뉴스 데이터의 긍부정 비율과 주가의 상관관계 분석

소프트웨어융합학과 이상민
경제학과 장현지
소프트웨어융합학과 김효준

목차

01

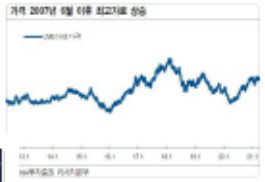
연구 배경

01. 연구 배경

뉴시스 | 4일 전 | 네이버뉴스

NH證 "고려아연, 아연 가격 강세로 **주가** **상승** 기대"

NH투자증권은 4일 고려아연에 대해 아연 가격이 강세를 보이고 있어 **주가** **상승**이



핀포인트뉴스 | 1일 전

아톤, 지난해 호실적에도 **주가**는 **하락**...이유있나?

아톤이 지난해 호실적을 7일 공시했지만, **하락**세에 접어든 **주가**에 영향을 주지 못했다. 이날 한국거

한국경제 | 6일 전 | 네이버뉴스

현대차, 러시아 공장 가동 일시 중단 소식에 **주가** **하락**

현대차가 러시아 공장 가동을 일시 중단했다는 소식에 **주가**가 **하락**하고 있다. 2일

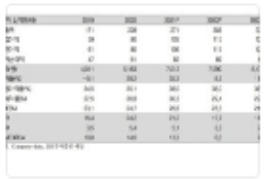
국제뉴스 | 1일 전

[급등주]효성오앤비 **주가** 24% **상승**세, 국제 곡물가격 **상승** 여파

효성오앤비 **주가** 24% **상승**곡선을 그리고 있다.

뉴시스 PiCK | 5일 전 | 네이버뉴스

IBK證 "티씨케이, 특허 무효에 과도한 **주가** **하락**...투자매력↑"



에너지경제 | 1일 전

[종합**주가**지수] 코스피 '2.2%' 넘게 급락...SK하이닉스·LG화학 등 ...

61%) 등 20위권 전 종목이 **하락**했다. 특히 삼성전자는 장중 한때 6만9900원까지 떨어져 작년 11월

금강일보 | 9시간 전

[주식] 셀트리온 **주가**, 초반 **하락** 딛고 상승세 보여

셀트리온 **주가**, 초반 **하락** 딛고 상승세 보여 사진=셀트리온 셀트리온 **주가**가 상승세를 보이며 투자자들의 눈길을 모으고 있다. 8일 오전 10시 18분께 셀트리온은 ...

주가

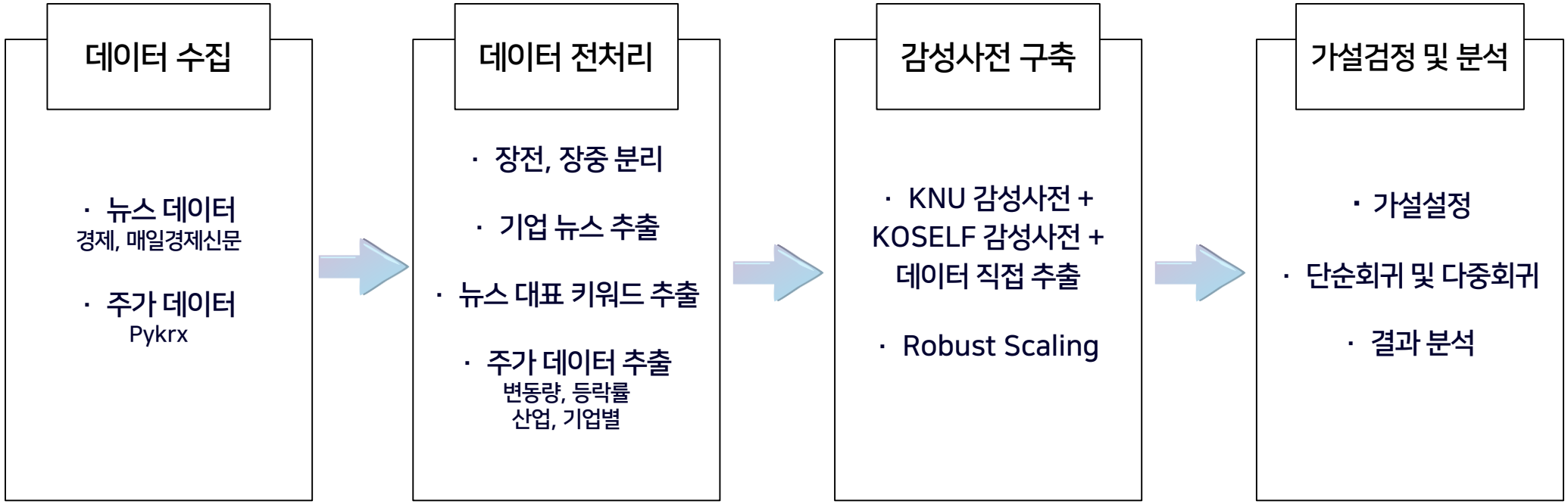
뉴스의 긍부정 비율에 따라 주가는 어떻게 변할까?

