  float32
 8   volume     807311 non-null  int64
 9   date       807311 non-null  object
 10  year       807311 non-null  int16
 11  month      807311 non-null  int8
 12  day        807311 non-null  int8
 13  dayofweek  807311 non-null  int8
dtypes: float32(1), int16(1), int32(7), int64(1), int8(3), object(1)
memory usage: 47.0+ MB
```

2차 프로젝트까지의 상황

index	알고리즘	결과(RMSE)	향후방향
1	NaiveBayes	10203.81835	추천알고리즘에 적합하지 않다고 판단(제거)
2	KNN	10646.62173	추천알고리즘에 적합하지 않다고 판단(제거)
3	PCA->KNN	11019.41044	추천알고리즘에 적합하지 않다고 판단(제거)
4	최근접이웃 협업 필터링(코사인유사도)		4개 알고리즘 가장 적합하다고 판단(향후 추가 분석 진행)
5	잠재요인 협업필터링(surprise)	199795.16245	추천알고리즘에 적합하지 않다고 판단(제거)

아이템 기반 최근접 협업 필터링

- 사용자와 아이템의 평점 같은 축적된 데이터로 다른 사용자가 사용하지 않은 아이템을 평가하는 방식
- 사용자 기반(user - user), 아이템 기반(item - item)
- 측정 수단으로 유클리드, 코사인, 맨해튼 등의 유사도
- 행은 사용자, 열은 아이템으로된 행렬데이터 사용

데이터 행렬 변환

- 종목명, 수정주가, 날짜만 모아서 피벗테이블로 변경

