Newbie Investors! Should You Sell or Hold?

Buy/Sell Recommendation For Stock Market Beginners

2023-1 Team timeseries
Kim Min/ Kim Dong-hwan/ Seo Yoo-jin/ Lee Soo-rin/ Jang Da-yeon

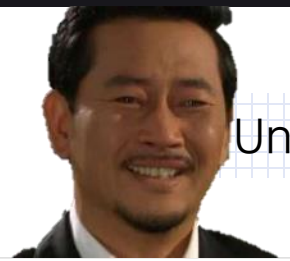
INDEX

1

Introduction

1 Introduction

Background of Topic Selection



Unprecedented Stock Market Volatility Due to COVID-19

1 Introduction

Background of Topic Selection

세계비즈 | 2022.06.27.

벼랑끝에 몰린 영끌·빚투족…금리오르고 주식·가상화폐 ↓

폭락에 **투자**자 피눈물…주담대 이자 부담 **주식**, 코인, 부동산 **투자**로 **피해**를 본 영끌·빚투족들이 연일… **코로나**19 이후 꾸준히 **주식투자**를 한 동학개미로서 자부심…

全南日報 전남일보 | 2022.06.28.

연이은 주식·가상화폐 하락에 '도박중독자' 증가

최근 광주·전남에서 **주식**과 가상화폐 **투자** 실패에 **피해**를 호소하는 사례가 증가하고 있다. 특히… **코로나**19 유행 전과 비교했을 때, 상담 건수가 2배 이상 증가한 …

아주경제 | 2022.08.25.

"고수익 보장, ○○주식·△△코인 사세요"…투자 사기 '횡행' [불법 …

◆글 싣는 순서 ① "고수익 보장, ○○**주식**·△△코인 사세요"…**투자** 사기 '횡행' ② 코인 **투자 피해**액 1년… **투자**자들에 따르면 이들 일당은 "**코로나**로 인해 **투자**설명…

全南日報 전남일보 | 2022.06.30.

코로나 폐업, 빚투 실패… 사회적 안전망 강화 필요

● **주식**·코인 **투자**자 '파탄 지경'빚투와 영끌(영혼까지 끌어모으다) 열풍이 한창이던 **주식**과 가상화폐 시장이 지난 5월부터 폭락하기 시작해 많은 **피해**자를 낳고 …

더스쿠프 | 2022.08.11.

[Weekly Global] 투자 귀재도 '하락장'에선…

버크셔 해서웨이가 **주식투자**로 57조원에 이르는 순손실을 냈다.[사진=뉴시스] **투자** 귀재도 '하락장'에선… 강서구 더스쿠프 기자 ksg@thescoop.co.kr [中 하이난 …

조세일보 PiCK | 2022.11.17. | 네이버뉴스

與 "금투세 유예하고 주식시장 개선해야"

2020년에는 **코로나**19 위기 극복을 위해 유동성이 풍부했던 시기지만 지금은 180도 다른 상황"이라고 강조했다. 금투세 도입과 주가 하락의 상관관계에 대해서는…

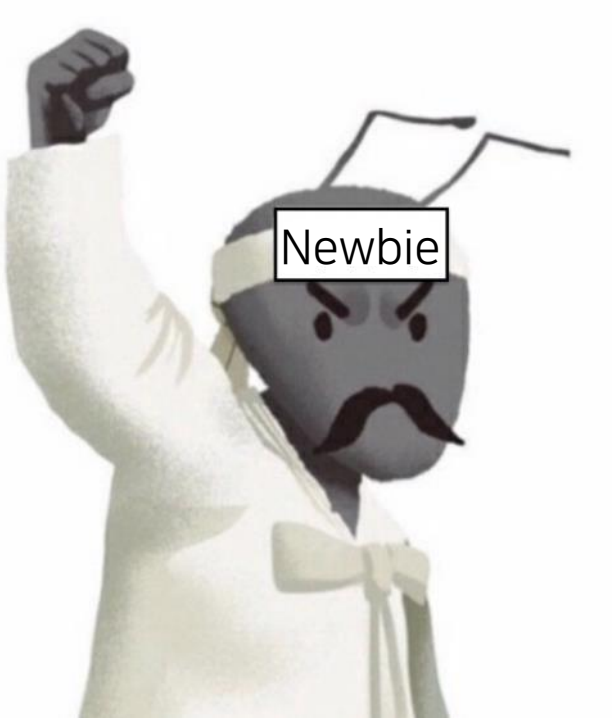
Interest in Stock Investment Rose During the Market Boom

As the Market Declined, Many Investors Suffered Significant Losses

1 Introduction

Background of Topic Selection

They say stocks are hot these days...
should I give it a try too??



Ah, everyone's like this because they don't study.
You've got to read the market trends!

what exactly are these market trends...
and how do you read them?

Especially for Beginners, It's Hard to Find Reliable Investment Information

Many Struggle with Knowing When to Buy or Sell

1 Introduction

Background of Topic Selection

How can we help novice investors who lack stock-related information?

Financial statements, cash flow, market conditions... it's all so complex! 😔

It would be great to have

simple and easy-to-understand information or indicators to refer to when investing!

1 Introduction

Background of Topic Selection

Team Time Series Topic Analysis

Buy/Sell Recommendation Service For Stock Market Beginners

Input information affecting stock price fluctuations into the model

→ the model learns to recommend buy/sell decisions

Providing investors with recommendations to buy or sell based on the current situation!

▶ Offering a indicator that provides insights/information for investment decisions

1 Introduction

Background of Topic Selection

~ Overview of Topic Analysis ~

Data Collection → Data Preprocessing → Variable Selection → Modeling → Performance Evaluation of Predictions

2

Data overview

2 Data overview

Collected Data

SK Hynix

Shinhan Financial Group

Hyundai Motor Company

Selected as analysis targets for

building a robust model applicable to various stocks

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Common indicators

- ✓ KOSPI data
- ✓ Bitcoin trading data
- ✓ Economic sentiment index
- ✓ News sentiment index
- ✓ Industrial production index
- ✓ Consumer price index
- ✓ Consumer confidence index
- ✓ Consumer sentiment index
- ✓ Unemployment rate
- ✓ Bank of Korea base rate
- ✓ Exchange rate

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Google Trends data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Data related to **Bitcoin** and **economic indicators** that are included **commonly** in modeling for all stocks.

Common indicators

- ✓ KOSPI data
- ✓ Bitcoin trading data
- ✓ Economic sentiment index
- ✓ News sentiment index
- ✓ Industrial production index
- ✓ Consumer price index
- ✓ Consumer confidence index
- ✓ Consumer sentiment index
- ✓ Unemployment rate
- ✓ Bank of Korea base rate
- ✓ Exchange rate

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Common indicators

- ✓ KOSPI data
- ✓ Bitcoin trading data
- ✓ Economic sentiment index
- ✓ News sentiment index
- ✓ Industrial production index
- ✓ Consumer price index
- ✓ Consumer sentiment index
- ✓ Unemployment rate
- ✓ Bank of Korea base rate
- ✓ Exchange rate

Transaction and **sentiment**-related data that varies for each individual stock

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Web Scraping

Common indicators

- ✓ KOSPI data
- ✓ Bitcoin trading data
- ✓ Economic sentiment index
- ✓ News sentiment index
- ✓ Industrial production index
- ✓ Consumer price index
- ✓ Consumer confidence index
- ✓ Consumer sentiment index
- ✓ Unemployment rate
- ✓ Bank of Korea base rate
- ✓ Exchange rate

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

Utilizing data that reflects public opinion

to grasp fluctuations based on sentiment and investor psychology!

2 Data overview

Collected Data

Individual indicators

- ✓ Stock price trend data
- ✓ Investor transaction performance data
- ✓ Foreign ownership data
- ✓ Short selling data
- ✓ Domestic news data
- ✓ English news data
- ✓ Naver Stock Discussion Forum data
- ✓ Naver search volume data

- ✓ ata
- ✓ Bitcoin trading data
- ✓ Economic sentiment index
- ✓ News sentiment index
- ✓ Industrial production index

Common indicators

- ✓ Consumer price index
- ✓ Consumer confidence index
- ✓ Consumer sentiment index
- ✓ Unemployment rate
- ✓ Bank of Korea base rate
- ✓ Exchange rate

Data collection period: 2017/06/08 ~ 2023/03/31

The data collection period is aligned with the launch date of the Naver Stock Discussion Forum service, which started on June 8, 2017.