02

데이터 분석

02. 분석 과정

도식화


데이터 수집
· 뉴스 데이터
경제, 매일경제신문
· 주가 데이터
Pykrx
데이터 전처리
· 장전, 장중 분리
· 기업 뉴스 추출
· 뉴스 대표 키워드 추출
· 주가 데이터 추출
변동량, 등락률
산업, 기업별
감성사전 구축
· KNU 감성사전 +
KOSELF 감성사전 +
데이터 직접 추출
· Robust Scaling
가설검정 및 분석
· 가설설정
· 단순회귀 및 다중회귀
· 결과 분석

02. 분석 과정

데이터 수집

뉴스 데이터

대용량 뉴스 크롤러 오픈소스 활용

- 날짜, 카테고리, 신문사, 제목, 본문, 기사

2021.03.01	경제	조선일보	美 뉴욕 증	다우존스	https://news.naver.com/main/r
2021.03.01	경제	머니투데이	속보뉴욕증	머니투데이	https://news.naver.com/main/r
2021.03.01	경제	머니투데이	노래 들으(	머니투데이	https://news.naver.com/main/r
2021.03.01	경제	매일경제	인플레이션	한국가스공	https://news.naver.com/main/r
2021.03.01	경제	매일경제	3월 증시 오	금리변수 (	https://news.naver.com/main/r
2021.03.01	경제	한국경제	씨티 비트코	미국 투자	https://news.naver.com/main/r
2021.03.01	경제	YTN	2월 중 무역	앵커 지난	https://news.naver.com/main/r
2021.03.01	경제	한국일보	페라리 SU	마세라티 (	https://news.naver.com/main/r
2021.03.01	경제	한국경제	아영FBC C	주류업체 (	https://news.naver.com/main/r

주가 데이터

pykrx 오픈소스 활용

- 기업별 주가, 코스피 주가, 산업군별 주가

```
1  from pykrx import stock
2  from pykrx import bond
3
4  df = stock.get_index_ohlcv("20181228", "20220228", "1001")
5  df.head()
```

코스피 날짜	시가	고가	저가	종가	거래량	거래대금
2018-12-28	2036.70	2046.97	2035.41	2041.04	352677713	4120695824217
2019-01-02	2050.55	2053.45	2004.27	2010.00	326367773	4295871822881
2019-01-03	2011.81	2014.72	1991.65	1993.70	427976017	5358519356361
2019-01-04	1992.40	2011.56	1984.53	2010.25	408990897	5490147620731
2019-01-07	2034.24	2048.06	2030.90	2037.10	440191435	5301385184683

https://github.com/lumyjuwon/KoreaNewsCrawler - 대용량 뉴스 크롤러 오픈소스

https://github.com/sharebook-kr/pykrx - krx주식정보 스크래핑

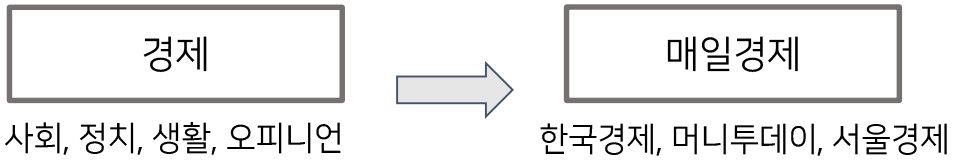
02. 분석 과정

데이터 전처리

뉴스 데이터 전처리

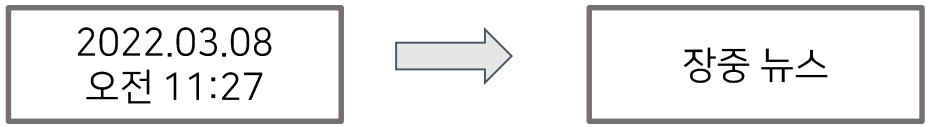
1) 경제 카테고리와 매일경제 신문사 데이터 활용

- 주가에 큰 영향을 주는 토픽은 경제라고 생각함
- 매일경제 신문사가 가장 뉴스 개수가 많고, 질이 좋은 신문사라고 생각함



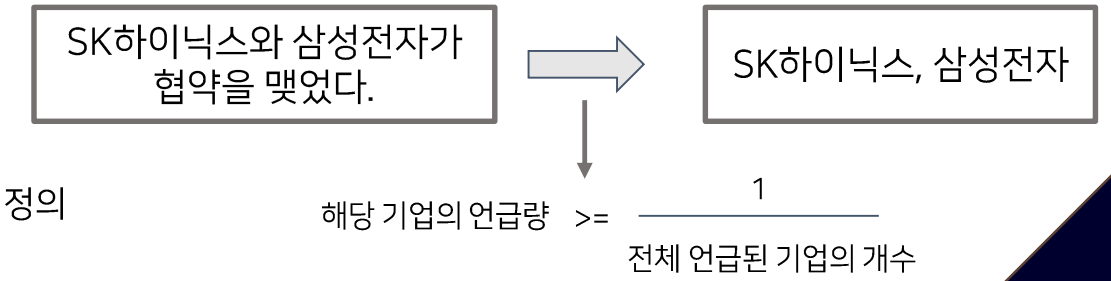
2) 날짜를 장전과 장중으로 분리

- 장중 : 해당 날짜의 시간이 09-15:30 사이이면 장중
- 장전 : 해당 날짜의 시간이 18시 이후면 다음날의 장전,
  해당 날짜의 시간이 09시 이전이면 해당 날짜의 장전



3) 뉴스에서 대표 회사 추출

- KRX 종목 리스트 데이터 이용
- 해당 기업 이름이 뉴스 원문에 포함되어 있으면 기업명 리스트에 삽입
- (해당 기업의 개수) >= 1/(전체 언급된 기업명 개수) 이면, 해당 회사의 뉴스로 정의



