분석 보고서 - 민경

서론

- **분석 주제**
  - 유튜브 데이터(영상개수, 조회수, 좋아요수 등) 와 주가 간의 상관성 분석
- **주제 선정 배경**
  - 최근 국제정서에 크게 영향을 미치는 사건들이 발생하면서 주가 또한 크게 변하고 있음 . 주가는 이런 사건들에 민감하게 반응하는 경향이 있고 , 유튜브 또한 이러한 사건 발생 시 업로드 양이 증가하는 경향이 있으므로 두 데이터 간의 상관성이 있을 것이라 생각해 상관성 분석을 진행해보고자 함.
- **가정**
  - 유튜브 데이터와 주식 데이터 간의 상관 관계가 있을 것이다.

분석 과정

분석 대상 종목 선정

- 최근 국제 정서에 가장 큰 영향을 미친 사건으로 '코로나' 와 '러시아-우크라이나 전쟁' 을 선정
- 코로나 - '항공' 관련 주, 러시아-우크라이나 전쟁 - '석유/가스' 관련 주 로 분석 대상 선정
- 각 카테고리에서 시가 총액이 높은 기업들, 그 중에서도 경향성이 비슷한 주들로 선정 - 나중에 EDA 관련 내용 추가

▼ 항공주

- 대한항공 - 003490
- 제주항공 - 089590
- 진에어 - 272450
- 에어부산 - 298690

▼ 석유/가스 주

1. SK이노베이션 - 096770
2. S-Oil - 010950
3. HD현대 - 267250
4. GS - 078930
5. LX인터내셔널 - 001120

유튜브 검색 키워드 및 필터 설정

유튜브 데이터 수집

상관성 분석

▼ 키워드 : 항공 +코로나 -스포츠 / 필터 : 관련성

	viewCount	likeCount	favoriteCount	commentCount	key	title	channel	url	date
0	4826	35.00000	0	17.00000	항공 +코로나 -스포츠	"다시 하늘 날고 싶어요"... 코로나 1년 벼랑 끝 항공산업 / YTN	YTN	https://www.youtube.com/watch?v=L-0_i8hiZbY	2021-01-19
1	40679	453.00000	0	32.00000	항공 +코로나 -스포츠	[다큐온] 🛩국내 공항과 항공, 여행업계의 새로운 도전의 현장 "코로나를...	KBS 다큐	https://www.youtube.com/watch?v=SPNXUaSEfMU	2021-12-03
2	7578	90.00000	0	4.00000	항공 +코로나 -스포츠	코로나19의 직격타를 맞은 항공업계의 현황과 향후 전망 🚲	JOB+채용	https://www.youtube.com/watch?v=iX0cxWiSJ9M	2021-03-05

항공+코로나-스포츠 관련성 순 유튜브 데이터

유튜브 데이터 전처리

1. likeCount, favoriteCount 중 likeCount 사용
   a. favorite Count - 모든 값이 0임

```
youtube['favoriteCount'].value_counts() # 0만 존재
```

```
0    592
Name: favoriteCount, dtype: int64
```

2. commentCount열 결측치 처리
   a. url을 통해 확인한 결과 댓글 사용이 중지된 영상의 경우 Null값으로 처리됐음을 알 수 있음 → 0으로 대체