2 Data overview

Collected Data

[Stock price trend]

- (Sources) Korea Exchange
- 1433 rows
- 11 columns

[Investor transaction performance]

- (Sources) Korea Exchange
- 1433 rows
- 5 columns

[Foreign ownership]

- (Sources) Korea Exchange
- 1433 rows
- 8 columns

[Short selling]

- (Sources) Korea Exchange
- 1433 rows
- 9 columns

[Domestic news]

- (Sources) Bigkinds
- about 80000 rows
- 19 columns

[English news]

- (Sources) CNN
- 약 17033 rows
- 2 columns

[Naver Stock Discussion Forum]

- (Sources) Naver Finance
- about 1000 rows
- 2 columns

[Naver search volume]

- (Sources) Naver Datalab
- 2131 rows
- 2 columns

2 Data overview

Collected Data

[KOSPI]

- (Sources) Korea Exchange
- 1433 rows
- 12 columns

[Bitcoin]

- (Sources) invest.com
- 2124 rows
- 7 columns

[Economic sentiment index]

- (Sources) Bank of Korea Economic Statistics System
- 71 rows (Monthly)
- 3 columns

[News sentiment index]

- (Sources) Bank of Korea Economic Statistics System
- 2124 rows
- 2 columns

[Industrial production index]

- (Sources) Bank of Korea Economic Statistics System
- 70 rows (Monthly)
- 2 columns

[Consumer price index]

- (Sources) Bank of Korea Economic Statistics System
- 71 rows (Monthly)
- 2 columns

2 Data overview

Collected Data

[Consumer confidence index]

- (Sources) invest.com
- 71 rows (월별)
- 5 columns

[Consumer sentiment index]

- (Sources) Bank of Korea Economic Statistics System
- 2124 rows
- 2 columns

[Unemployment rate]

- (Sources) KOSIS
- 71 rows (월별)
- 4 columns

[Bank of Korea base rate]

- (Sources) Bank of Korea Economic Statistics System
- 2124 rows
- 2 columns

[Exchange rate]

- (Sources) Bank of Korea Economic Statistics System
- 2124 rows
- 5 columns

LET'S GO

3

Data Preprocessing

3 Data Preprocessing

Sentiment analysis of Korean news articles

Sentiment Analysis

The technology of analyzing and determining the emotions of the person who wrote the document

e.g., Analysis of user product reviews on online shopping platforms

3 Data Preprocessing

Sentiment analysis of Korean news articles

Sentiment Analysis

The technology of analyzing and determining the emotions of the person who wrote the document

e.g., Analysis of user product reviews on online shopping platforms

Headline of the news article	Result
현대차 기아, 유럽 전기차 접수 판매 1위 휩쓸어	positive (1)
'그냥 쉰' 무기력한 청년 50만명, 이대론 미래 없다	negative (-1)
동물성 원료를 비건으로 바꾸는 기술, 비건 화학	neutral (0)

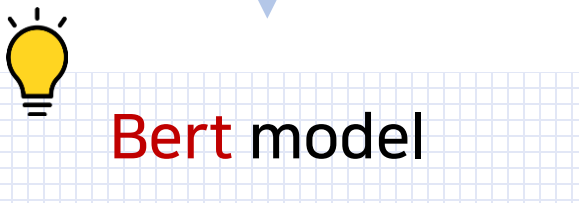
3 Data Preprocessing

Sentiment analysis of Korean news articles

Sentiment Analysis

The technology of analyzing and determining the emotions of the person who wrote the document

e.g., Analysis of user product reviews on online shopping platforms



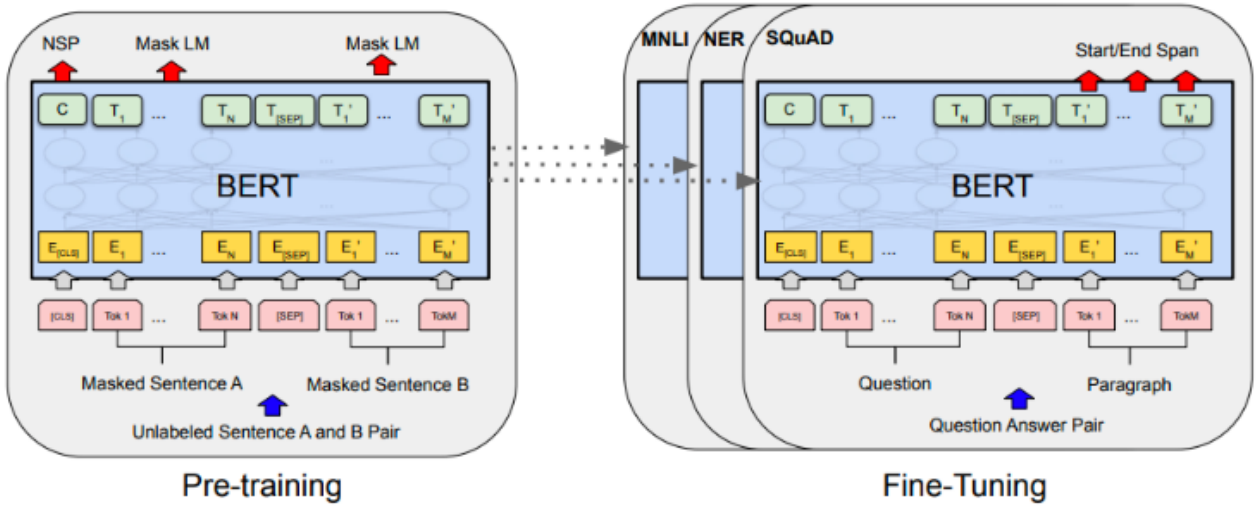
3 Data Preprocessing

Sentiment analysis of Korean news articles

Bert

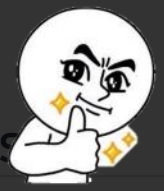
Training a language model using pre-trained large-scale unlabeled data with no labeling.

A structure where multiple layers of encoder are added to the Transformer architecture.



3 Data Preprocessing

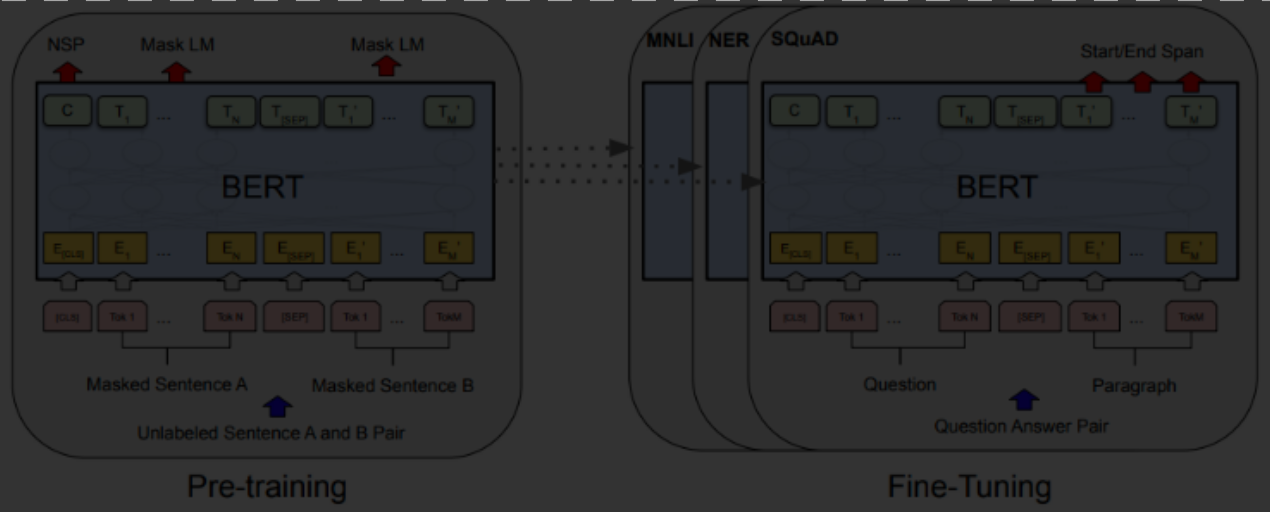
Sentiment analysis of Korean news articles



Advantages of the BERT model

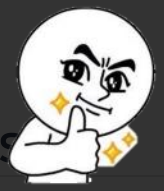
Bert

1. Utilizing embeddings that reflect context
2. Using a Subword tokenizer to tokenize frequently used words and less common words differently.



3 Data Preprocessing

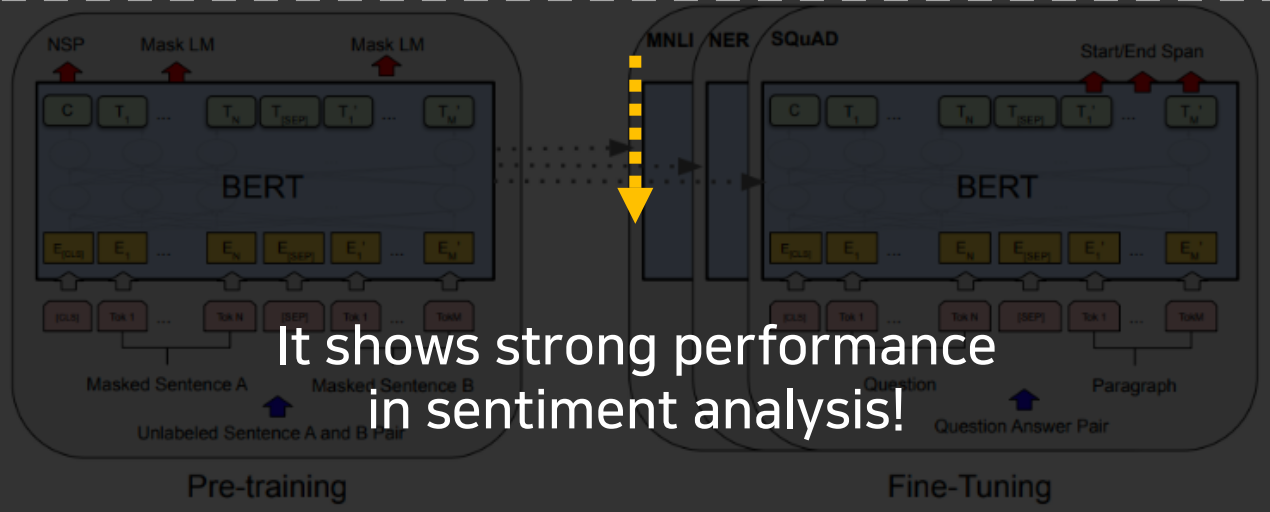
Sentiment analysis of Korean news articles



Advantages of the BERT model

Bert

1. Utilizing embeddings that reflect context
2. Using a Subword tokenizer to tokenize frequently used words and less common words differently.