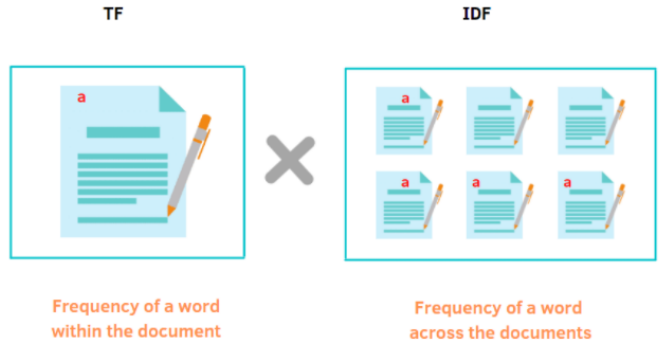
02. 분석 과정

데이터 전처리

뉴스의 대표 키워드 추출

1) TF-IDF Clustering을 활용한 키워드 추출

- mecab을 활용해 명사 형태소만 추출
- TF-IDF 값 산출 후 군집화
- 군집 내 대표 키워드 산출
- 군집을 labeling하여 대표 키워드를 해당 뉴스의 키워드로 정의



문서1 -> Cluster1
문서2 -> Cluster1
문서3 -> Cluster2
…

Cluster1 대표 키워드
- 금리, 시장, 환율 …

Cluster2 대표 키워드
- 코로나, 마스크, 바이러스 …

실험결과

- 뉴스마다 특징이 달라 Cluster 개수를 결정 불가능(Perplexity, Coherence 등)
- 뉴스마다 핵심 키워드가 다르기 때문에 TF-IDF Clustering방법은 채택하지 않음.

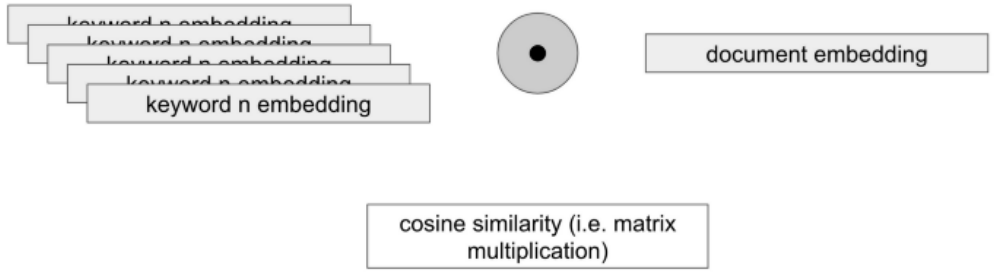
02. 분석 과정

데이터 전처리

뉴스의 대표 키워드 추출

2) KeyBERT를 활용한 키워드 추출

- 형태소 분석기를 통해 명사만 추출
- 벡터화 후, n-gram을 1~3으로 설정하여 단어 추출
- 다국어 SBERT를 활용해 문서와 가장 유사한 키워드 추출



아래 기법 추가 실험

- Max Sum Similarity : 데이터 쌍 사이의 최대 합 거리는 데이터 쌍 간의 거리가 최대화되는 데이터 쌍으로 정의
- Maximal Marginal Relevance : 중복을 최소화하고 결과의 다양성을 극대화 하는 방법

  문서와 가장 유사한 키워드 선택하고, 선택한 키워드와 비슷하지 않은 새 후보를 반복적으로 선택

국토교통부가 버스 안전사고 예방을 위해 관계기관과 차량 안전장치 정상 작동 여부 등에 대한 집중 점검에 나선다. 국토부는 1일 17개 시도 버스 업계 한국교통안전공단 등과 영상회의를 열고 대책을 논의했다고 밝혔다. 이번 회의는…

['교통 정책 시내버스', '버스 업계 지자체', '시내버스 승객', '버스 교통사고 예방', '버스 승객 사망']

02. 분석 과정

데이터 전처리

주가 데이터 전처리

1) 주가 변동량

- 해당 날짜의 시가 - 전 날의 종가(장전 데이터와 비교)
- 해당 날짜의 종가 - 해당 날짜의 시가(장중 데이터와 비교)

2) 카테고리별 주가 데이터 추출

- 기업별 주가 데이터
- 인덱스(산업군, 대형주 등)별 주가 데이터

코스피 날짜	시가	고가	저가	종가	거래량	거래대금	변동량
2019-01-02	2050.55	2053.45	2004.27	2010.00	326367773	4295871822881	9.51
2019-01-03	2011.81	2014.72	1991.65	1993.70	427976017	5358519356361	1.81
2019-01-04	1992.40	2011.56	1984.53	2010.25	408990897	5490147620731	-1.30
2019-01-07	2034.24	2048.06	2030.90	2037.10	440191435	5301385184683	23.99
2019-01-08	2038.68	2042.70	2023.59	2025.27	397831202	4826641977635	1.58

02. 분석 과정

데이터 전처리

긍/부정 사전

1) KNU 감성사전

- 긍정: '가격이 싸다', '가까이 사귀어', '가까이하다', '가능성이 늘어나다'
- 부정: '거슬리다', '거역하다', '거역함', '거의 없다', '거절하거나'

2) 경제 도메인 텍스트

- 긍정: '가치', '가치 있는', '강세', '개선', '개선된', '개선되는', '경신'
- 부정: '결국', '결함', '공허한', '과적', '극심한', '둔화', '마이너스'

3) 뉴스 기사에서 직접 추출한 텍스트

- 긍정: '신사업 진출', '고성장', '공급 증가', '매출 증대', 'D램 수요 증가'
- 부정: '공급 부족', '불황', '수출 규제', '우려', '소비 위축', '', '침체'