	viewCount	likeCount	commentCount	key	title	channel	url	date
180	689	13.00000	NaN	항공 +코로나 -스포츠	대한항공(003490) 코로나 이후, 수혜 가능 커	토마토증권통 TomatoTV	https://www.youtube.com/watch?v=iZrbxQGuY4U	2020-11-17
256	8326	114.00000	NaN	항공 +코로나 -스포츠	[대한항공 뉴스룸] 우리는 '연결' 합니다.	대한항공 뉴스룸	https://www.youtube.com/watch?v=eGIlPIGBsw8	2020-05-19
297	6347	192.00000	NaN	항공 +코로나 -스포츠	[수익극대화 포트전략 이창원] 위드 코로나 대장 아시아나항공! 일진다이아 대응!	MBN골드	https://www.youtube.com/watch?v=qw2wMNB23e0	2021-08-24
300	6347	192.00000	NaN	항공 +코로나 -스포츠	[수익극대화 포트전략 이창원] 위드 코로나 대장 아시아나항공! 일진다이아 대응!	MBN골드	https://www.youtube.com/watch?v=qw2wMNB23e0	2021-08-24
308	374	NaN	NaN	항공 +코로나 -스포츠	급등주 - 대한항공우, 항공, 여행, 코로나, 실적, 주식투자, 삼성전자, 재테크,...	평택존놈TV	https://www.youtube.com/watch?v=oxXiLA5IbeM	2022-11-04
414	430	27.00000	NaN	항공 +코로나 -스포츠	대한항공 / 파버나인 / 지니뮤직 / 마이크로디지탈 / 여행주 / 코로나 / 항공주...	오공주차트룸	https://www.youtube.com/watch?v=GPno_1I2AWA	2021-05-24
425	275	9.00000	NaN	항공 +코로나 -스포츠	[제주항공] 위드코로나 기대/항공주 기술적 반등!? 과감한 매도가 수익을 만든다	영스톡	https://www.youtube.com/watch?v=bbUu_37RYTY	2022-02-08
446	1114	44.00000	NaN	항공 +코로나 -스포츠	한국공항 코로나 백신효과로 항공지상 조업 수요 회복 및 대한항공 아시아나항공 합병 효...	태린이아빠	https://www.youtube.com/watch?v=7KorBKBweDA	2021-06-01
478	256	8.00000	NaN	항공 +코로나 -스포츠	[대한항공 주가] 위드코로나 대장주 자리매김, 탄소중립 항공유 도입 긍정적 #대한항...	더블유경제TV	https://www.youtube.com/watch?v=ysbyDwjFtZo	2021-09-14
496	48605	1183.00000	NaN	항공 +코로나 -스포츠	베트남 코로나 확진자 85명 이제는 걷잡을 수 없나?베트남 항공은 왜 환불 안해줘?(...	CONG TUBE 2020	https://www.youtube.com/watch?v=2EGHXzla90w	2020-03-20

3. 일자별로 그룹화 후 영상 개수(num_video) 열 추가

```
# 일자별로 그룹화
youtube_date = youtube.groupby(by='date').sum()
youtube_date['num_video'] = youtube.groupby(by='date').size()
youtube_date.head(3)
```

date	viewCount	likeCount	commentCount	num_video
2020-01-22	854	4.00000	5.00000	1
2020-01-29	90	3.00000	0.00000	1
2020-02-01	1144	12.00000	4.00000	2

4. 주가 데이터 기준 유튜브 영상이 존재하지 않는 날 결측치 보간
    a. 사람들의 관심도가 영상이 올라온 이후로 점차 선형적으로 증감한다고 가정하여 선형 보간법을 활용해 결측치 보간
        - 결측치 처리 전

	Date	viewCount	likeCount	commentCount	num_video
23	2020-02-06	239.00000	0.00000	0.00000	1.00000
24	2020-02-07	NaN	NaN	NaN	NaN
25	2020-02-10	7350.00000	87.00000	6.00000	5.00000
26	2020-02-11	NaN	NaN	NaN	NaN
27	2020-02-12	138.00000	3.00000	0.00000	2.00000
28	2020-02-13	714.00000	3.00000	6.00000	1.00000
29	2020-02-14	2905.00000	34.00000	18.00000	1.00000

        - 결측치 처리 후

	Date	viewCount	likeCount	commentCount	num_video
23	2020-02-06	239.00000	0.00000	0.00000	1.00000
24	2020-02-07	3794.50000	43.50000	3.00000	3.00000
25	2020-02-10	7350.00000	87.00000	6.00000	5.00000
26	2020-02-11	3744.00000	45.00000	3.00000	3.50000
27	2020-02-12	138.00000	3.00000	0.00000	2.00000
28	2020-02-13	714.00000	3.00000	6.00000	1.00000
29	2020-02-14	2905.00000	34.00000	18.00000	1.00000

    b. 초기 데이터는 이전 데이터가 존재하지 않아 결측치 보간이 어렵기 때문에 제거
5. 조회 수, 좋아요 수, 댓글 수 → 모두 더한 youtube_performance(영상반응)열 생성