It shows strong performance in sentiment analysis!

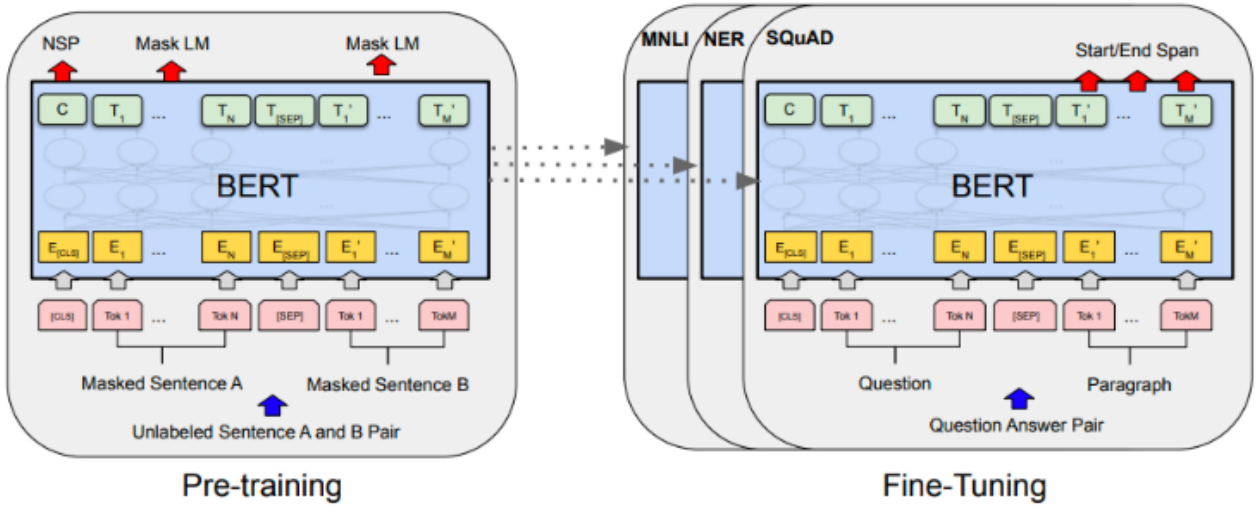
3 Data Preprocessing

Sentiment analysis of Korean news articles

Bert

Training a language model using pre-trained large-scale unlabeled data with no labeling.

A structure where multiple layers of encoder are added to the Transformer architecture.



3 Data Preprocessing

Sentiment analysis of Korean news articles

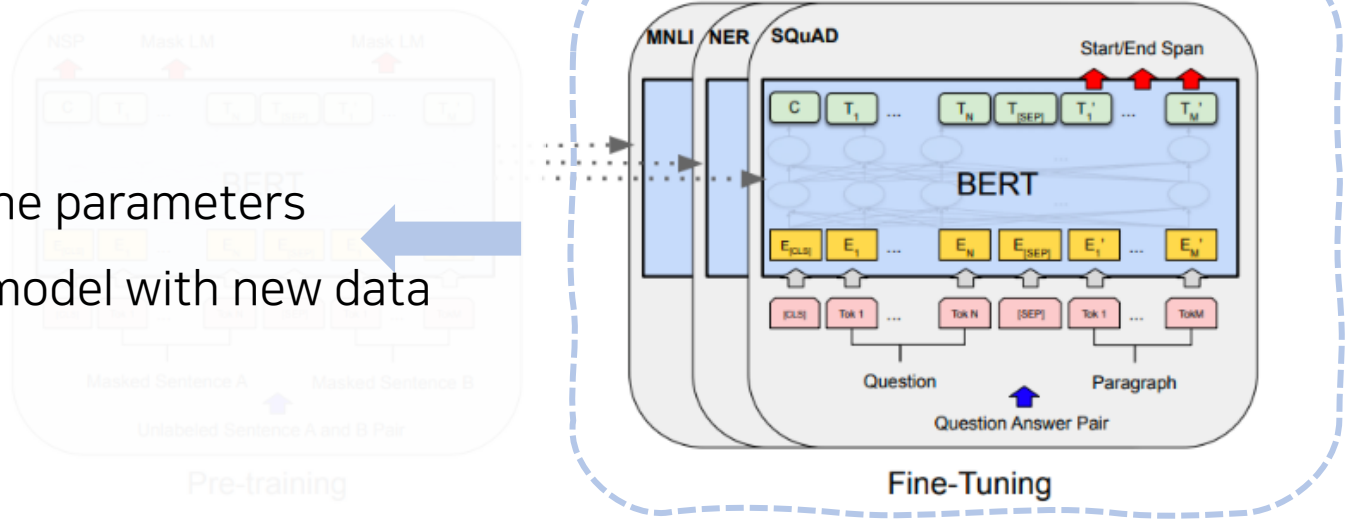
Bert

Training a language model using pre-trained large-scale unlabeled data with no labeling.

A structure where multiple layers of encoder are added to the Transformer architecture.

Fine-tuning the parameters

of the pre-trained model with new data



3 Data Preprocessing

Sentiment analysis of Korean news articles

train data A dataset consisting of 4,846 economic news articles labeled as 'neutral', 'negative', or 'positive'

Headline	label
Gran에 따르면 … 러시아로 옮길 계획 없다.	neutral
국제 전자산업 회사 … 직원 수를 줄였다고 보도했다.	negative
새로운 생산공장으로 인해 … 생산 수익성을 높일 것이다.	positive

3 Data Preprocessing

Sentiment analysis of Korean news articles

train data A dataset consisting of 4,846 economic news articles labeled as 'neutral', 'negative', or 'positive'

Headline	label
Gran에 따르면 … 러시아로 옮길 계획 없다.	neutral
국제 전자산업 회사 … 직원 수를 줄였다고 보도했다.	negative
새로운 생산공장으로 인해 … 생산 수익성을 높일 것이다.	positive

↓ negative: -1 neutral: 0 positive : 1

Headline	label
Gran에 따르면 … 러시아로 옮길 계획 없다.	0
국제 전자산업 회사 … 직원 수를 줄였다고 보도했다.	-1
새로운 생산공장으로 인해 … 생산 수익성을 높일 것이다.	1

3 Data Preprocessing

Sentiment analysis of Korean news articles

Data Headline data from 2017-06-08 ~ to 2023-03-31.

e.g., Shinhan Financial Group News Headlines

Date	Headline
2017-06-08	[fnRASSI] 장마감, 거래소 하락 종목 (신한 -8.4%)
2017-06-08	'오늘의 증시 메모 [6월 8일]
⋮	⋮
2023-03-31	신한금융 "데이터센터 전력 재생에너지로 조달 "
2023-03-31	신한금융, 데이터센터 전력 100% 재생 에너지 추진

3 Data Preprocessing

Sentiment analysis of Korean news articles

Data Headline data from 2017-06-08 ~ to 2023-03-31.

e.g., Shinhan Financial Group News Headlines

Date	Headline	label
2017-06-08	[fnRASSI] 장마감, 거래소 하락 종목 (신한 -8.4%)	-1
2017-06-08	'오늘의 증시 메모 [6월 8일]	0
⋮	⋮	
2023-03-31	신한금융 "데이터센터 전력 재생에너지로 조달 "	1
2023-03-31	신한금융, 데이터센터 전력 100% 재생 에너지 추진	1

Labeling through the trained model

3 Data Preprocessing

Sentiment analysis of Korean news articles

Data Headline data from 2017-06-08 ~ to 2023-03-31.

e.g., Shinhan Financial Group News Headlines

news.groupby('날짜')['label'].mean()

Adding a sentiment score column

by grouping by date.