→ KNU 감성사전 단어는 0.5점, 경제 도메인 및 뉴스 기사에서 직접 추출한 텍스트 1점 으로 반영

02. 분석 과정

데이터 전처리

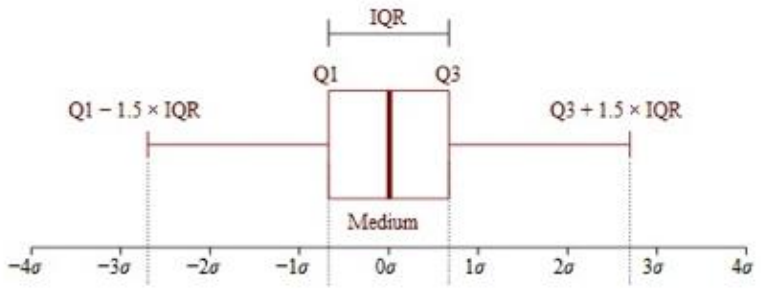
긍/부정 비율 및 스케일링

1) 긍부정 비율

긍정점수 / (긍정점수 + 부정점수) = 긍부정 비율
긍정점수 = 10, 부정점수 = 5, 긍부정 비율= 10/15 = 0.666…

2) Robust Scaler

긍부정 비율이 0.5 이상(긍정)에 몰려있는 경향을 띠어서 Robust Scaler를 통한 scaling을 진행



Source: statisticshowto

$$\frac{x_i - Q_2(x)}{Q_3(x) - Q_1(x)}$$

02. 분석 과정

데이터 전처리

최종 데이터

	날짜	신문사	제목	원문	기준일자	오전오후	시간	장전장중	회사	키워드	긍정점수	부정점수	긍부정비율	RobustSclaer
0	2019.01.01. 오후 9:34	매일경제	이마트 노브랜드 초콜릿 백화현상에 전량 회수	이마트가 자체 브랜드로 만들어 판매하는 노브랜드 다크 초콜릿 일부 제품이 변질돼 판매...	20190102	오후	21	장전	이마트	초콜릿 일부 제품 항의 이마트 상품 구매 고객 제품 품질 상품 소비자 혼란	2.0	2.5	0.444444	-0.748538
1	2019.01.01. 오후 7:18	매일경제	반도체경기 둔화에 가격 인하 요청 쇄도...장비 발주도 재검토	지난달 반도체 수출 8.3%↓ 2016년 9월 이후 첫 역성장 삼성 원가 절감·재고축소...	20190102	오후	19	장전	LG SK SK하이닉스 레이 삼성전자	반도체 수출 마이너스 대한 대응 마련 중국 추격 무관 경제 먹구름 작년 시장 한국	8.0	13.5	0.372093	-1.037944
2	2019.01.01. 오후 7:17	매일경제	정유·유화도 초호황 끝...수요 감소로 혹한기 대비	작년 4분기 실적 적자 가능성 철강은 중국發 공급과잉 우려 비상 걸린 수출한국 중국...	20190102	오후	19	장전	레이	시장 침체 수요 석유화학 업계 새해 중국 글로벌 철강업 충격 부족	4.0	11.5	0.258065	-1.494058

03

가설검정

03. 가설 검정

분석 방법 : 회귀분석

독립변수와 종속변수 사이의 상관관계를 나타내는 통계분석기법

독립변수가 종속변수에 미치는 영향을 확인, 추정된 회귀모형을 통해 종속변수의 예측치를 구할 수 있음

1) 단순회귀분석

- 독립 변수가 단일 개일 때 사용
- $y = \alpha + \beta x + \varepsilon_i$

2) 다중회귀분석

- 독립 변수가 두 개 이상일 때 사용
- $y = \alpha + \beta x + \varepsilon_i \; y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

03. 가설 검정

분석 방법 : 회귀분석

변수 설정

- (장전) X 변수 : 뉴스의 긍부정비율, Y 변수 : 등락률, 종가
- (장중) X 변수 : 뉴스의 긍부정비율, Y 변수 : 종가-시가, 종가

03. 가설 검정

가설 설정

거시적 분석

1) 국내 금리 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다.

2) 국제 분쟁 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다.

3) 환율 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다.

4) 국제유가 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다.

03. 가설 검정

결과 분석

① 장전(X변수 : 분쟁, 금리, 환율, 유가, Y변수 : 코스피등락률)

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.212
Model:	OLS	Adj. R-squared:	0.147
Method:	Least Squares	F-statistic:	3.226
Date:	Tue, 08 Mar 202	Prob (F-statistic):	0.0198
Time:	15:52:38	Log-Likelihood:	-98.252
No. Observations:	53	AIC:	206.5
Df Residuals:	48	BIC:	216.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1539	0.389	0.396	0.694	-0.628	0.936
분쟁	0.4244	0.446	0.951	0.346	-0.473	1.322
금리	0.6430	0.337	1.907	0.063	-0.035	1.321
환율	-0.6700	0.245	-2.72	0.009	-1.164	-0.176
유가	-0.0864	0.368	-0.23	0.816	-0.827	0.654

Omnibus:	18.628	Durbin-Watson:	1.814
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43.819
Skew:	0.901	Prob(JB):	3.05e-10
Kurtosis:	7.073	Cond. No.	3.86

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.212
Model:	OLS	Adj. R-squared:	0.163
Method:	Least Squares	F-statistic:	4.381
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.0082
Time:	15:53:59	Log-Likelihood:	-98.28
No. Observations:	53	AIC:	204.6
Df Residuals:	49	BIC:	212.4
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1943	0.346	0.562	0.577	-0.500	0.889
분쟁	0.4378	0.438	0.999	0.323	-0.443	1.319
금리	0.6446	0.334	1.931	0.059	-0.026	1.316
환율	-0.6848	0.235	-2.915	0.005	-1.157	-0.213

Omnibus:	18.734	Durbin-Watson:	1.825
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43.407
Skew:	0.916	Prob(JB):	3.75e-10
Kurtosis:	7.037	Cond. No.	3.39