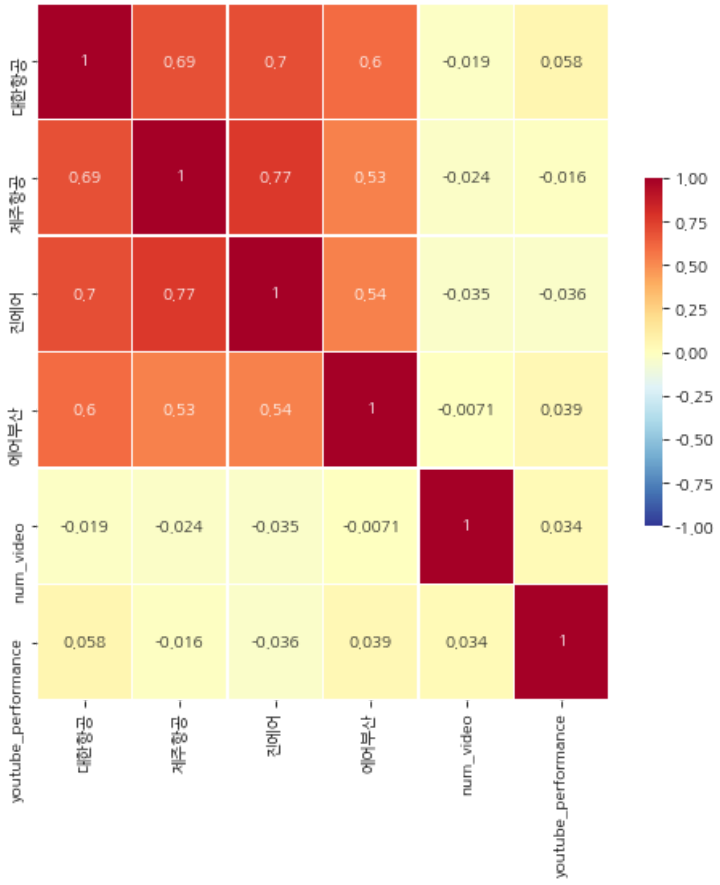
```
youtube_linear.head(3)
```

	Date	num_video	youtube_performance
14	2020-01-22	1.00000	863.00000
15	2020-01-23	1.00000	606.33333
16	2020-01-28	1.00000	349.66667