Date	Sentiment score
2017-06-08	0.148148
2017-06-09	0.142857
⋮	⋮
2023-03-30	0.166667
2023-03-31	0.384615

3 Data Preprocessing

Sentiment analysis of Foreign news articles

Data

The data comprises a total of 9,995 rows, crawled from CNN from July 2017 to March 2023

Date	Headline
2017-07-19	US general warns of … control killer robots
2017-07-21	Pompeo signals want for N. Korea regime change
⋮	⋮
2023-03-25	How AI turned the ancient sport of Go upside down
2023-03-29	US & South Korea stage joint military drills

3 Data Preprocessing

Sentiment analysis of Foreign news articles

Data

The data comprises a total of 9,995 rows, crawled from CNN from July 2017 to March 2023

Date	Headline
2017-07-19	US general warns of … control killer robots
2017-07-21	Pompeo signals want for N. Korea regime change
⋮	⋮
2023-03-25	How AI turned the ancient sport of Go upside down
2023-03-29	US & South Korea stage joint military drills

3 Data Preprocessing

Sentiment analysis of Foreign news articles

Data

The data comprises a total of 9,995 rows, crawled from CNN from July 2017 to March 2023

D	
2017-07-19	US general warns of ··· control killer robots
2017-07-21	Pompeo signals want for N. Korea regime change
⋮	⋮
2023-03-25	How AI turned the ancient sport of Go upside down
2023-03-29	US & South Korea stage joint military drills

What is nltk???

A Python package for natural language processing, text analysis, and text mining

Performing sentiment analysis on

English sentences using nltk.sentiment.vader

Using the compound score ranging from -1 to 1.

Sentiment analysis using nltk

3 Data Preprocessing

Sentiment analysis of Foreign news articles

Data

The data comprises a total of 9,995 rows, crawled from CNN from July 2017 to March 2023

Date	Headline	score
2017-07-19	US general warns of … control killer robots	-0.6908
2017-07-21	Pompeo signals want for N. Korea regime change	0.0772
⋮	⋮	
2023-03-25	How AI turned the ancient sport of Go upside down	0
2023-03-29	US & South Korea stage joint military drills	0

Get a score using nltk.sentiment.vader

3 Data Preprocessing

Sentiment analysis of Foreign news articles

Data

The data comprises a total of 9,995 rows, crawled from CNN from July 2017 to March 2023

article.groupby('날짜')['score'].mean()

Adding a sentiment score column

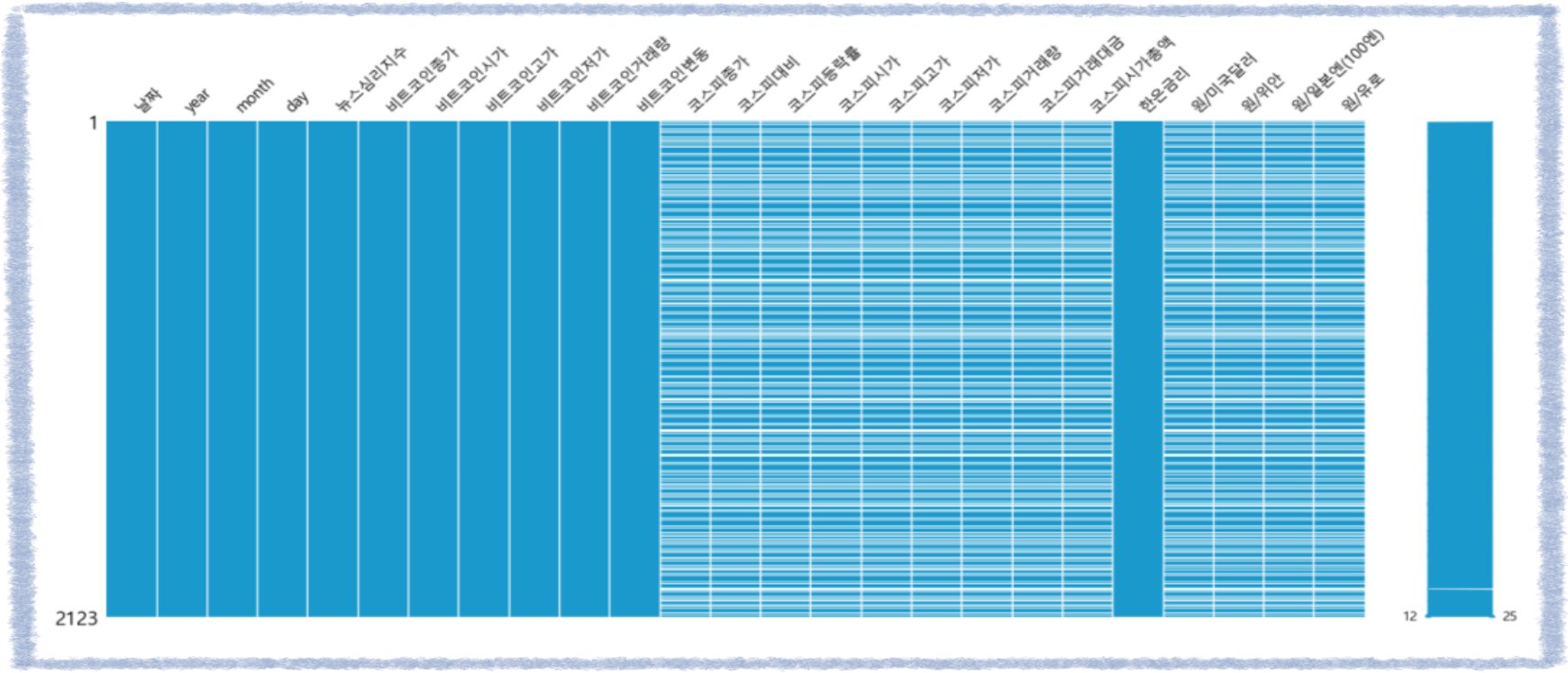
by grouping by date

Date	Sentiment score
2017-07-14	0.365775
2017-07-15	0.261127
⋮	⋮
2023-03-30	0.120400
2023-03-31	0.079550