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.195
Model:	OLS	Adj. R-squared:	0.163
Method:	Least Squares	F-statistic:	6.073
Date:	Tue, 08 Mar 20	2 Prob (F-statistic):	0.0043
Time:	15:54:38	Log-Likelihood:	-98.817
No. Observations:	53	AIC:	203.6
Df Residuals:	50	BIC:	209.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0106	0.278	-0.038	0.970	-0.569	0.548
금리	0.7169	0.326	2.200	0.032	0.062	1.372
환율	-0.6610	0.234	-2.828	0.007	-1.130	-0.192

Omnibus:	23.673	Durbin-Watson:	1.890
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65.753
Skew:	1.126	Prob(JB):	5.27e-15
Kurtosis:	7.970	Cond. No.	2.20

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.118
Model:	OLS	Adj. R-squared:	0.100
Method:	Least Squares	F-statistic:	6.797
Date:	Tue, 08 Mar 20	2 Prob (F-statistic):	0.0119
Time:	15:55:30	Log-Likelihood:	-101.26
No. Observations:	53	AIC:	206.5
Df Residuals:	51	BIC:	210.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1941	0.275	-0.705	0.484	-0.747	0.358
환율	-0.6308	0.242	-2.607	0.012	-1.117	-0.145

Omnibus:	34.139	Durbin-Watson:	1.847
Prob(Omnibus):	0.000	Jarque-Bera (JB):	140.493
Skew:	1.550	Prob(JB):	3.11e-31
Kurtosis:	10.349	Cond. No.	1.90

03. 가설 검정

결과 분석

② 장중(X변수 : 분쟁, 금리, 환율, 유가, Y변수 : 코스피 종가)

OLS Regression Results

Dep. Variable:	종가_x	R-squared:	0.105
Model:	OLS	Adj. R-squared:	0.030
Method:	Least Squares	F-statistic:	1.401
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.248
Time:	16:01:33	Log-Likelihood:	-392.60
No. Observations:	53	AIC:	795.2
Df Residuals:	48	BIC:	805.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2497.0467	100.458	24.857	0.000	2295.062	2699.032
분쟁	48.9542	115.208	0.425	0.673	-182.688	280.596
금리	-17.8796	87.064	-0.205	0.838	-192.933	157.173
환율	96.3363	63.380	1.520	0.135	-31.099	223.771
유가	124.1962	95.103	1.306	0.198	-67.021	315.413

Omnibus:	6.826	Durbin-Watson:	0.227
Prob(Omnibus):	0.033	Jarque-Bera (JB):	5.230
Skew:	0.647	Prob(JB):	0.0732
Kurtosis:	2.168	Cond. No.	3.86

OLS Regression Results

Dep. Variable:	종가_x	R-squared:	0.104
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	1.891
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.143
Time:	16:03:03	Log-Likelihood:	-392.62
No. Observations:	53	AIC:	793.2
Df Residuals:	49	BIC:	801.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2499.2138	98.921	25.265	0.000	2300.424	2698.003
분쟁	43.9336	111.479	0.394	0.695	-180.092	267.959
환율	95.8261	62.710	1.528	0.133	-30.194	221.846
유가	124.5955	94.149	1.323	0.192	-64.604	313.795

Omnibus:	6.750	Durbin-Watson:	0.222
Prob(Omnibus):	0.034	Jarque-Bera (JB):	5.293
Skew:	0.658	Prob(JB):	0.0709
Kurtosis:	2.183	Cond. No.	3.76

OLS Regression Results

Dep. Variable:	종가_x	R-squared:	0.101
Model:	OLS	Adj. R-squared:	0.065
Method:	Least Squares	F-statistic:	2.806
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.0700
Time:	16:04:22	Log-Likelihood:	-392.70
No. Observations:	53	AIC:	791.4
Df Residuals:	50	BIC:	797.3
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2474.8736	76.616	32.302	0.000	2320.986	2628.762
환율	99.3314	61.549	1.614	0.113	-24.294	222.957
유가	119.5582	92.486	1.293	0.202	-66.206	305.322

Omnibus:	6.943	Durbin-Watson:	0.212
Prob(Omnibus):	0.031	Jarque-Bera (JB):	5.440
Skew:	0.668	Prob(JB):	0.0659
Kurtosis:	2.177	Cond. No.	2.65

OLS Regression Results

Dep. Variable:	종가_x	R-squared:	0.071
Model:	OLS	Adj. R-squared:	0.053
Method:	Least Squares	F-statistic:	3.890
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.0540
Time:	16:04:38	Log-Likelihood:	-393.57
No. Observations:	53	AIC:	791.1
Df Residuals:	51	BIC:	795.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2429.0905	68.385	35.521	0.000	2291.803	2566.378
환율	118.5636	60.116	1.972	0.054	-2.124	239.251

Omnibus:	6.267	Durbin-Watson:	0.208
Prob(Omnibus):	0.044	Jarque-Bera (JB):	5.721
Skew:	0.730	Prob(JB):	0.0572
Kurtosis:	2.322	Cond. No.	1.90

03. 가설 검정

결과 분석

거시적 분석

	장전	장중
1) 국내 금리 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다. ⇨	채택	X
2) 국제 분쟁 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다. ⇨	X	채택
3) 환율 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다. ⇨	X	X
4) 국제유가 뉴스의 긍부정 비율이 높을수록 코스피 지수는 상승할 것이다. ⇨	X	X

03. 가설 검정

가설 설정

산업 분석

수주 뉴스의 긍부정 비율이 높을수록 건설업 업종 기업들의 주가가 상승할 것이다.

