분석 보고서 - 경진

주제 - 유튜브 데이터를 활용한 주가 데이터 상관성 분석 및 시각화

서론

- 다양한 사건들과 상황에 따라 주가에 영향을 미치는 경우가 많이 생긴다. 사건(상황)과 주가 사이의 존재하는 매개변수를 사람들의 관심도라고 생각한다. 이 관심도와 주가가 얼마나 많은 상관관계를 알아보기 위해 유튜브 데이터를 활용하여 분석 및 시각화를 진행하고자 한다.
- 가정 : 사람들의 관심도를 알 수 있는 유튜브 데이터와 주가 데이터 사이에는 상관관계가 존재할 것이다.

분석과정

▼ 항공주

1) 대한항공 - 003490

2) 제주항공 - 089590

3) 진에어 - 272450

4) 에어부산 - 298690

▼ 석유관련주

1) SK이노베이션 - 096770

2) S-Oil - 010950

3) HD현대 - 267250

4) GS - 078930

5) LX인터내셔널 - 001120

1) 주가데이터 전처리

- 2020.01.02 ~ 2022.12.29 기간, '종가', '등락률', '거래량' 3개의 기준 사용
  - 종가, 거래량은 종목당 크기가 다르기 때문에 정규화진행 (MinMaxScaler)
- 주가데이터 시각화 및 3개의 기준 최대, 최소 시기 파악
  - 항공주 - 최대시기, 최소시기

	종가	등락률	거래량
대한항공	2021-06-11	2020-03-25	2020-11-16
제주항공	2021-06-11	2020-03-25	2022-11-22
진에어	2021-05-17	2020-03-25	2022-06-13
에어부산	2020-01-03	2020-11-16	2020-11-16

	종가	등락률	거래량
대한항공	2020-03-19	2020-03-19	2020-01-10
제주항공	2020-03-23	2020-03-19	2021-08-27
진에어	2020-03-19	2020-03-19	2020-09-23
에어부산	2022-10-31	2020-03-19	2021-05-27

- 석유관련주 - 최대시기, 최소시기

	종가	등락률	거래량
SK이노베이션	2021-02-02	2020-03-25	2020-08-05
S-Oil	2022-06-10	2020-04-02	2020-11-10
HD현대	2021-05-11	2020-02-06	2021-04-28
GS	2020-01-07	2020-03-25	2020-11-30
LX인터내셔널	2022-09-15	2020-04-06	2020-04-07

	종가	등락률	거래량
SK이노베이션	2020-03-19	2020-03-19	2022-12-21
S-Oil	2020-03-23	2020-03-19	2021-12-22
HD현대	2020-03-23	2020-03-19	2021-04-08
GS	2020-09-24	2020-03-23	2022-12-15
LX인터내셔널	2020-03-19	2020-03-19	2020-01-21

2) 유튜브 키워드 전처리 및 주가 데이터와 병합

▼ 사용 키워드

- 항공주 : '항공사+주가'
  - '항공사'로 키워드를 진행했을 때, 주가와 관련 없는 내용들이 너무 많아서 +주가를 추가하였음
- 석유관련주 : '유가전쟁'

i) 기간별 영상 수

- 1주 간격으로 업로드 된 영상 개수 얻음

```
keyword_num = keyword_df_use.copy()
keyword_num.drop_duplicates(inplace = True)
keyword_num['CNT'] = 1
keyword_num = keyword_num.resample('1W-MON')['CNT'].agg(np.sum).fillna(0) # 1주 간격 영상 수
```

	CNT
date	
2020-01-06	0
2020-01-13	0
2020-01-20	0

	CNT
date	
2022-12-12	5
2022-12-19	1
2022-12-26	7

- 주가 데이터와 join='outer'를 이용하여 병합한 뒤 'bfill'을 이용하여 중간 중간 결측치들을 그 주의 값으로 채워줌
  - 마지막 22.12.26이후에 데이터는 23.01.02 데이터 값으로 채워줌

```
close_merge = pd.concat([close_sc,keyword_num],axis=1, join='outer')
```

```
close_merge.CNT[-1] = 7  # 22.12.26 그 다음주 영상개수
```

```
close_merge.CNT.fillna(method='bfill', inplace=True)  # 중간 중간 결측치들은 그 주의 값으로 채워줌
```

```
close_merge
```