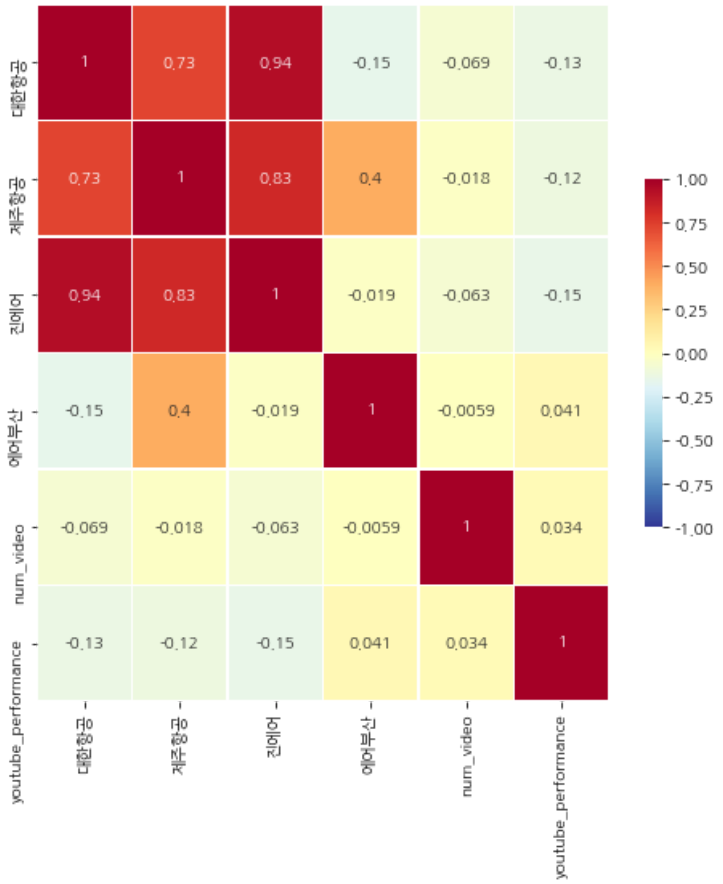
상관성 분석

- 유튜브 데이터와 주식 데이터 통합 (inner join) 해 데이터 프레임 생성
- 등락률, 종가, 거래량 총 3개의 데이터 프레임 생성

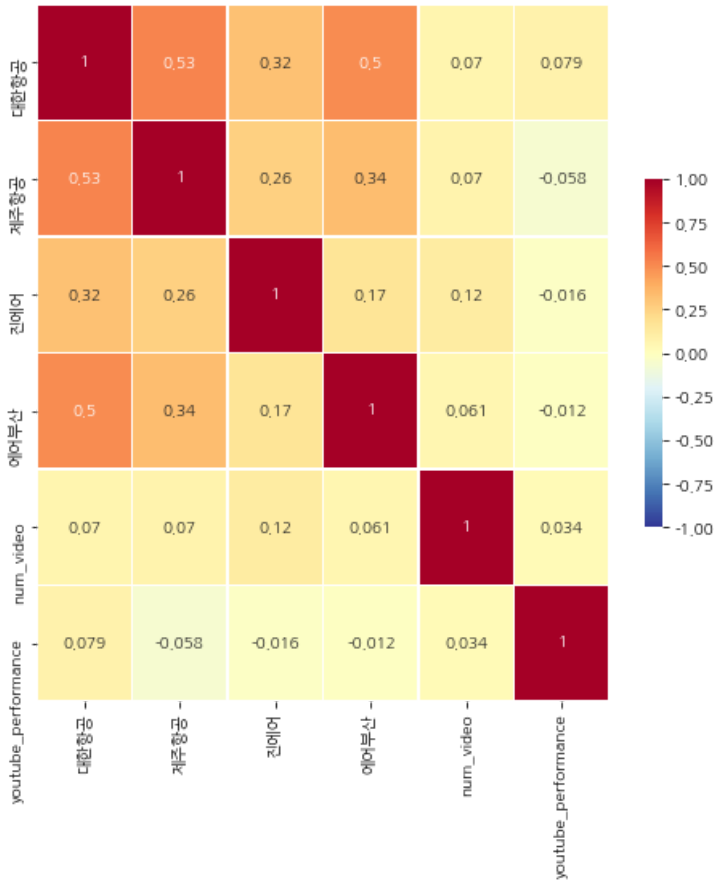
1. 유튜브 데이터 - 등락률



2. 유튜브 데이터 - 종가


대한항공
제주항공
진에어
에어부산
num_video
youtube_performance
1
0.73
0.94
-0.15
-0.069
-0.13
0.73
1
0.83
0.4
-0.018
-0.12
0.94
0.83
1
-0.019
-0.063
-0.15
-0.15
0.4
-0.019
1
-0.0059
0.041
-0.069
-0.018
-0.063
-0.0059
1
0.034
-0.13
-0.12
-0.15
0.041
0.034
1
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

3. 유튜브 데이터 - 거래량



▼ 키워드 : 국제유가 / 필터 : 조회수