3 Data Preprocessing

Removing missing values and interpolating

Visualizing daily data missing values

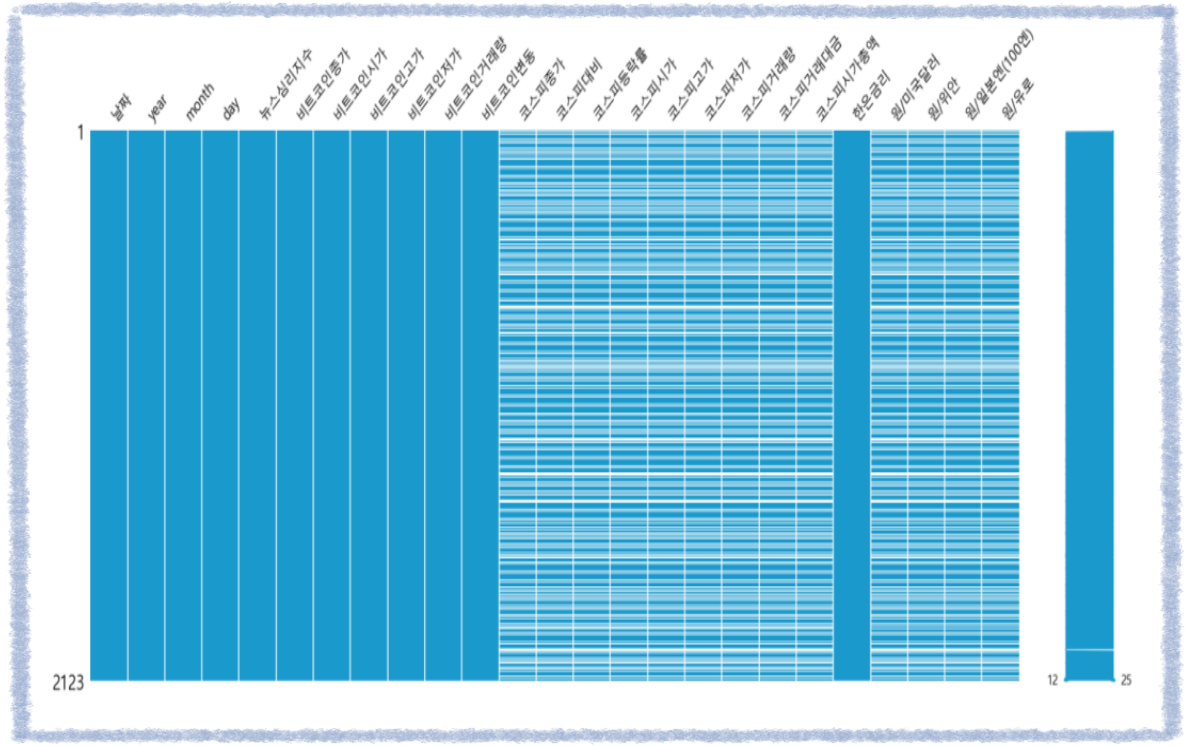


Find out the presence of a significant amount of missing values

3 Data Preprocessing

Removing missing values and interpolating

Visualizing daily data missing values



Drop

✓ Weekend data

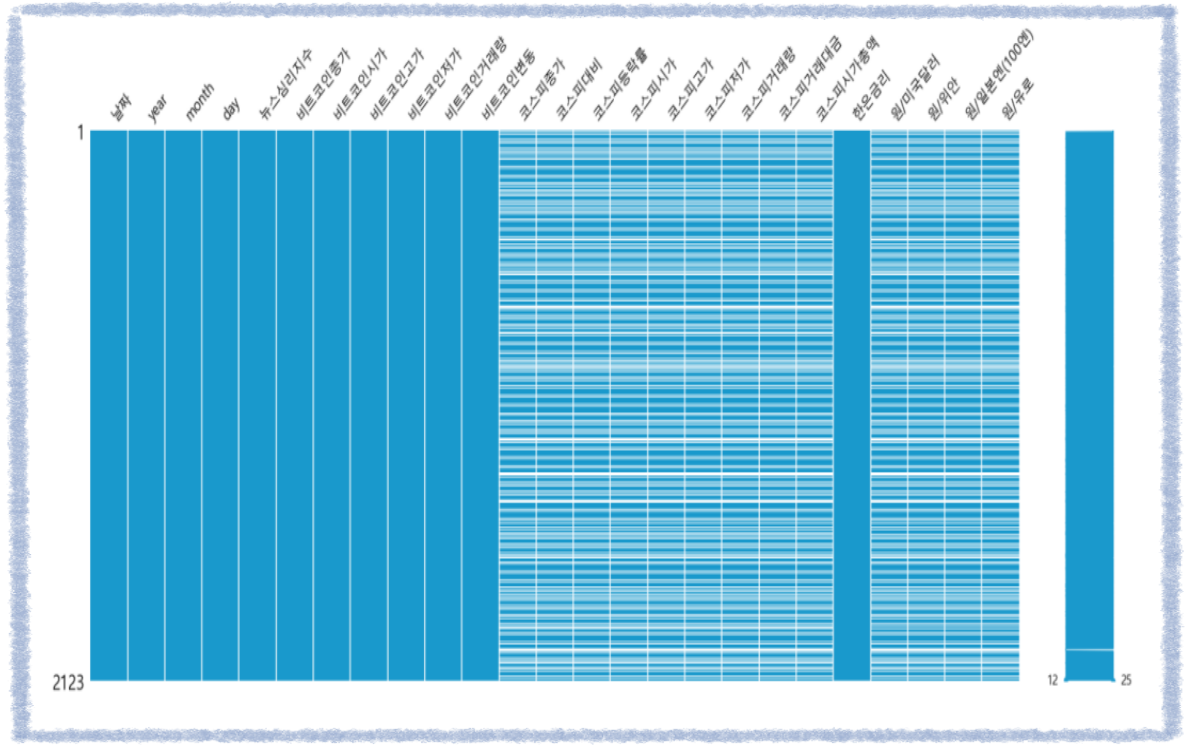
→ Weekend Data – Missing Values Due to Market Closure

→ Drop

3 Data Preprocessing

Removing missing values and interpolating

Visualizing daily data missing values



Drop

- ✓ Weekend data
  - → Weekend Data – Missing Values Due to Market Closure
  - → Drop

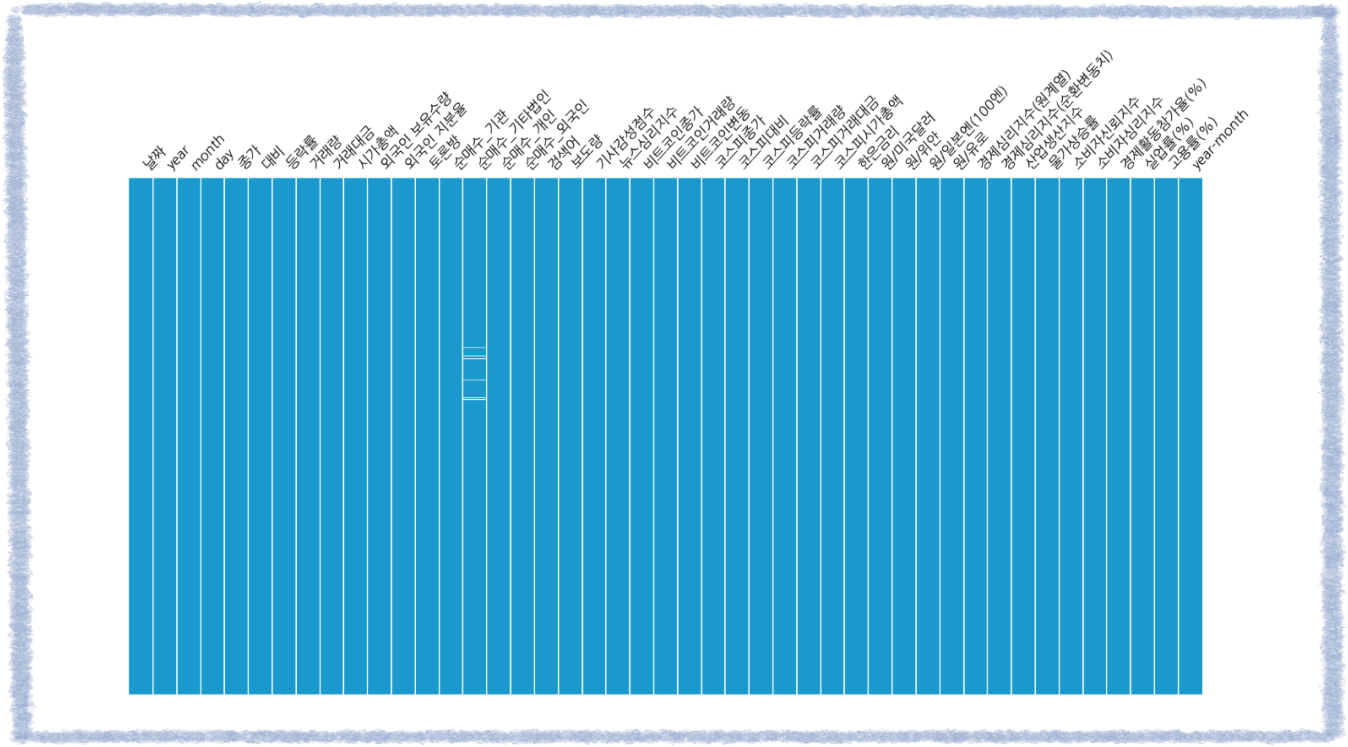
Interpolation

- ✓ Article coverage volume → Replace to 0
- ✓ Discussion forum → Replace to 0
- ✓ Sentiment score → mean