	대한항공	제주항공	진에어	에어부산	CNT
2020-01-02	0.359760	0.790576	0.432947	0.985532	0.0
2020-01-03	0.345859	0.764270	0.430777	1.000000	0.0
2020-01-06	0.325688	0.701284	0.411100	0.946215	0.0
2020-01-07	0.334527	0.719624	0.417659	0.929545	0.0
2020-01-08	0.320626	0.661927	0.428607	0.867477	0.0
...	...	...	...	...	...
2022-12-23	0.606713	0.324404	0.592432	0.040365	7.0
2022-12-26	0.602819	0.327838	0.589909	0.061334	7.0
2022-12-27	0.598925	0.375919	0.615136	0.072342	7.0
2022-12-28	0.602819	0.417130	0.630272	0.096980	7.0
2022-12-29	0.550251	0.393090	0.592432	0.081254	7.0

ii) 관심도 (조회수/댓글수) 변수생성

- 조회수, 좋아요수, 댓글수 모두 영상게시기간이 길수록 값이 커진다고 생각하여, 사람들의 관심도를 확인하기 위해 새로운 변수(조회수 / 댓글수)를 사용해봄

```
keyword_df_use2['interest'] = keyword_df_use2['viewCount'] / keyword_df_use2['commentCount']
keyword_df_use2 = keyword_df_use2[['interest']]
keyword_df_use2.head(3)
```

	interest
date	
2022-12-20	364.000000
2023-02-06	649.800000
2022-05-23	329.909091

- 두 col의 나눗셈을 계산하기 위해 조회수 및 댓글수의 '0'값도 결측치로 처리하여 결측치들을 모두 제거함. 또한 같은 시기의 여러 영상이 있으면, 모든 관심도 값을 더해줌.

```
keyword_df_use2 = keyword_df_use2.groupby(level=0).agg(np.sum)  # 같은 시기의 interest는 더해줌
keyword_df_use2.columns = ['interest']
```

- 주가데이터와 병합하기 위해 2020.01.02 ~ 2022.12.29의 관심도값의 새로운 dataframe을 생성하고, 관심도는 연속형 변수라고 생각하여 결측치들을 시기별 평균값으로 대체함(interpolate이용)

```
date_idx = pd.date_range('2019-10-11', '2022-12-30')
date_df = pd.DataFrame(range(len(date_idx)), index=date_idx)
date_merge = pd.concat([date_df, keyword_df_use2], axis=1, join='outer')
```

```
date_merge.iloc[:, 1].interpolate(inplace=True)  # 결측치 있는 시기 시기별평균값으로 대체
```

```
date_merge = pd.DataFrame(date_merge.iloc[:, 1], columns=['interest'])
```

```
date_merge = date_merge['2020-01-02' : '2022-12-29']
date_merge.head()  # 2020.01.02 ~ 2022.12.29 데이터 추출
```

	interest
2020-01-02	303.300757
2020-01-03	305.247283
2020-01-04	307.193808
2020-01-05	309.140334
2020-01-06	311.086859

- 주가데이터와 join='inner'를 이용하여 병합함

```
close_merge2 = pd.concat([close_sc, date_merge],axis=1, join='inner')  # 공통시기일때만 병합
```

```
close_merge2
```