date	2000-01-04	2000-01-05	2000-01-06	2000-01-07	2000-01-10	2000-01-11	2000-01-12	2000-01-13	2000-01-14	2000-01-17	...	2021-04-05	2021-04-06	2021-04-07	2021-04-08	2021-04-09	2021-04-12	2021-04-13	2021-04-14	2021-04-15	2021-04-16
name																					
70	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	91700.0	92100.0	91200.0	91400.0	91700.0	91300.0	91500.0	92700.0	97700.0	98200.0
80	3338.0	3097.0	3105.0	2957.0	3179.0	3253.0	3223.0	3171.0	3216.0	3356.0	...	36250.0	36650.0	36350.0	35500.0	35900.0	36200.0	36100.0	35550.0	35100.0	35350.0
100	23510.0	23074.0	21562.0	22570.0	22738.0	22906.0	22368.0	21965.0	21965.0	22167.0	...	62800.0	61800.0	62700.0	62300.0	63100.0	65800.0	65900.0	64600.0	65800.0	67000.0
120	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	180500.0	183500.0	183000.0	179000.0	178000.0	178000.0	175000.0	176000.0	174500.0	179500.0
150	20558.0	20789.0	19557.0	19980.0	19826.0	19557.0	20019.0	18556.0	19480.0	19172.0	...	49750.0	49750.0	50300.0	49700.0	49500.0	49500.0	49250.0	49900.0	51600.0	52000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