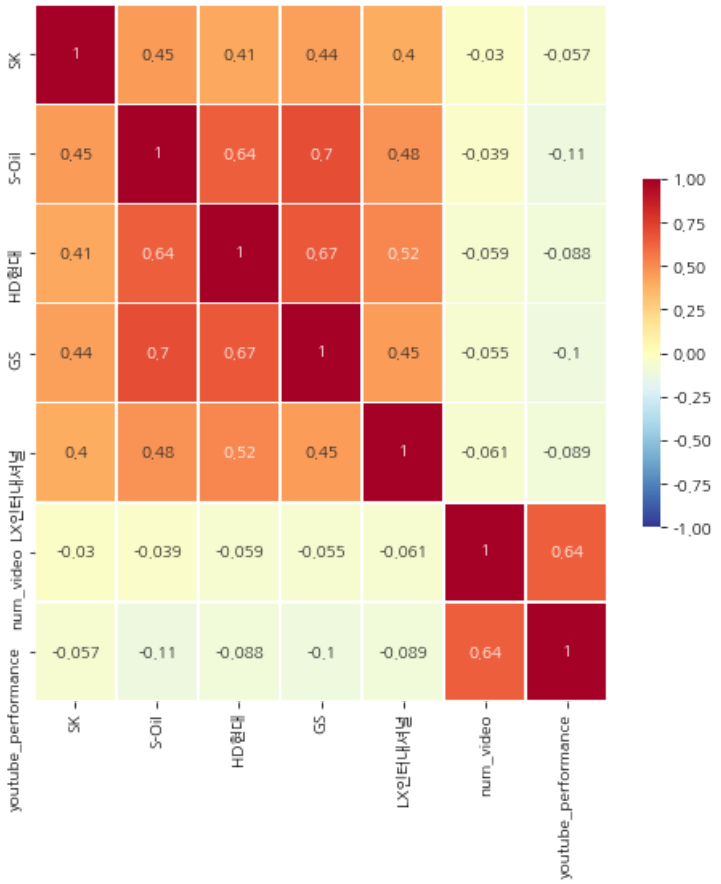
	viewCount	likeCount	favoriteCount	commentCount	key	title	channel	url	date
0	1773801	11583.00000	0	3050.00000	국제유가	밤사이 국제유가 '폭락'..WTI 100달러 붕괴 (2022.07.06...	MBCNEWS	https://www.youtube.com/watch?v=vnl2xFjWUz8	2022-07-06
1	786038	4538.00000	0	386.00000	국제유가	국제유가를 흔들 수 있는 원유매장량이 많은 나라들	보통남자	https://www.youtube.com/watch?v=_LD-qsg_R9I	2021-06-11
2	530700	2981.00000	0	3329.00000	국제유가	국제유가 떨어질 때도... 기름값은 40일째 '요지부동' / JTBC 뉴스룸	JTBC News	https://www.youtube.com/watch?v=GcHEOZiW-sA	2022-06-14