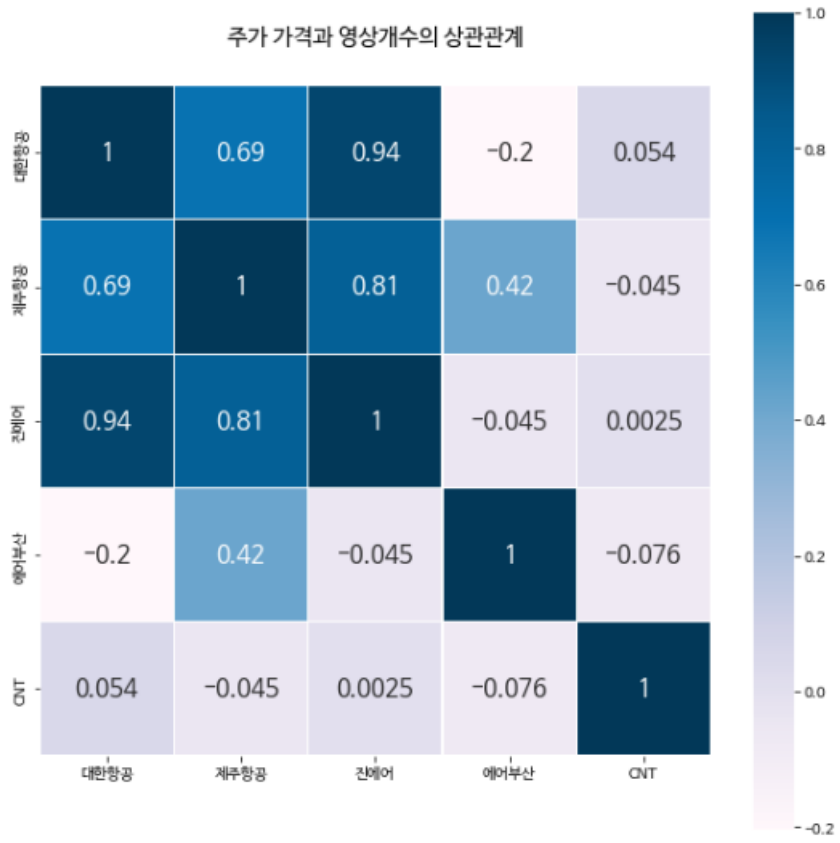
	대한항공	제주항공	진에어	에어부산	interest
2020-01-02	0.359760	0.790576	0.432947	0.985532	303.300757
2020-01-03	0.345859	0.764270	0.430777	1.000000	305.247283
2020-01-06	0.325688	0.701284	0.411100	0.946215	311.086859
2020-01-07	0.334527	0.719624	0.417659	0.929545	313.033385
2020-01-08	0.320626	0.661927	0.428607	0.867477	314.979911

3) 상관분석

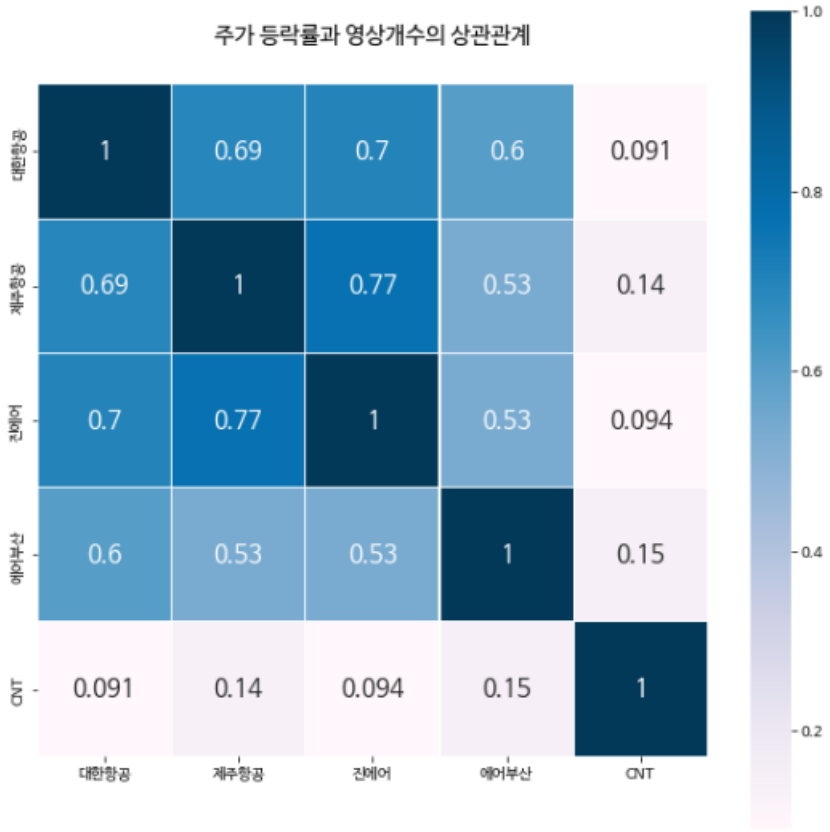
- 종가기준, 등락률 기준 -> 항공주, 석유관련주의 종목들이 모두 종가, 등락률일 때의 큰 상관관계를 갖고 있으므로 두 기준으로 진행했음.