03. 가설 검정

결과 분석

① 장전

x: 긍부정비율, y : 등락률

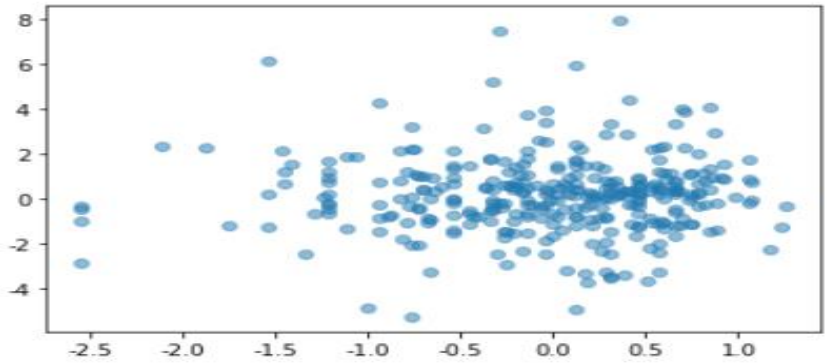
(0.013477771330847 41, 0.8072943858176649)

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.05959
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.807
Time:	15:17:58	Log-Likelihood:	-647.51
No. Observations:	330	AIC:	1299.
Df Residuals:	328	BIC:	1307.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1534	0.095	1.612	0.108	-0.034	0.341
긍부정비율	0.0336	0.137	0.244	0.807	-0.237	0.304

Omnibus:	42.941	Durbin-Watson:	1.715
Prob(Omnibus):	0.000	Jarque-Bera (JB):	127.533
Skew:	0.565	Prob(JB):	2.03e-28
Kurtosis:	5.828	Cond. No.	1.45



x: 긍부정비율, y : 종가

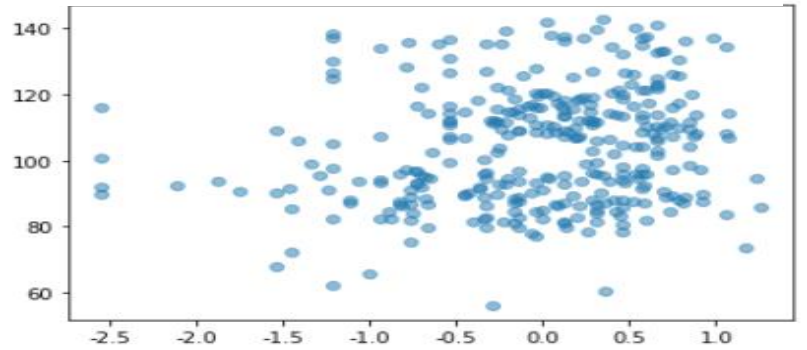
(0.18438048992563977, 0.0007636924557025387)

OLS Regression Results

Dep. Variable:	종가_x	R-squared:	0.034
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	11.54
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.000764
Time:	15:18:54	Log-Likelihood:	-1405.6
No. Observations:	330	AIC:	2815.
Df Residuals:	328	BIC:	2823.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	104.3463	0.947	110.202	0.000	102.484	106.209
긍부정비율	4.6463	1.368	3.398	0.001	1.956	7.337

Omnibus:	8.332	Durbin-Watson:	0.086
Prob(Omnibus):	0.016	Jarque-Bera (JB):	5.731
Skew:	0.186	Prob(JB):	0.0570
Kurtosis:	2.472	Cond. No.	1.45



03. 가설 검정

결과 분석

② 장중

x: 긍부정비율, y : 종가-시가

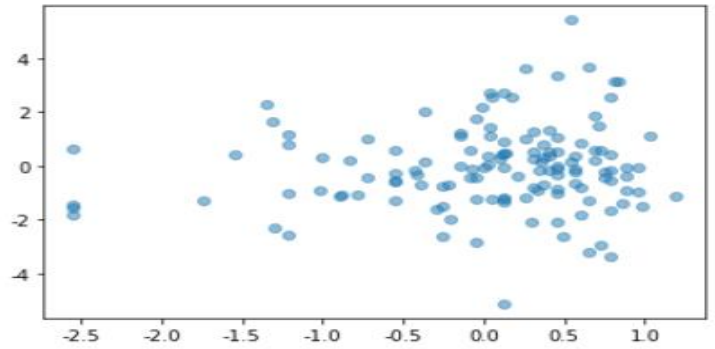
(0.114560834323660O7, 0.187484440299947 62)

OLS Regression Results

Dep. Variable:	해당날짜변동량	R-squared:	0.013
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	1.755
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.187
Time:	15:22:39	Log-Likelihood:	-249.16
No. Observations:	134	AIC:	502.3
Df Residuals:	132	BIC:	508.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0603	0.135	-0.446	0.656	-0.328	0.207
긍부정비율	0.2359	0.178	1.325	0.187	-0.116	0.588

Omnibus:	6.541	Durbin-Watson:	2.000
Prob(Omnibus):	0.038	Jarque-Bera (JB):	8.625
Skew:	0.250	Prob(JB):	0.0134
Kurtosis:	4.138	Cond. No.	1.32



x: 긍부정비율, y : 종가

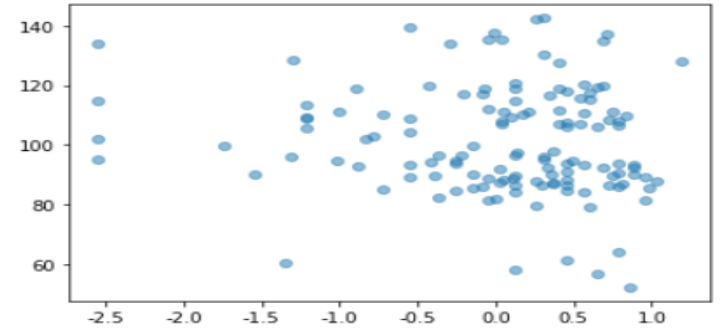
(-0.120045507753714, 0.16709146922487986)