3 Data Preprocessing

Removing missing values and interpolating

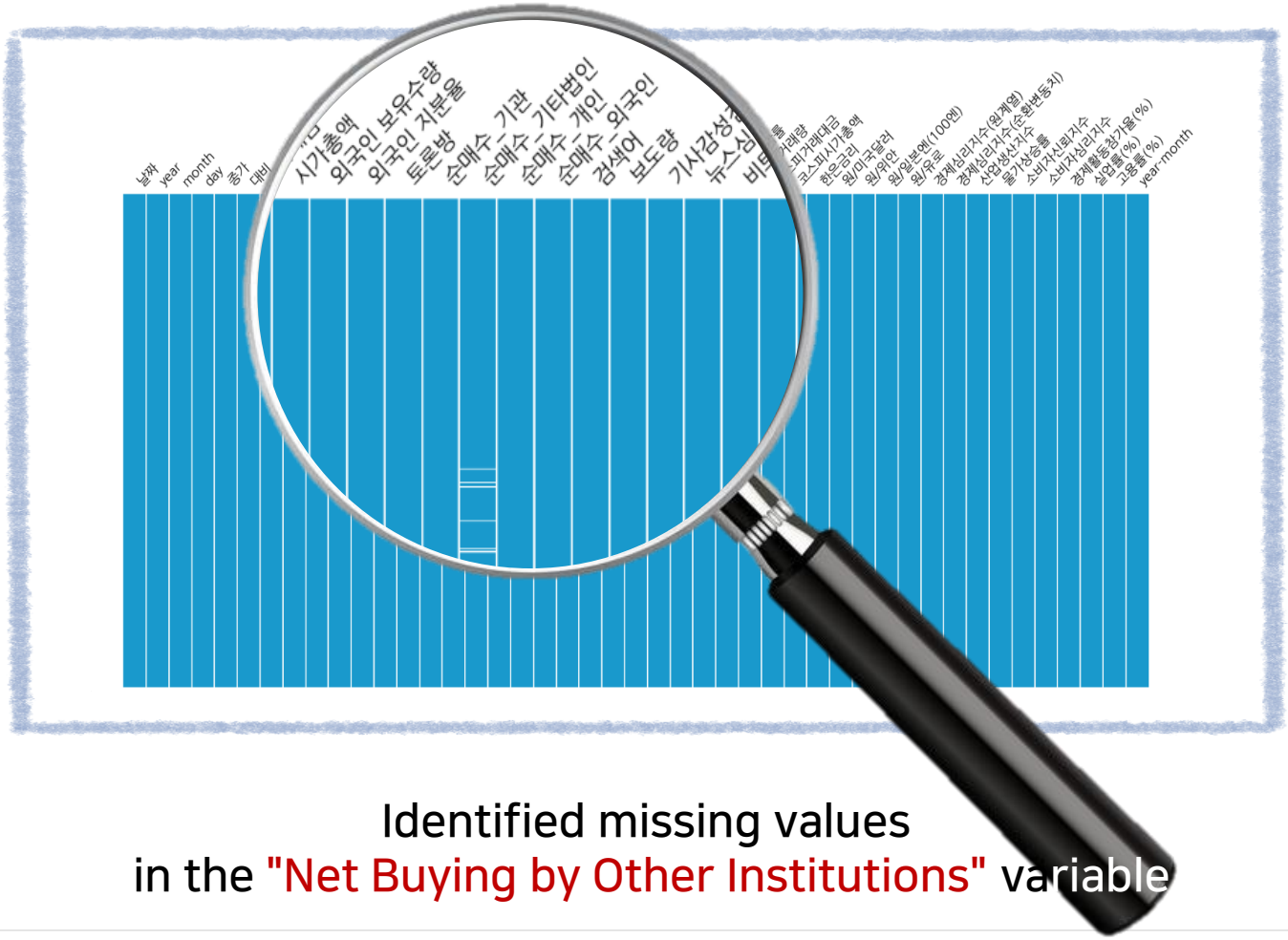
Visualizing daily data missing values



3 Data Preprocessing

Removing missing values and interpolating

Visualizing daily data missing values

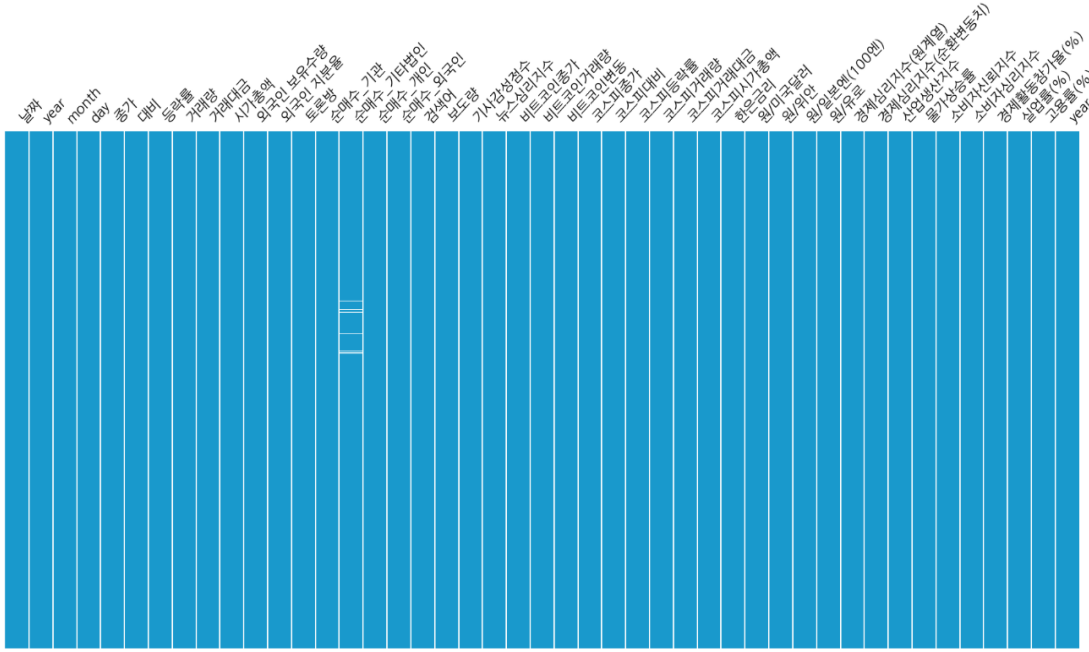


Identified missing values
in the "Net Buying by Other Institutions" variable

3 Data Preprocessing

Removing missing values and interpolating

Visualizing daily data missing values



Interpolation

✓ Net_buying_other_institutions

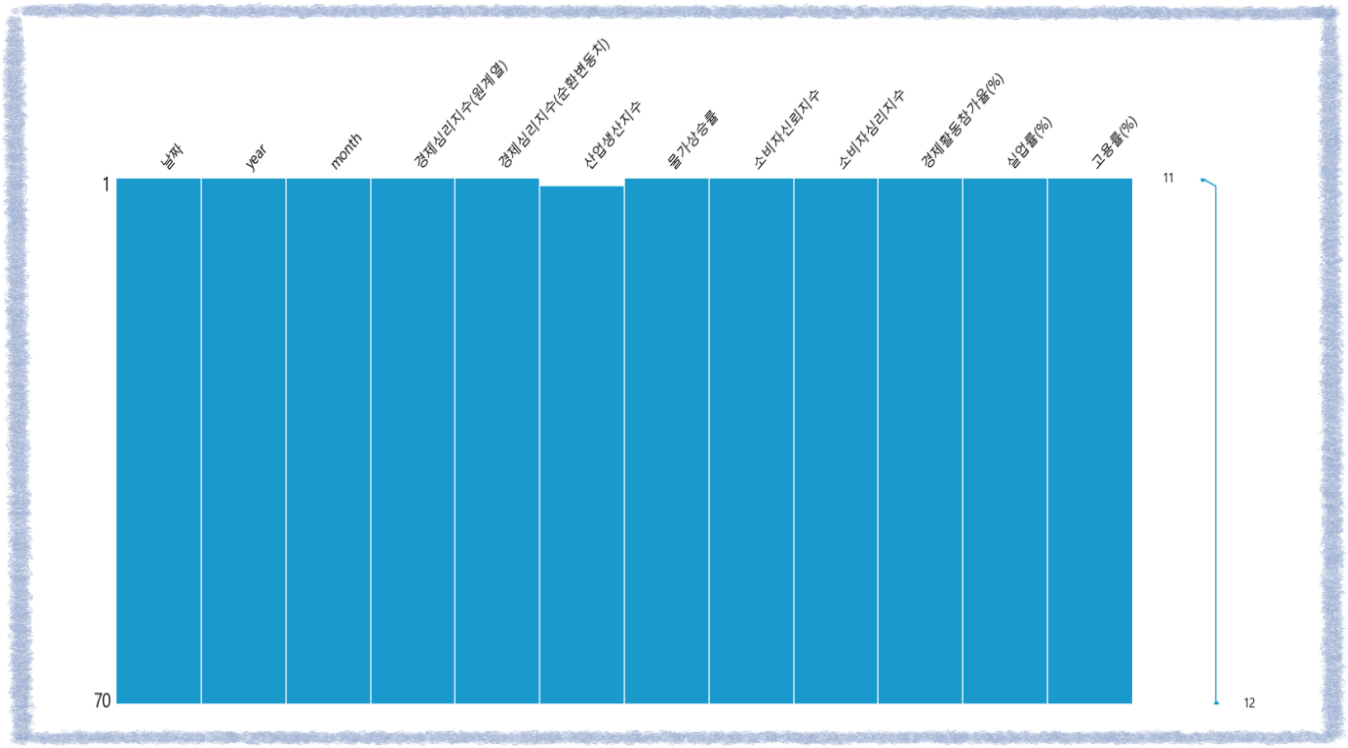
⟶ Interpolating using cubic spline method