282330	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	159000.0	160500.0	161000.0	159000.0	162000.0	163500.0	162000.0	162500.0	159500.0	0.0
284740	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	42000.0	41700.0	42300.0	42000.0	41900.0	41950.0	42400.0	43900.0	43600.0	0.0
285130	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	259000.0	251000.0	250000.0	255000.0	260000.0	263500.0	256797.0	277000.0	274500.0	0.0
294870	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	29350.0	29200.0	29400.0	29800.0	29100.0	28450.0	27950.0	28500.0	28200.0	0.0
316140	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	10050.0	10100.0	10650.0	10600.0	10300.0	10400.0	10350.0	10400.0	10400.0	0.0

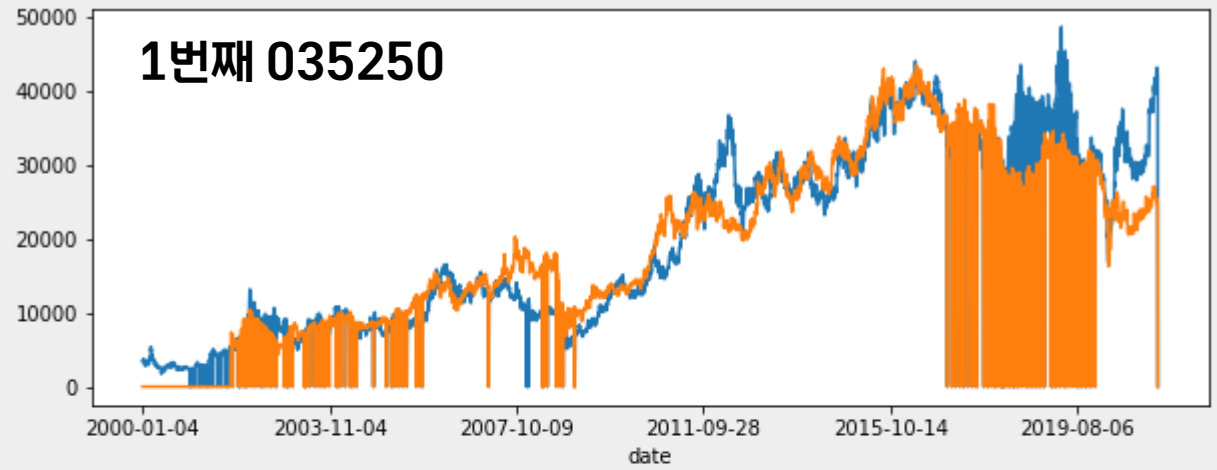
200 rows × 5432 columns

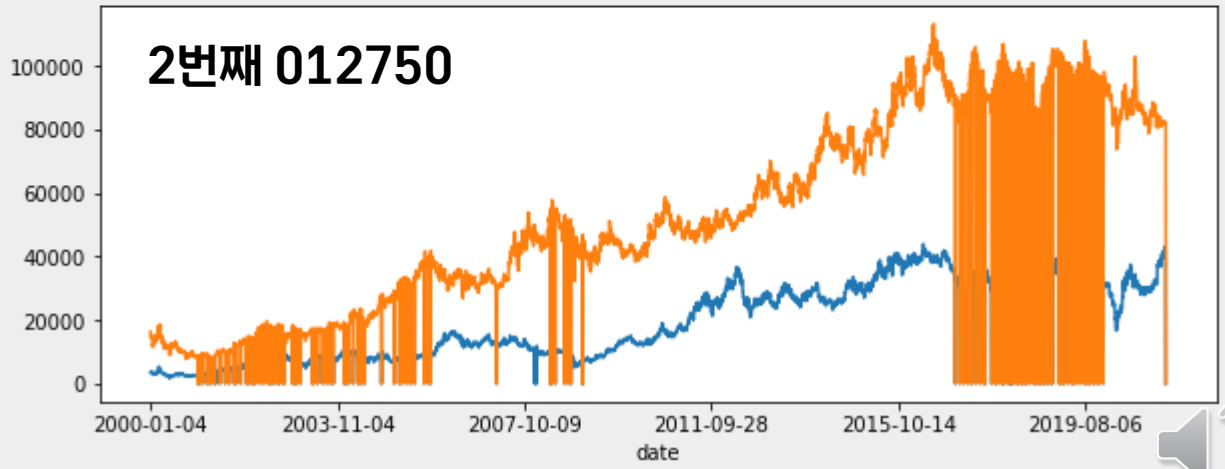
2차 프로젝트. Pairwise_코사인 유사도

```
item_cs_df[20000].sort_values(ascending=False)[1:6]