i) 항공주

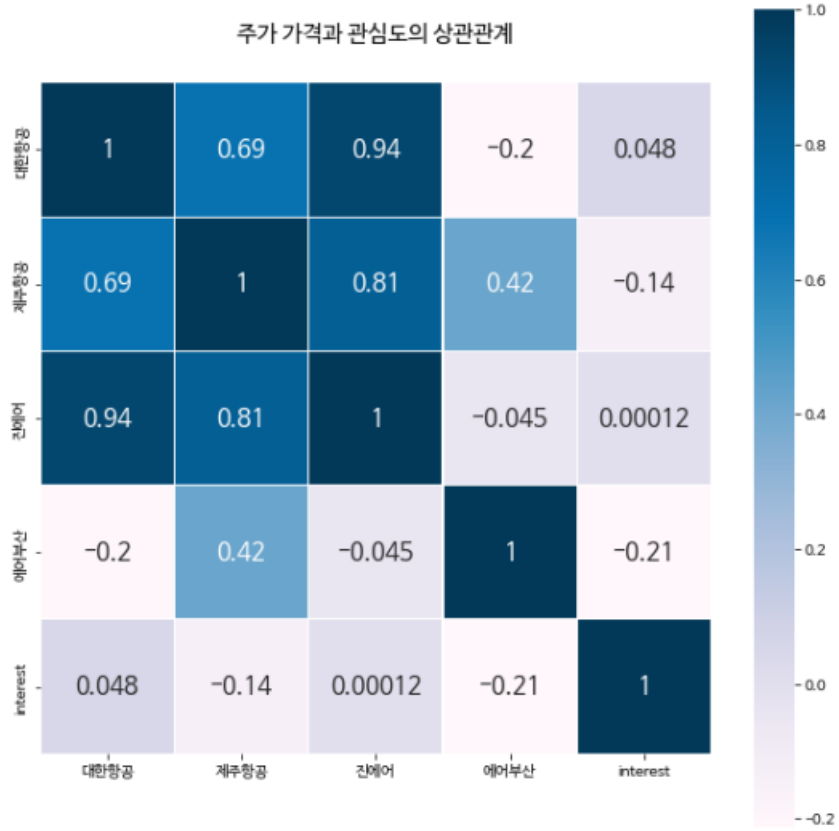
- 종가와 영상 수의 상관관계



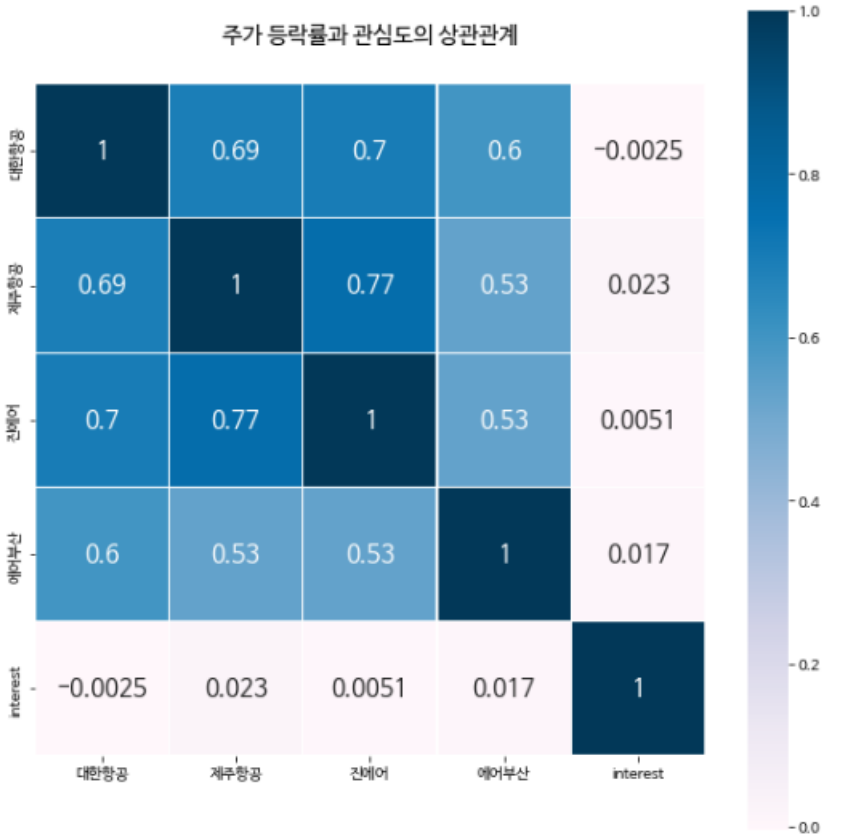
- 등락률과 영상 수의 상관관계



- 종가와 관심도의 상관관계

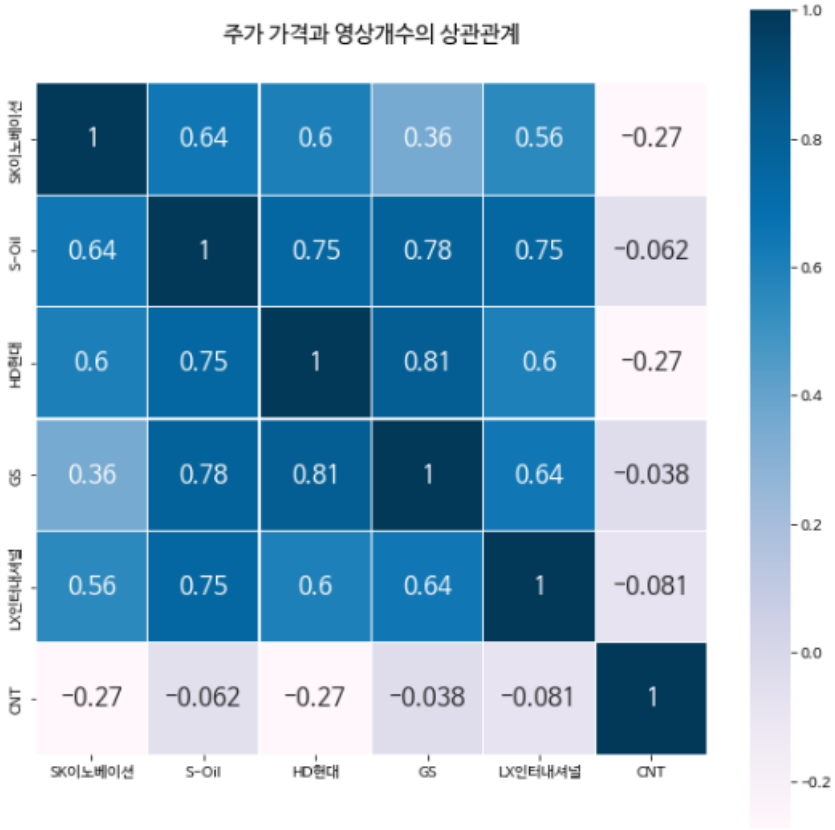


- 등락률과 관심도의 상관관계

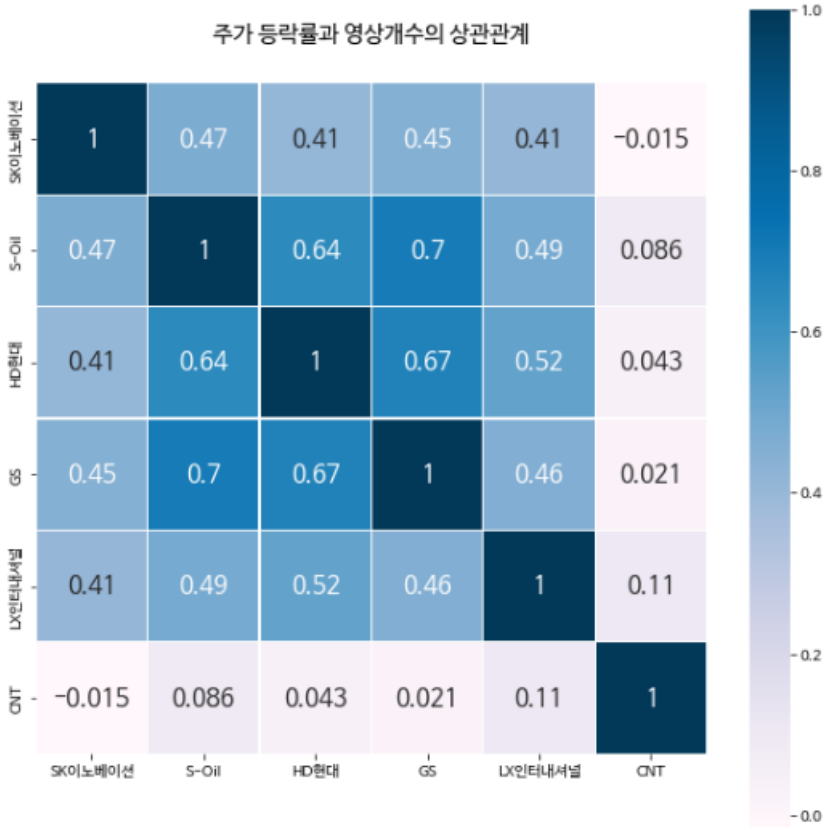


ii) 석유관련주

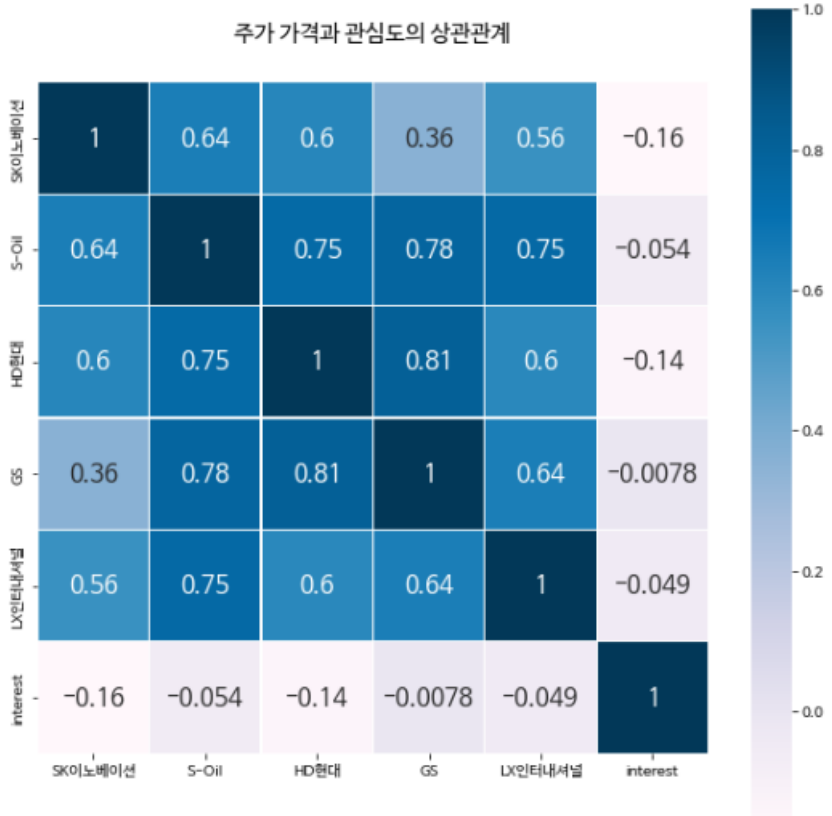