OLS Regression Results

Dep. Variable:	종가_x	R-squared:	0.014
Model:	OLS	Adj. R-squared:	0.007
Method:	Least Squares	F-statistic:	1.930
Date:	Tue, 08 Mar 202	Prob (F-statistic):	0.167
Time:	15:24:46	Log-Likelihood:	-577.09
No. Observations:	134	AIC:	1158.
Df Residuals:	132	BIC:	1164.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	101.1516	1.564	64.681	0.000	98.058	104.245
긍부정비율	-2.8590	2.058	-1.389	0.167	-6.930	1.212

Omnibus:	0.674	Durbin-Watson:	0.098
Prob(Omnibus):	0.714	Jarque-Bera (JB):	0.421
Skew:	0.128	Prob(JB):	0.810
Kurtosis:	3.098	Cond. No.	1.32



03. 가설 검정

결과 분석

산업 분석

	장전	장중
수주 뉴스의 긍부정 비율이 높을수록, 건설업 업종 기업들의 주가가 상승할 것이다.	X	X

03. 가설 검정

가설 설정

기업 분석

삼성 전자 뉴스의 긍부정 비율이 높을수록 삼성전자의 주가는 상승할 것이다.

03. 가설 검정

결과 분석

① 장전

x: 긍부정비율, y : 등락률

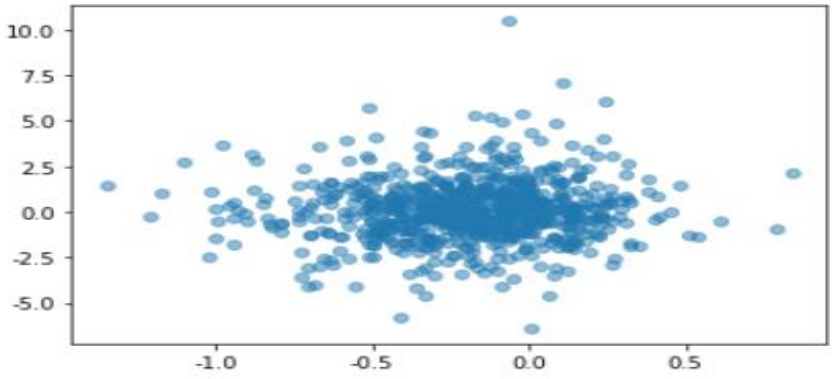
(0.04858413993960554, 0.17525651987867513)

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	1.841
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.175
Time:	14:38:04	Log-Likelihood:	-1497.7
No. Observations:	780	AIC:	2999.
Df Residuals:	778	BIC:	3009.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1469	0.071	2.068	0.039	0.007	0.286
긍부정비율	0.2749	0.203	1.357	0.175	-0.123	0.673

Omnibus:	92.373	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	336.942
Skew:	0.514	Prob(JB):	6.82e-74
Kurtosis:	6.051	Cond. No.	3.56



x: 긍부정비율, y : 종가

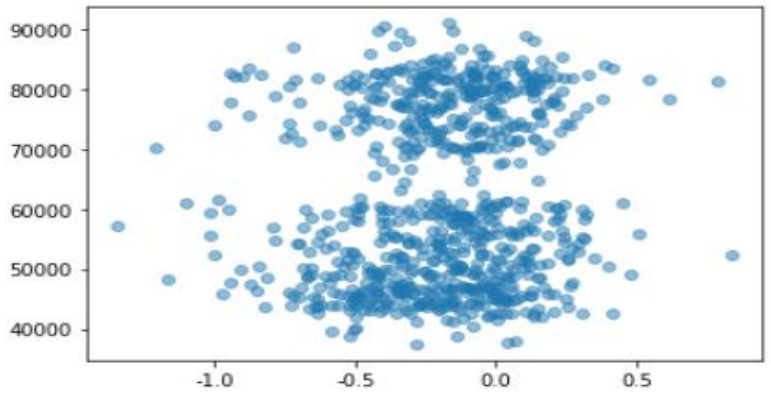
(0.10292719623094279, 0.004006672300634685)

OLS Regression Results

Dep. Variable:	종가	R-squared:	0.011
Model:	OLS	Adj. R-squared:	0.009
Method:	Least Squares	F-statistic:	8.330
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.00401
Time:	14:43:20	Log-Likelihood:	-8580.2
No. Observations:	780	AIC:	1.716e+04
Df Residuals:	778	BIC:	1.717e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.269e+04	623.441	100.552	0.000	6.15e+04	6.39e+04
긍부정비율	5135.3427	1779.249	2.886	0.004	1642.645	8628.041

Omnibus:	10589.971	Durbin-Watson:	0.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	72.794
Skew:	0.247	Prob(JB):	1.56e-16
Kurtosis:	1.587	Cond. No.	3.56