```

```
name
35250    0.981656
12750    0.978017
10130    0.973473
5830     0.970318
33780    0.968675
Name: 20000, dtype: float64
```

20000

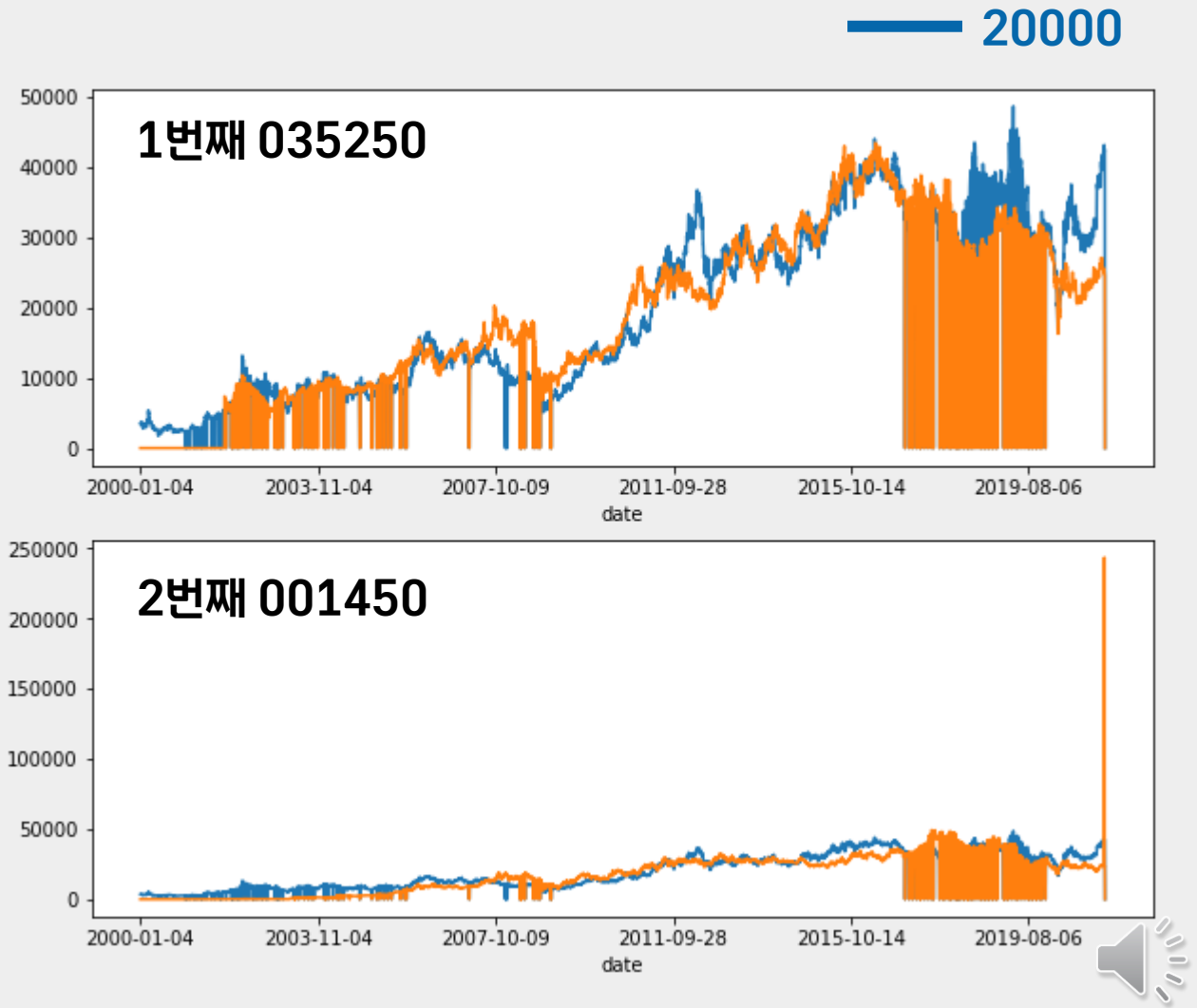




2차 프로젝트. Pairwise_유클리드 거리

```
item_ed_df[20000].sort_values(ascending=True)[1:6]
```

```
name
35250    324497.79
1450     489934.55
1680     615976.48
69260    673997.57
9830     700198.55
Name: 20000, dtype: float64
```

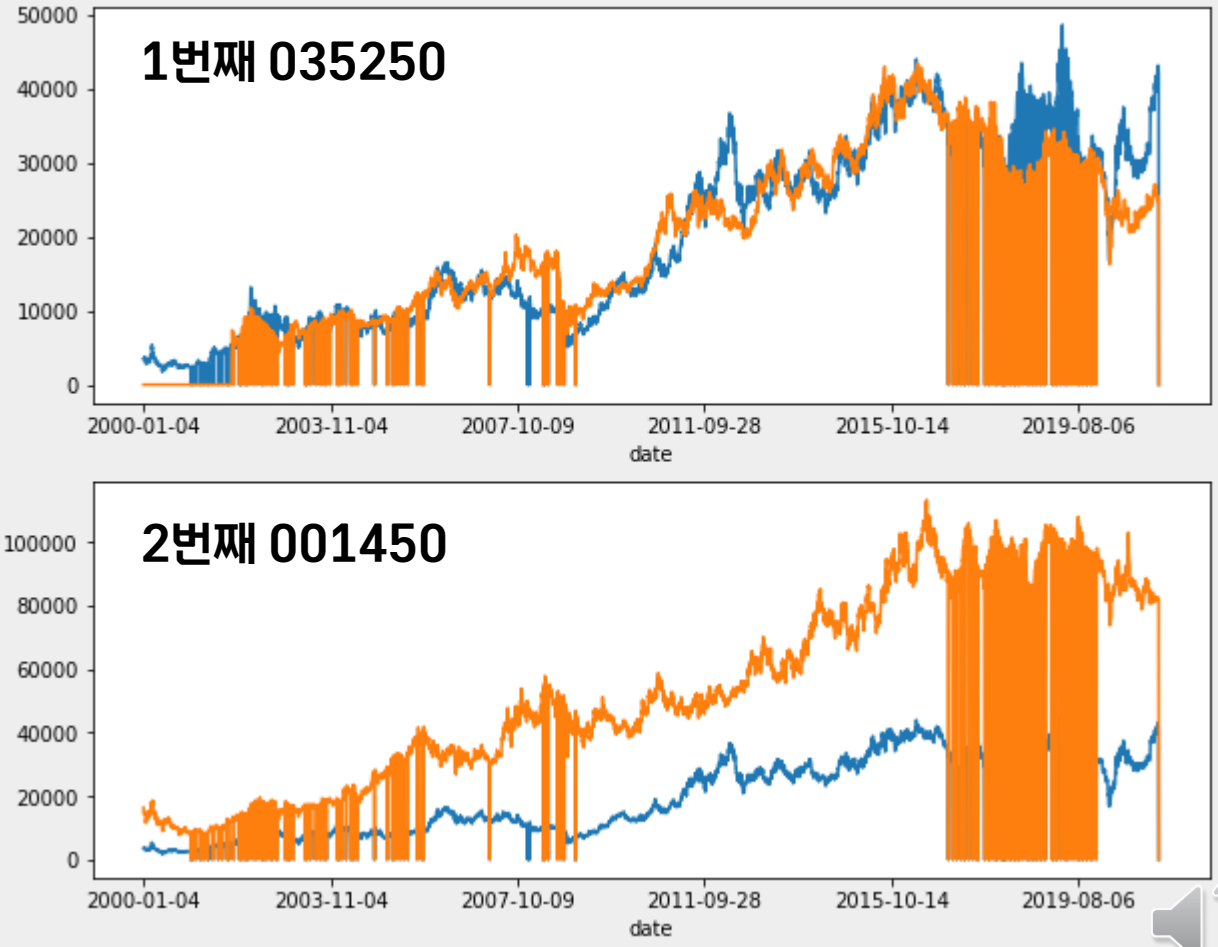


2차 프로젝트. Pairwise_맨해튼 거리

```
item_mt_df[20000].sort_values(ascending=True)[1:6]
```

```
name
35250    16320436.00
1450     27067584.00
1680     36843555.00
9830     38043045.00
69260    39851265.00
Name: 20000, dtype: float64
```

20000

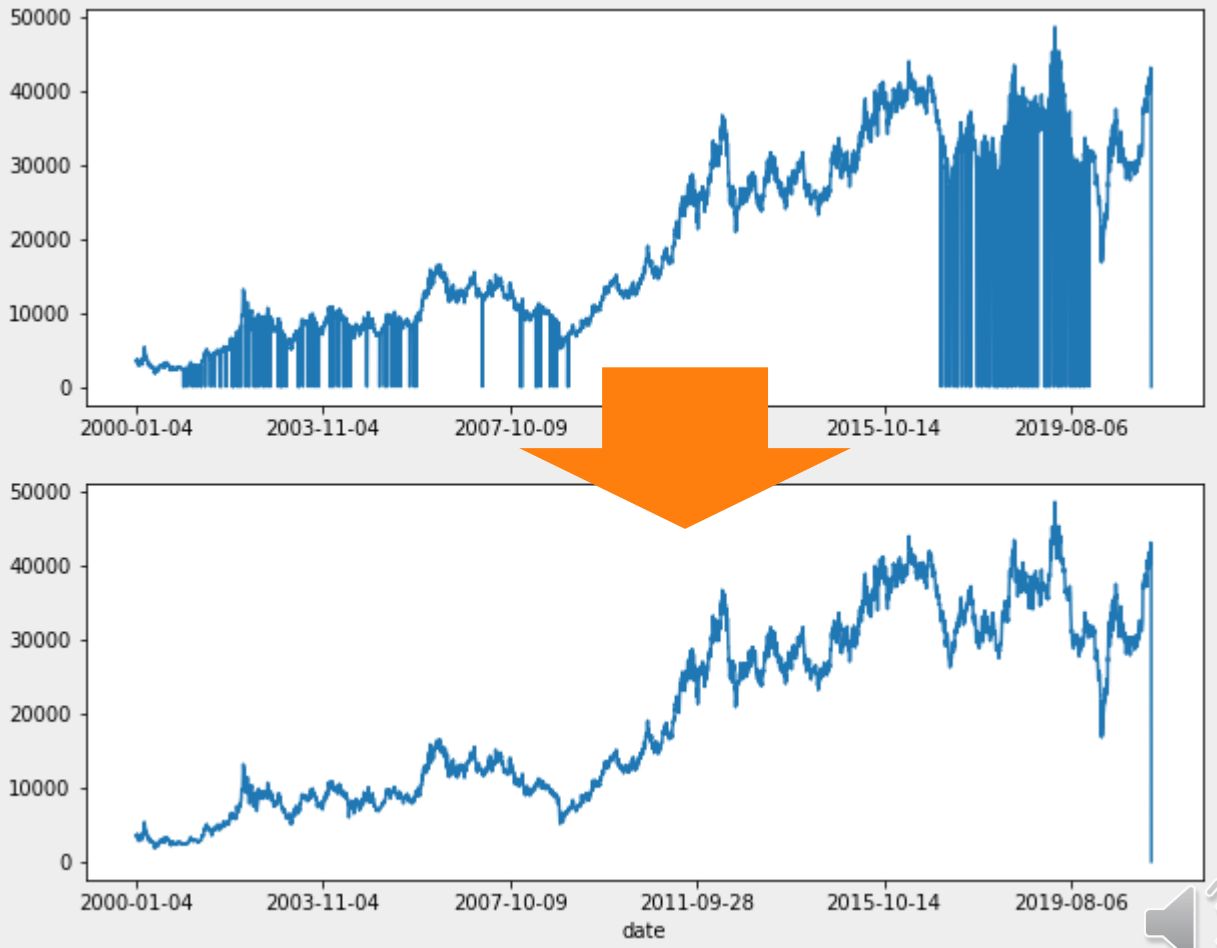


3차 프로젝트. 데이터 추가 전처리

- NA값을 0으로 처리한게 그래프상으로 깔끔하지 않음
- 0으로 처리 된 부분을 해당 주가의 앞과 뒤의 값 중간값으로 변경

```
mid_df=new_df.copy()

for i in range(len(mid_df.columns)):
    for j in range(1, len(mid_df.index)-1):