third-degree polynomial curve

3 Data Preprocessing

Removing missing values and interpolating

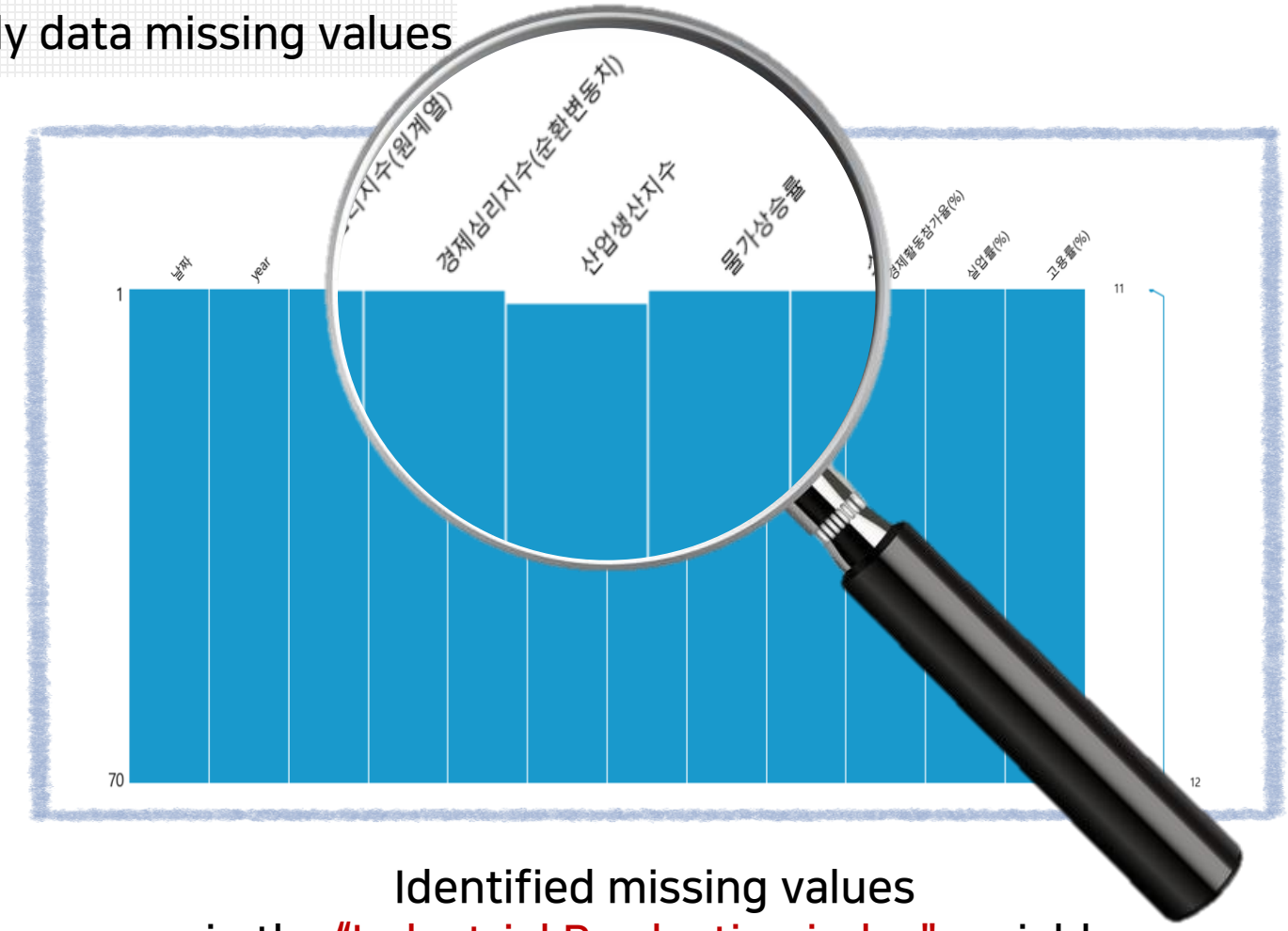
Visualizing monthly data missing values



3 Data Preprocessing

Removing missing values and interpolating

Visualizing monthly data missing values

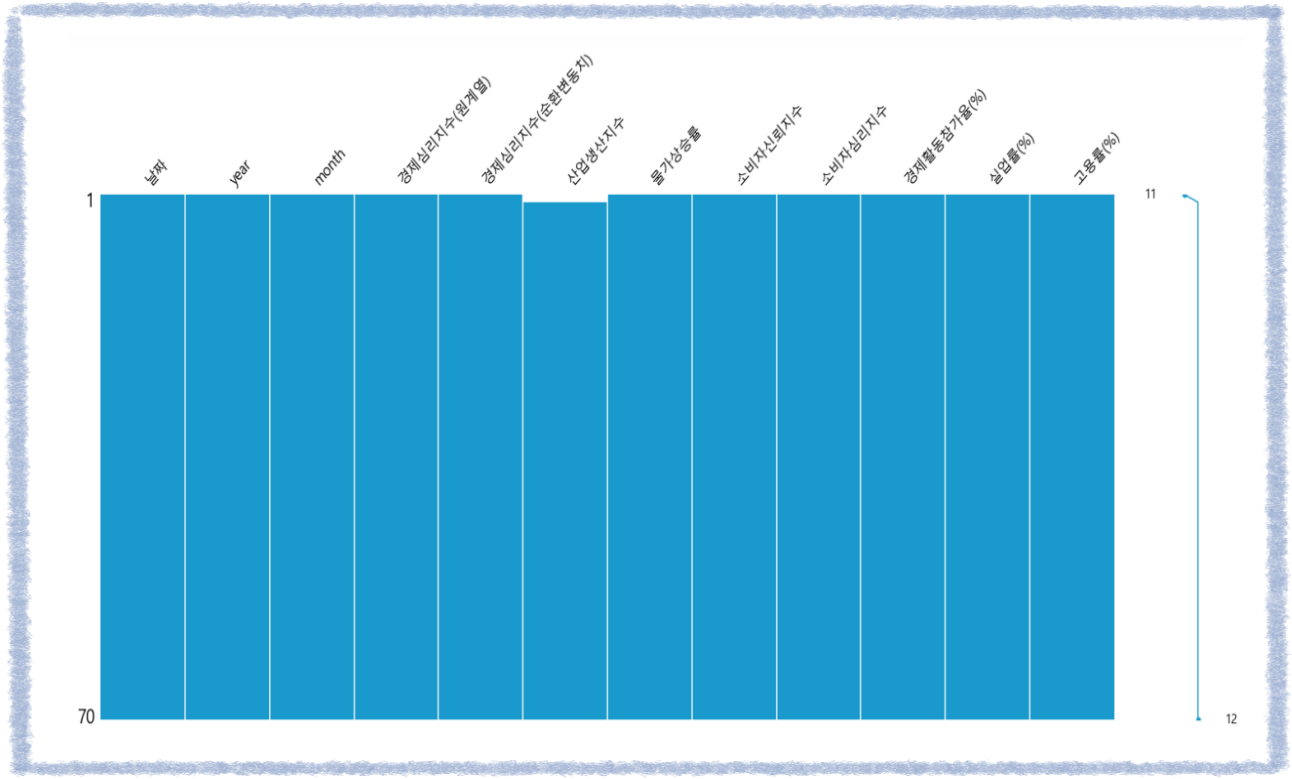


Identified missing values
in the “Industrial Production index" variable

3 Data Preprocessing

Removing missing values and interpolating

Visualizing monthly data missing values



Interpolation

✓ Industrial Production Index

→ Average of the previous month,
the month before the previous month,
and the same month of the previous year

3 Data Preprocessing

Final data set

Removing column from individual indicators

Removing a total of 14 unnecessary variables from
individual datasets of Shinhan Financial Group, SK Hynix, and Hyundai Motor

Short_Selling_Volume_Total_Trading_Volume/Balance Quantity/Total Trading Amount/Balance Amount, Opening Price, High Price, Listed Shares, Bitcoin Opening Price, Bitcoin High Price, Bitcoin Low Price, KOSPI Opening Price, KOSPI High Price, KOSPI Low Price

3 Data Preprocessing

Final data set

Removing column from individual indicators

Removing a total of 14 unnecessary variables from
individual datasets of Shinhan Financial Group, SK Hynix, and Hyundai Motor

Standardizing data types to numeric

Converting 8 variables with object data type to numeric in the daily dataset

3 Data Preprocessing

Final data set

Removing column from individual indicators

Removing a total of 14 unnecessary variables from
individual datasets of Shinhan Financial Group, SK Hynix, and Hyundai Motor

Standardizing data types to numeric

Converting 8 variables with object data type to numeric in the daily dataset

Bitcoin closing price, Bitcoin opening price, Bitcoin highest price, Bitcoin lowest price,
Bitcoin trading volume, Bitcoin volatility, KRW/USD, KRW/JPY (100JPY)

3 Data Preprocessing

Final data set

Removing column from individual indicators

Removing a total of 14 unnecessary variables from

individual datasets of Shinhan Financial Group, SK Hynix, and Hyundai Motor

Standardizing data types to numeric

Converting 8 variables with object data type to numeric in the daily dataset

Merge data

merge individual indicators and common indicators data

SK Hynix, Shinhan Financial Group,
Hyundai Motor Company

3 Data Preprocessing

Final data set

Removing column from individual indicators

Removing a total of 14 unnecessary variables from
individual datasets of Shinhan Financial Group, SK Hynix, and Hyundai Motor

Standardizing data types to numeric

Converting 8 variables with object data type to numeric in the daily dataset

Merge data

merge individual indicators and common indicators data

Sorting

Sort by ascending order of ‘date’ column

3 Data Preprocessing

Final data set

SK Hynix 1432rows 44columns

날짜	year	month	day	종가	대비	등락률	거래량	거래대금	시가총액	...	원/유로	경제심리지수(원계열)	경제심리지수(순환변동치)	산업생산지수	물가상승률	소비자신뢰지수	소비자심리지수	경제활동참가율(%)	실업률(%)
2017-06-08	2017	6	8	56700.0	200.0	0.35	3943311.0	2.232900e+11	4.127773e+13	...	1263.48	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-09	2017	6	9	58300.0	1600.0	2.82	4799560.0	2.770599e+11	4.244254e+13	...	1259.72	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-12	2017	6	12	57500.0	-800.0	-1.37	2650404.0	1.520987e+11	4.186014e+13	...	1256.53	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-13	2017	6	13	58800.0	1300.0	2.26	3431811.0	1.987173e+11	4.280654e+13	...	1261.29	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-14	2017	6	14	59300.0	500.0	0.85	4441904.0	2.638198e+11	4.317054e+13	...	1265.33	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

3 Data Preprocessing

Final data set

Hyundai

1432rows 44columns

날짜	year	month	day	종가	대비	등락률	거래량	거래대금	시가총액	...	원/유로	경제심리지수(원계열)	경제심리지수(순환변동치)	산업생산지수	물가상승률	소비자신뢰지수	소비자심리지수	경제활동참가율(%)	실업률(%)
2017-06-08	2017	6	8	160000.0	1000.0	0.63	548233.0	8.753964e+10	3.524424e+13	...	1263.48	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-09	2017	6	9	160000.0	0.0	0.00	343795.0	5.502646e+10	3.524424e+13	...	1259.72	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-12	2017	6	12	162500.0	2500.0	1.56	390351.0	6.332846e+10	3.579493e+13	...	1256.53	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-13	2017	6	13	164500.0	2000.0	1.23	548527.0	9.082350e+10	3.623548e+13	...	1261.29	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-14	2017	6	14	169000.0	4500.0	2.74	638693.0	1.077933e+11	3.722672e+13	...	1265.33	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