03. 가설 검정

결과 분석

② 장중

x: 긍부정비율, y : 종가-시가

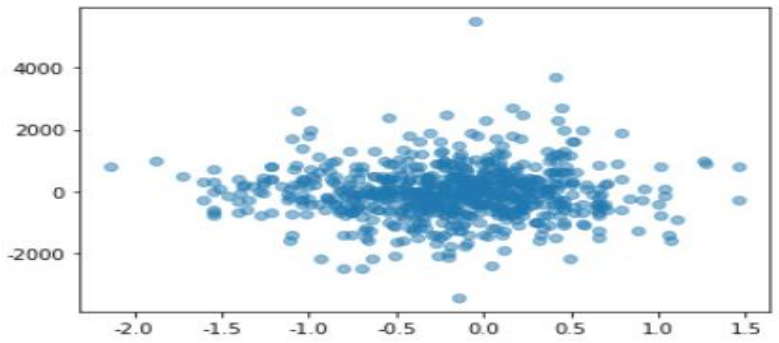
(0.01798243059053934, 0.6353023361575707)

OLS Regression Results

Dep. Variable:	해당날짜변동량	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	0.2251
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.635
Time:	14:50:08	Log-Likelihood:	-5693.3
No. Observations:	698	AIC:	1.139e+04
Df Residuals:	696	BIC:	1.140e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-26.7927	33.971	-0.789	0.431	-93.491	39.906
긍부정비율	28.1008	59.224	0.474	0.635	-88.178	144.379

Omnibus:	98.185	Durbin-Watson:	2.035
Prob(Omnibus):	0.000	Jarque-Bera (JB):	439.688
Skew:	0.553	Prob(JB):	3.33e-96
Kurtosis:	6.728	Cond. No.	1.95



x: 긍부정비율, y : 종가

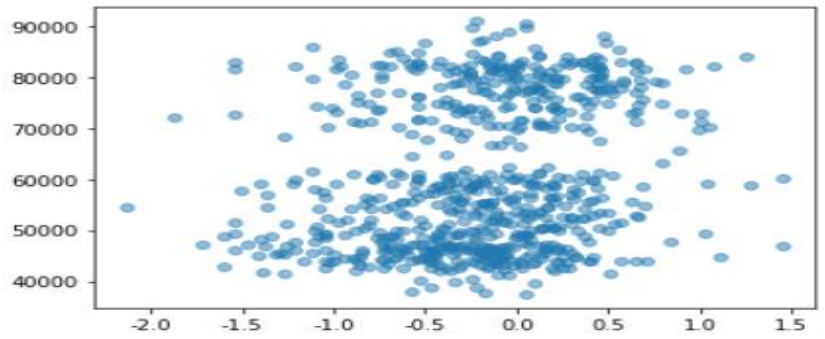
(0.2179462244294484, 5.961663467659075e-09)

OLS Regression Results

Dep. Variable:	종가	R-squared:	0.048
Model:	OLS	Adj. R-squared:	0.046
Method:	Least Squares	F-statistic:	34.71
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	5.96e-09
Time:	14:52:18	Log-Likelihood:	-7657.6
No. Observations:	698	AIC:	1.532e+04
Df Residuals:	696	BIC:	1.533e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.179e+04	566.623	109.053	0.000	6.07e+04	6.29e+04
긍부정비율	5819.6668	987.817	5.891	0.000	3880.208	7759.125

Omnibus:	451.761	Durbin-Watson:	0.089
Prob(Omnibus):	0.000	Jarque-Bera (JB):	55.713
Skew:	0.355	Prob(JB):	7.98e-13
Kurtosis:	1.812	Cond. No.	1.95



03. 가설 검정

결과 분석

기업 분석

	장전	장중
삼성 전자 뉴스의 긍부정 비율이 높을수록 삼성전자의 주가는 상승할 것이다.	X	X

03. 가설 검정

결과 분석

대, 소형주

1894 코스피 200 TOP 10

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.179
Model:	OLS	Adj. R-squared:	0.147
Method:	Least Squares	F-statistic:	5.488
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.00713
Time:	16:26:49	Log-Likelihood:	-104.56
No. Observations:	53	AIC:	215.1
Df Residuals:	50	BIC:	221.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0297	0.310	-0.096	0.924	-0.653	0.593
금리	0.7444	0.363	2.049	0.046	0.015	1.474
환율	-0.7062	0.261	-2.711	0.009	-1.229	-0.183

Omnibus:	18.224	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30.866
Skew:	1.061	Prob(JB):	1.98e-07
Kurtosis:	6.079	Cond. No.	2.20

1167 코스피 200 중소형주

OLS Regression Results

Dep. Variable:	등락률	R-squared:	0.110
Model:	OLS	Adj. R-squared:	0.075
Method:	Least Squares	F-statistic:	3.164
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.0536
Time:	16:26:44	Log-Likelihood:	-102.61
No. Observations:	53	AIC:	211.2
Df Residuals:	50	BIC:	217.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2034	0.299	0.681	0.499	-0.397	0.804
금리	0.6410	0.350	1.831	0.073	-0.062	1.344
환율	-0.4506	0.251	-1.795	0.079	-0.955	0.054

Omnibus:	11.908	Durbin-Watson:	1.514
Prob(Omnibus):	0.003	Jarque-Bera (JB):	28.435
Skew:	0.422	Prob(JB):	6.69e-07
Kurtosis:	6.488	Cond. No.	2.20

04

결론 및 제언

04. 결론 및 제언

1) 전체적으로 p-value값을 통해 가설을 검정할 수 있었지만, R-squared값이 유의미한 값이 나오지 않아 아쉬움

2) 경제 데이터 중에서도 노이즈한 데이터가 많이 있어 좀 더 질 좋은 데이터를 얻는다면 더 좋은 결과를 얻을 수 있을 것

3) BERT와 같은 pre-trained model을 사용하여 긍부정사전을 만들어보고자 함

4) 긍부정사전을 구축할 때 도메인에서 직접 추출한 데이터의 양을 늘린다면 더 유의미한 결과를 얻을 수 있을 것

Thank you