        if mid_df.iloc[j, i] == 0:
            mid_df.iloc[j, i] = (mid_df.iloc[j-1, i] + mid_df.iloc[j-1, i])/2
```

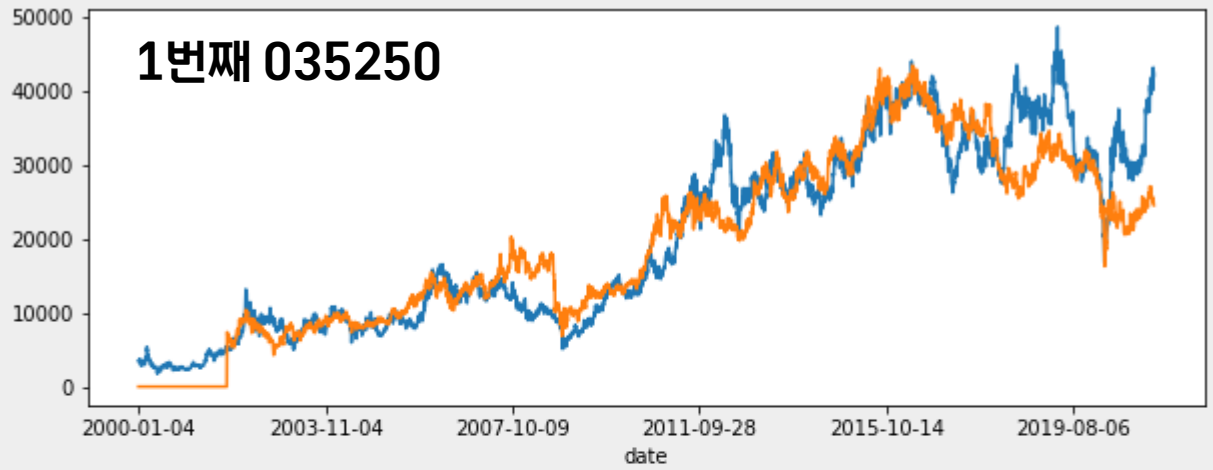


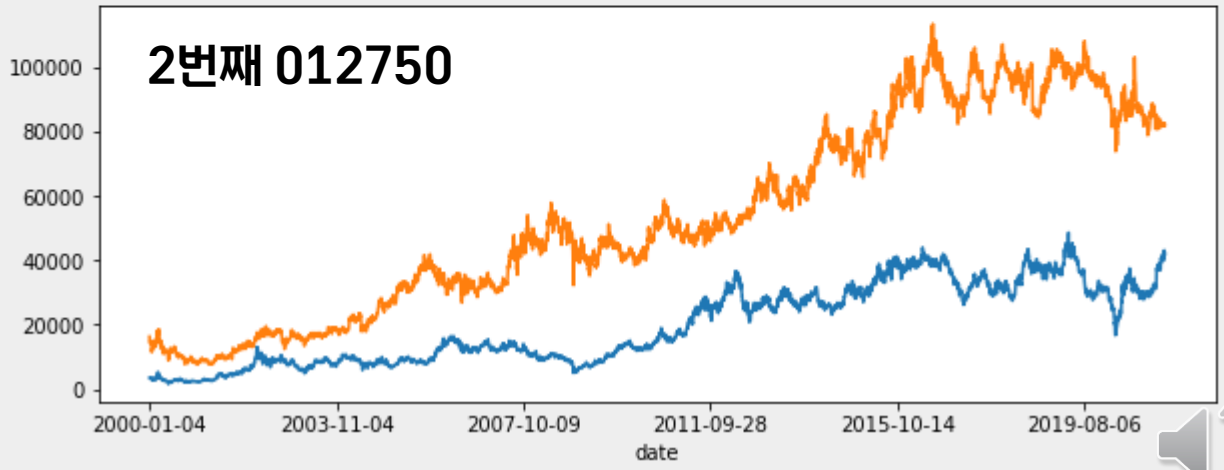
3차 프로젝트. Pairwise_코사인 유사도

```
item_cs_df1[20000].sort_values(ascending=False)[1:6]
```

```
name
35250   0.98
12750   0.98
10130   0.97
33780   0.97
5830    0.97
Name: 20000, dtype: float64
```

20000

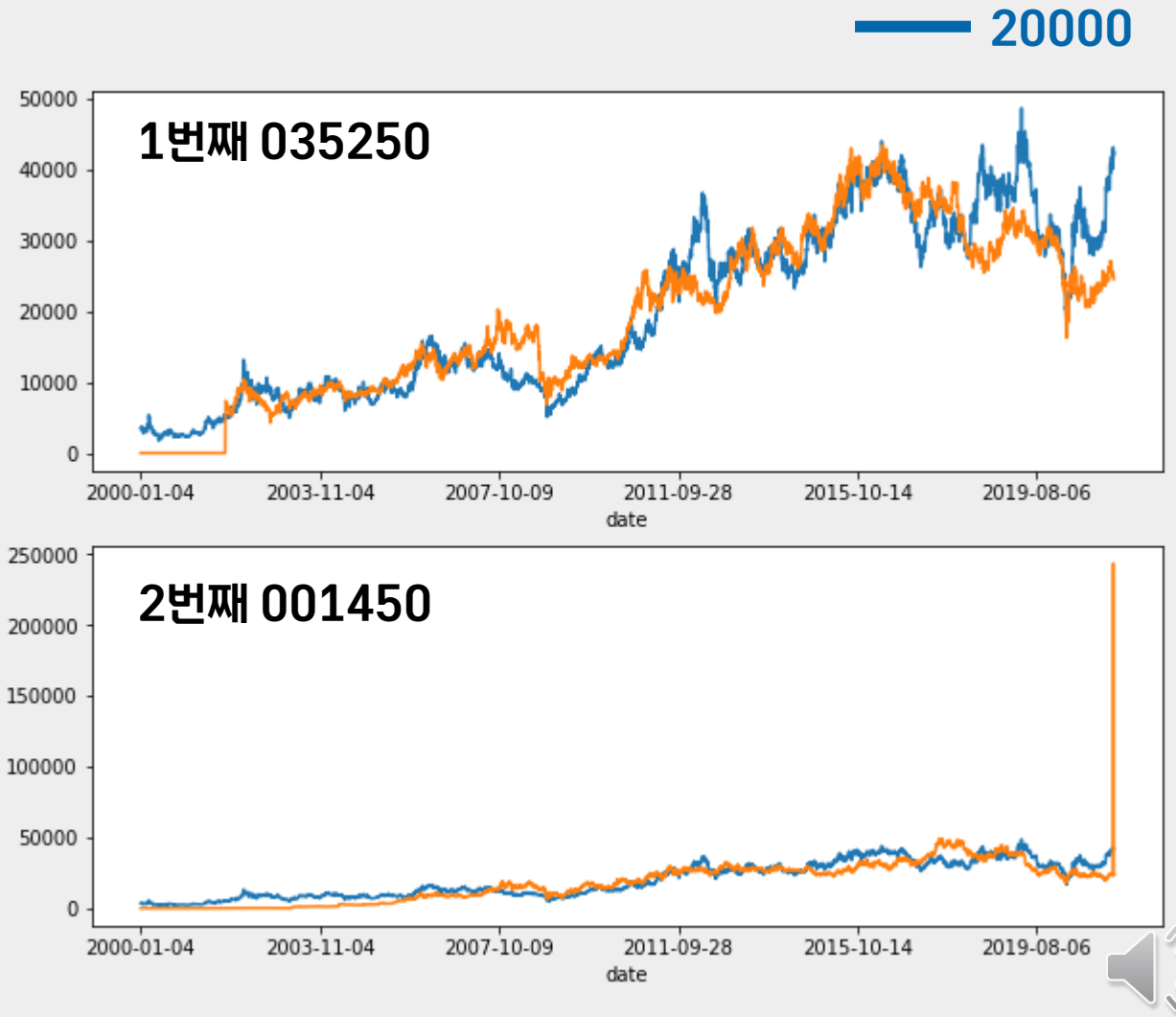




3차 프로젝트. Pairwise_유클리드 거리

```
item_ed_df1[20000].sort_values(ascending=True)[1:6]
```

```
name
35250    317759.63
1450     488853.87
1680     610721.59
69260    681442.60
5250     701731.46
Name: 20000, dtype: float64
```

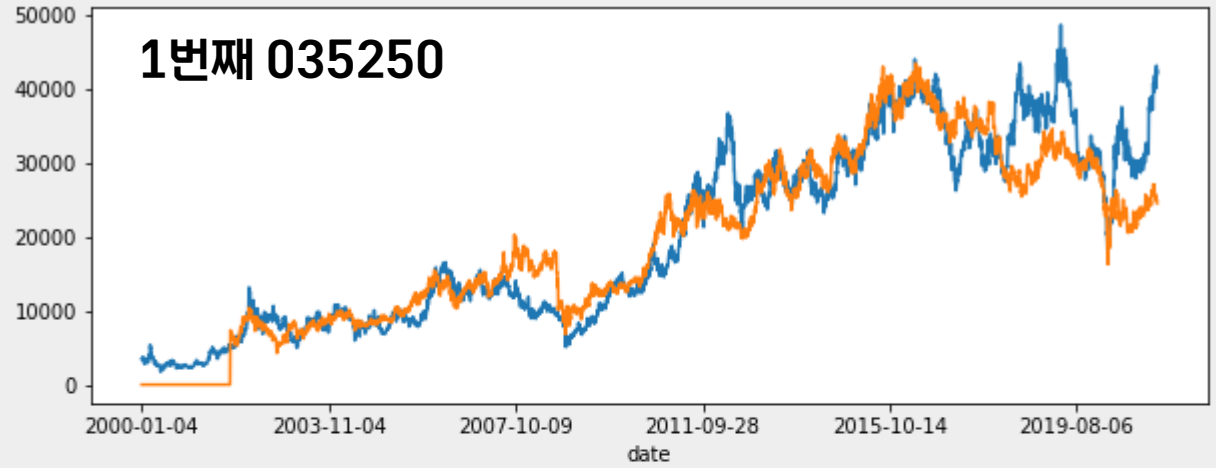


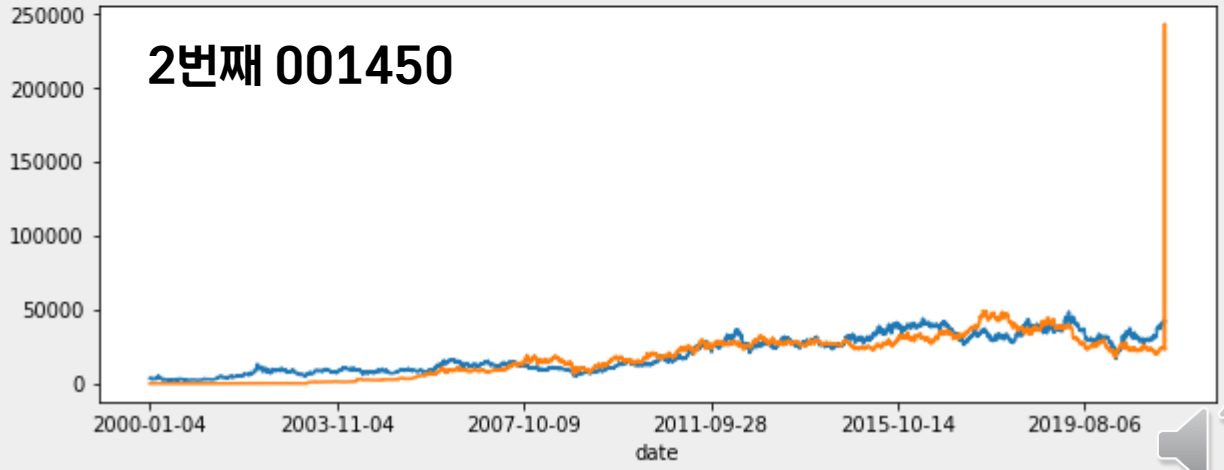
3차 프로젝트. Pairwise_맨해튼 거리

```
item_mt_df1[20000].sort_values(ascending=True)[1:6]
```

```
name
35250    16634539.00
1450     27844283.00
1680     37634867.00
9830     38491049.00