국제유가 조회수 순 유튜브 데이터

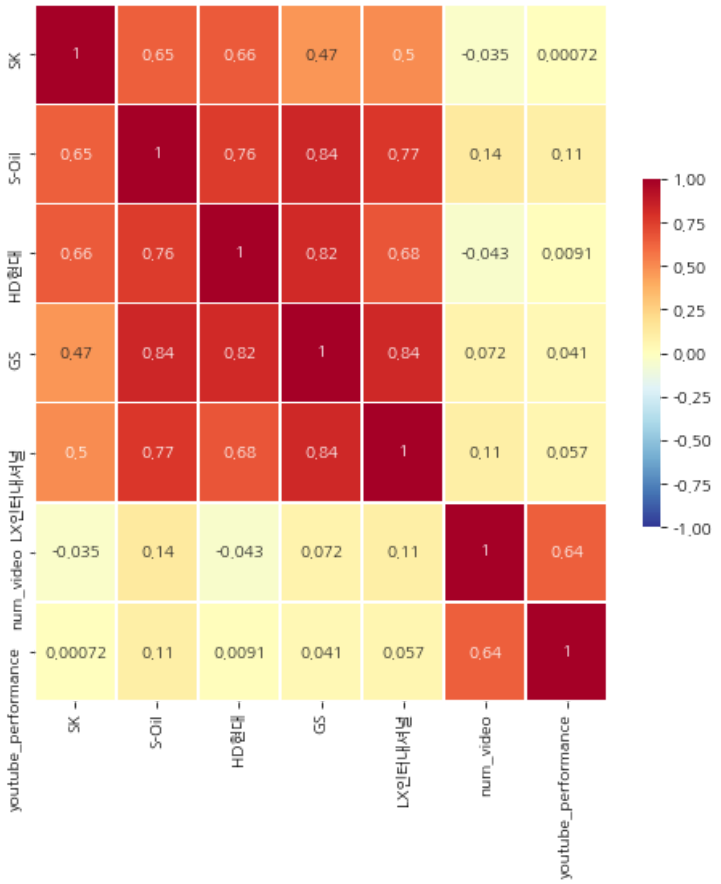
전처리 과정 - 위와 동일

상관성 분석

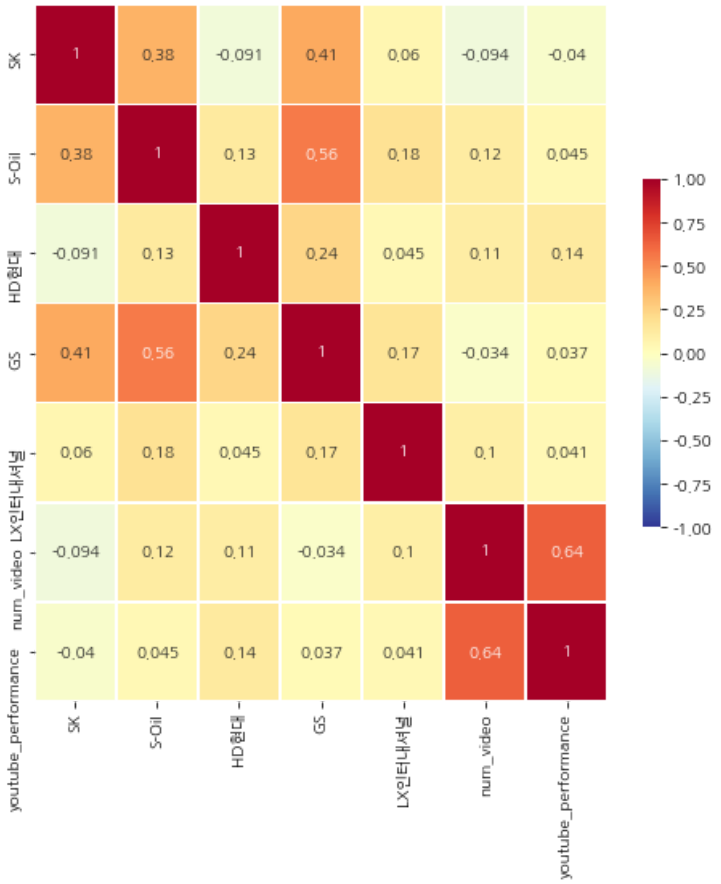
1. 유튜브 데이터 - 등락률


SK
S-Oil
HD현대
GS
LX인터내셔널
num_video
youtube_performance
1
0.45
0.41
0.44
0.4
-0.03
-0.057
0.45
1
0.64
0.7
0.48
-0.039
-0.11
0.41
0.64
1
0.67
0.52
-0.059
-0.088
0.44
0.7
0.67
1
0.45
-0.055
-0.1
0.4
0.48
0.52
0.45
1
-0.061
-0.089
-0.03
-0.039
-0.059
-0.055
-0.061
1
0.64
-0.057
-0.11
-0.088
-0.1
-0.089
0.64
1
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

2. 유튜브 데이터 - 종가


SK
S-Oil
HD현대
GS
LX인터내셔널
num_video
youtube_performance
1
0.65
0.66
0.47
0.5
-0.035
0.00072
0.65
1
0.76
0.84
0.77
0.14
0.11
0.66
0.76
1
0.82
0.68
-0.043
0.0091
0.47
0.84
0.82
1
0.84
0.072
0.041
0.5
0.77
0.68
0.84
1
0.11
0.057
-0.035
0.14
-0.043
0.072
0.11
1
0.64
0.00072
0.11
0.0091
0.041
0.057
0.64
1
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

3. 유튜브 데이터 - 거래량



결론

항공주

1. 등락률의 경우 네 종목 모두 큰 상관관계가 없음
2. 종가의 경우 대한항공(-0.13), 제주항공(-0.12), 진에어(-0.15)로 유튜브 반응과 약간의 음의 상관관계가 있지만 거의 없다고 할 수 있음
3. 거래량의 경우 진에어(0.12)가 유튜브 영상개수와의 상관계수가 가장 높았지만 그 수치가 낮아 상관성이 거의 없음

유가주

1. 등락률의 경우 네 종목 모두 큰 상관관계가 없음
2. 종가의 경우 S-Oil이 영상 개수와 0.14 , 영상 반응과 0.11로 가장 높은 상관관계를 가지지만 그 수치가 낮아 상관성이 거의 없음
3. 거래량 또한 S-Oil는 영상 개수와(0.12), HD현대는 영상개수(0.11), 유튜브 반응과(0.14) 약간의 양의 상관계수를 가지지만 그 수치가 낮음

분석 결과 처음의 가정(유튜브 데이터와 주식 데이터 간의 상관 관계가 있을 것이다)과 달리 큰 상관관계를 보이지 않음을 확인할 수 있었다.

한계점

- 유튜브 API를 활용해 프로젝트를 진행했는데 할당량의 한계와 검색 키워드 설정의 어려움으로 인해 데이터의 양이 적었고 수집한 데이터 또한 실제로 해당 키워드와 관련된 영상인지 하나하나 확인하기가 어려워 상관성이 낮게 나온 것으로 보임.
    - 데이터 수집량을 늘리고 주식 데이터와 관련성이 있는 기준(제목에 주식이 포함된 영상 또는 관련 종목의 이슈를 다루는 영상 등)을 세워서 유튜브 데이터의 품질 높이기
- 결측치의 경우 선형 보간법을 사용했는데 추후 유튜브 영상과 사람들의 관심도 간의 관계를 더 조사해 적절한 Metric에 따라 보간했으면 더 높은 상관성을 얻을 수 있었을 것 같다.