- 종가와 영상 수의 상관관계



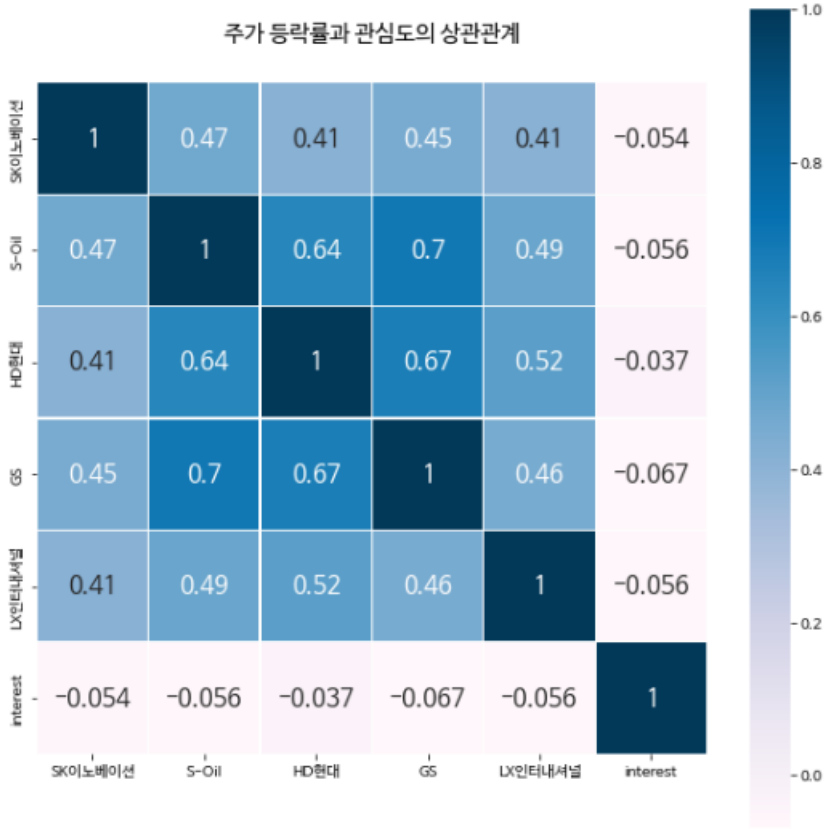
- 등락률과 영상 수의 상관관계



- 종가와 관심도의 상관관계



- 등락률과 관심도의 상관관계



분석결과

i) 항공주

- 종가와 영상개수의 큰 상관관계를 확인하지 못함
  - 0.05, -0.04, 0.002, -0.07
- 등락률과 영상개수도 마찬가지로 큰 상관관계를 확인하지 못하였지만, 종가에 비해 관계가 깊다고 확인됨
  - 0.09, 0.14, 0.09, 0.14