69260    41033688.00
Name: 20000, dtype: float64
```

20000





주식 종목 추천

데이터분석과 기계학습

201804236 이혜규

[2차프로젝트] 020000(스타일 플랫폼)과 유사한 종목

코사인유사도

1순위는 카지노 전문업체

2순위는 안심솔루션

35250	0.981656
12750	0.978017
10130	0.973473
5830	0.970318
33780	0.968675

유클리드 거리

1순위는 카지노 전문업체

2순위는 보험사

35250	324497.79
1450	489934.55
1680	615976.48
69260	673997.57
9830	700198.55

맨해튼 거리

1순위는 카지노 전문업체

2순위는 보험사

35250	16320436.00
1450	27067584.00
1680	36843555.00
9830	38043045.00
69260	39851265.00

020000(스타일 플랫폼)과 유사한 종목 **[3차프로젝트]**

코사인유사도

1순위는 카지노 전문업체

2순위는 안심솔루션

35250	0.98
12750	0.98
10130	0.97
33780	0.97
5830	0.97

유클리드 거리

1순위는 카지노 전문업체

2순위는 보험사

35250	317759.63
1450	488853.87
1680	610721.59
69260	681442.60
5250	701731.46

맨해튼 거리

1순위는 카지노 전문업체

2순위는 보험사

35250	16634539.00
1450	27844283.00
1680	37634867.00
9830	38491049.00
69260	41033688.00

1순위는 3개 알고리즘 모두 같은 종목, 2순위부터 달라진다. 특히 유클리드와 맨해튼 알고리즘은 비슷했고 4,5 순위부터 달라졌다.

주식 종목 추천

데이터분석과 기계학습

201804236 이혜규

[2차프로젝트] 035720(포털 및 정보매개 서비스업)과 유사한 종목

코사인유사도

1순위는 축전지 제조업

2순위는 모빌리티·반도체

6400	0.953749
11790	0.949308
51910	0.935552
36570	0.932560
6280	0.915330

유클리드 거리

1순위는 제약회사

2순위는 종합부품 제조회사

6280	3755225.82
9150	4229703.07
11070	4476415.80
11780	4685308.85
68270	4712554.73

맨해튼 거리

1순위는 종합부품 제조회사

2순위는 원유 정제회사

9150	180393556.00
10950	204515804.00
33780	205949353.00
12750	208477023.00
6280	211046025.00

035720(포털 및 정보매개 서비스업)과 유사한 종목 **[3차프로젝트]**

11790	0.96
6400	0.96
51910	0.94
36570	0.93
6280	0.92

코사인유사도

1순위는 모빌리티·반도체

2순위는 축전지 제조업

6280	3773041.87
9150	4248212.31
11070	4509176.15
11780	4712979.34
68270	4817038.37

유클리드 거리

1순위는 제약회사

2순위는 종합부품 제조회사

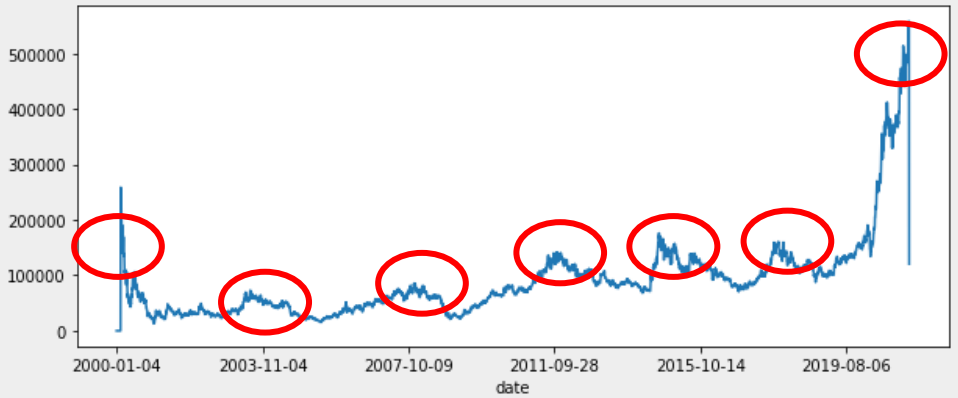
9150	181308628.00
10950	206448014.00
33780	206622508.00
12750	212371539.00
6280	216209914.00

맨해튼 거리

1순위는 종합부품 제조회사

2순위는 원유 정제회사

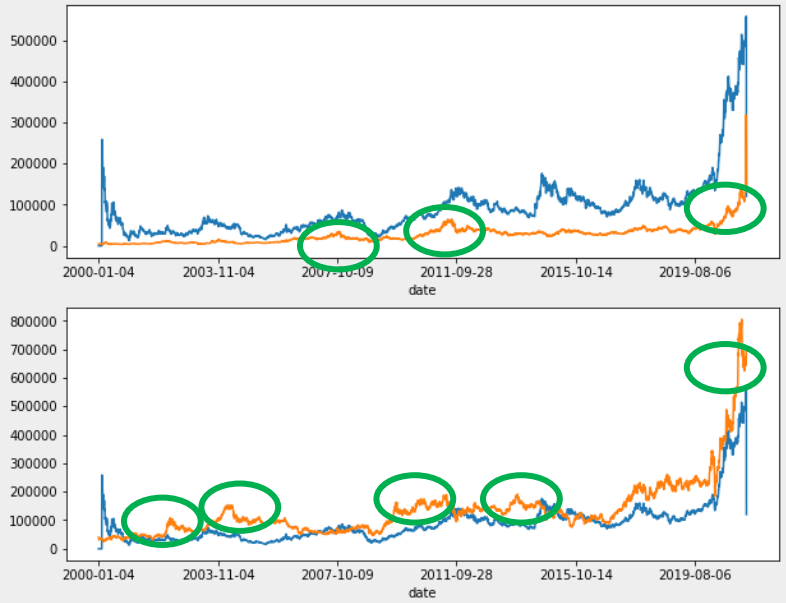
1순위는 3개 알고리즘 모두 다르다. 세 알고리즘이 각 다른 분석을 진행할때는 어떤 종목을 추천해야할까?