- 종가, 등락률과 관심도도 마찬가지로 큰 상관관계를 확인하지 못하였지만, 종가 기준 '에어부산'과 관심도의 관계가 -0.21로 가장 큰 상관관계를 보였다.

ii) 석유관련주

- 항공주와 마찬가지로 큰 상관관계를 확인하지 못하였지만, 종가와 영상 수의 상관관계에서 'SK이노베이션'과 'HD현대' 두 종목이 -0.27이라는 나름 큰 값을 얻은 것을 확인할 수 있다.

보완하면 좋은 부분

- 주요사건을 중점으로 분석을 실시하였기에, 그 사건을 중심으로 기간을 좁혔으면 더 좋은 상관관계가 나왔을 것이라고 생각함
- 상관관계와 인과관계는 엄연히 다른 개념. 따라서 상관분석에서 더 나아가 회귀분석을 통해 계수의 유의성을 판단하여 인과관계를 파악하면 좋을 것

- 사람들의 관심이 주가에 반영되는 시기를 고려하여, 전처리하고 분석했으면 더 좋은 결과가 나올 것이라고 생각함
- 관심도를 조회수/댓글수 로 측정하여 진행했는데, 다양한 변수들을 생성하여 분석하면 더 좋을 것이라고 생각함

한계

- 유튜브 API를 활용하는 점에서 크고 작은 사건들을 동시에 다룰 수 없었기에, 주가와 관련된 주요사건을 중점으로 분석할 수 밖에 없었음