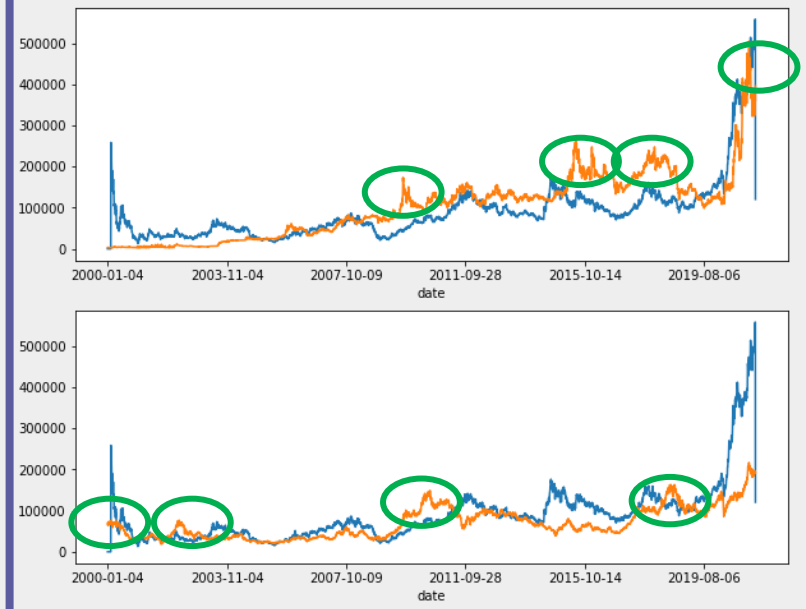
1. 각 1,2,3 순위에서 가장 많이 겹치는 종목 추천

2. 시각화로 평가

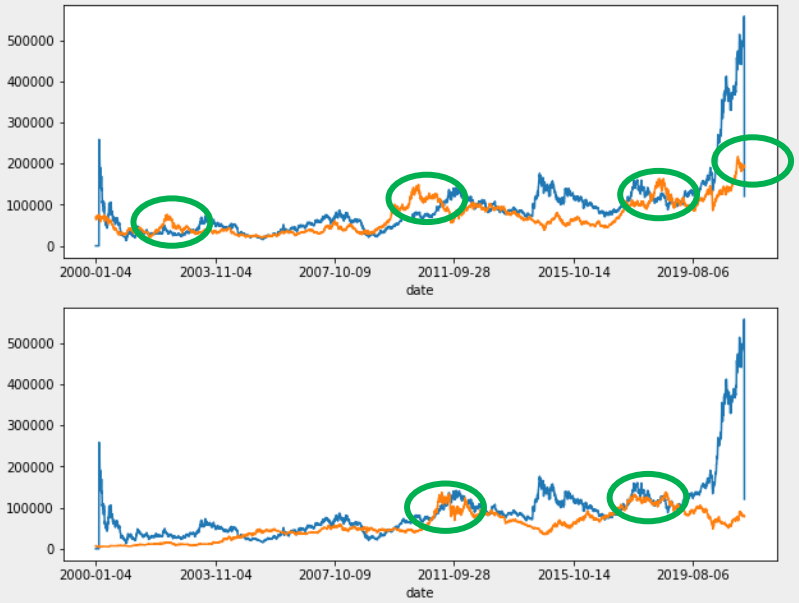
- 기존의 종목은 7개의 상한가 변동이 있다.



코사인



유클리드



맨해튼

느낀점

✓ 업종이 비슷한 종목끼리 추천으로 뜰 것으로 예상했다.

- 시가 총액의 규모가 비슷한 종목끼리 묶였다.

✓ 다른 수치적인 결과를 보는 분석보다 그래프로 보여 재밋는 분석이였다.

감사합니다