3 Data Preprocessing

Final data set

Shinhan

1432rows 44columns

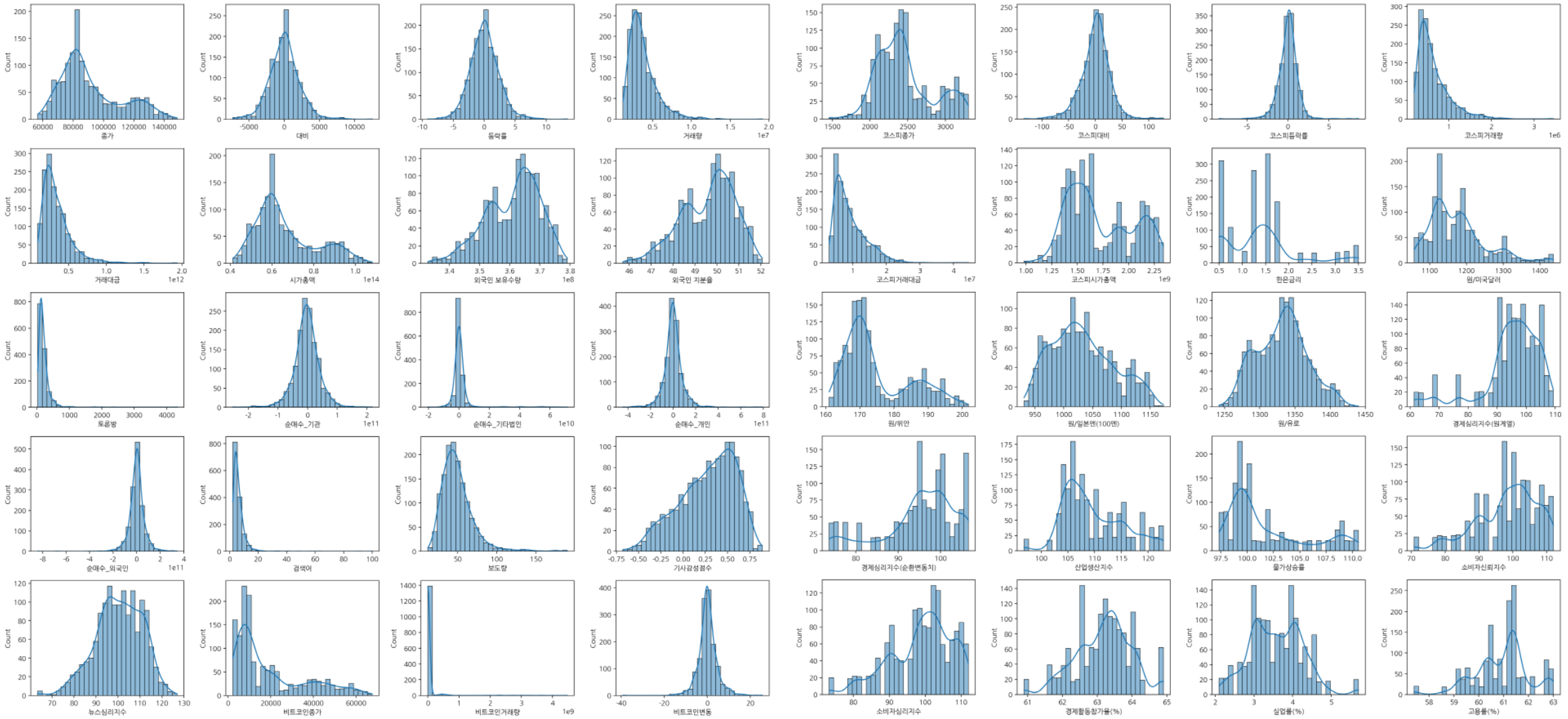
날짜	year	month	day	종가	대비	등락률	거래량	거래대금	시가총액	...	원/유로	경제심리지수(원계열)	경제심리지수(순환변동치)	산업생산지수	물가상승률	소비자신뢰지수	소비자심리지수	경제활동참가율(%)	실업률(%)
2017-06-08	2017	6	8	49400.0	100.0	0.20	1342162.0	6.623365e+10	2.342546e+13	...	1263.48	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-09	2017	6	9	49900.0	500.0	1.01	974960.0	4.861680e+10	2.366256e+13	...	1259.72	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-12	2017	6	12	50300.0	400.0	0.80	1279975.0	6.462560e+10	2.385224e+13	...	1256.53	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-13	2017	6	13	50100.0	-200.0	-0.40	437819.0	2.193719e+10	2.375740e+13	...	1261.29	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
2017-06-14	2017	6	14	50700.0	600.0	1.20	996192.0	5.008074e+10	2.404192e+13	...	1265.33	99.8	99.6	103.9	97.338	111.0	110.8	63.9	3.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

4

EDA

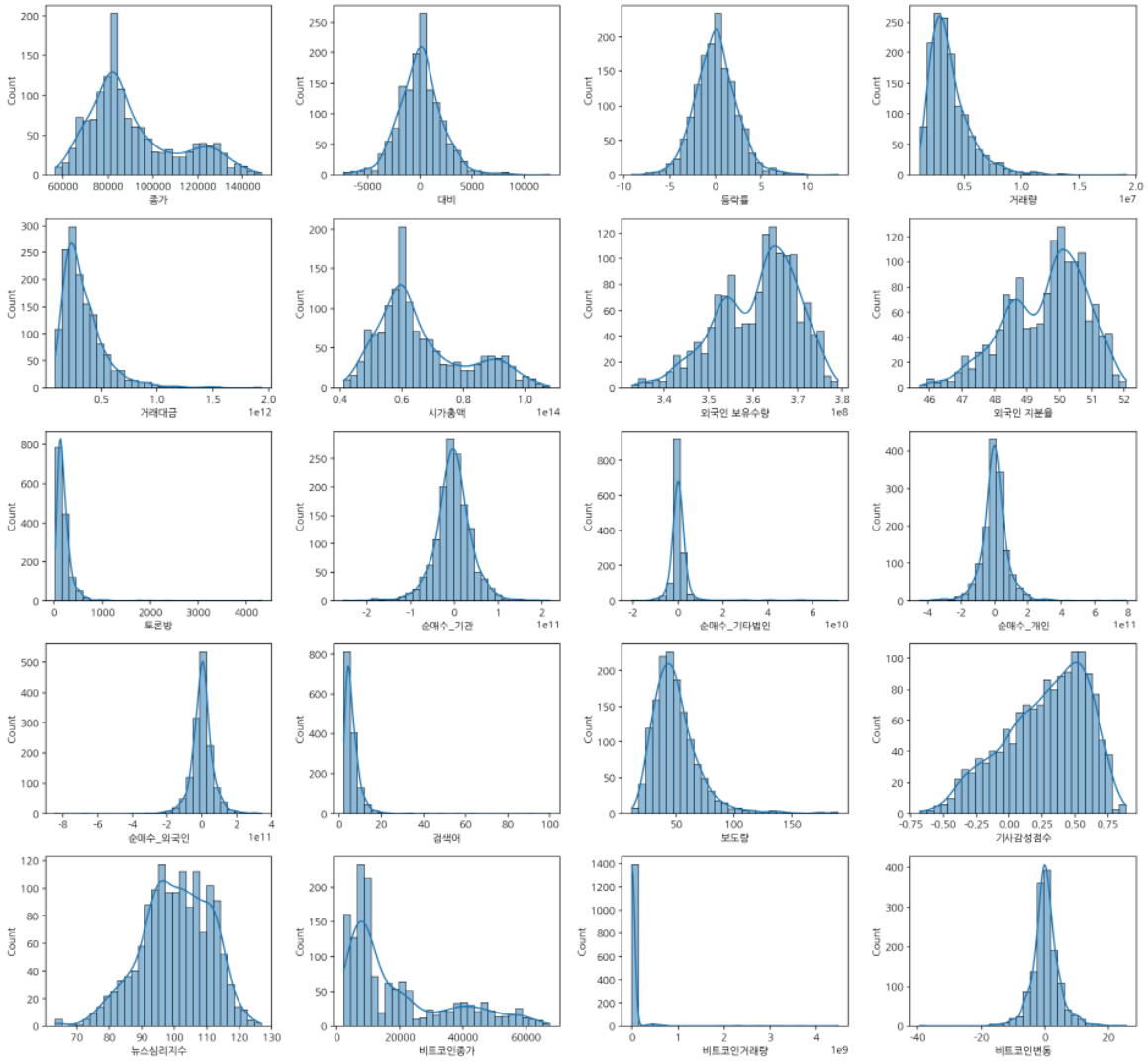
4 EDA : SKHynix

Distribution of X variables



4 EDA : SKHynix

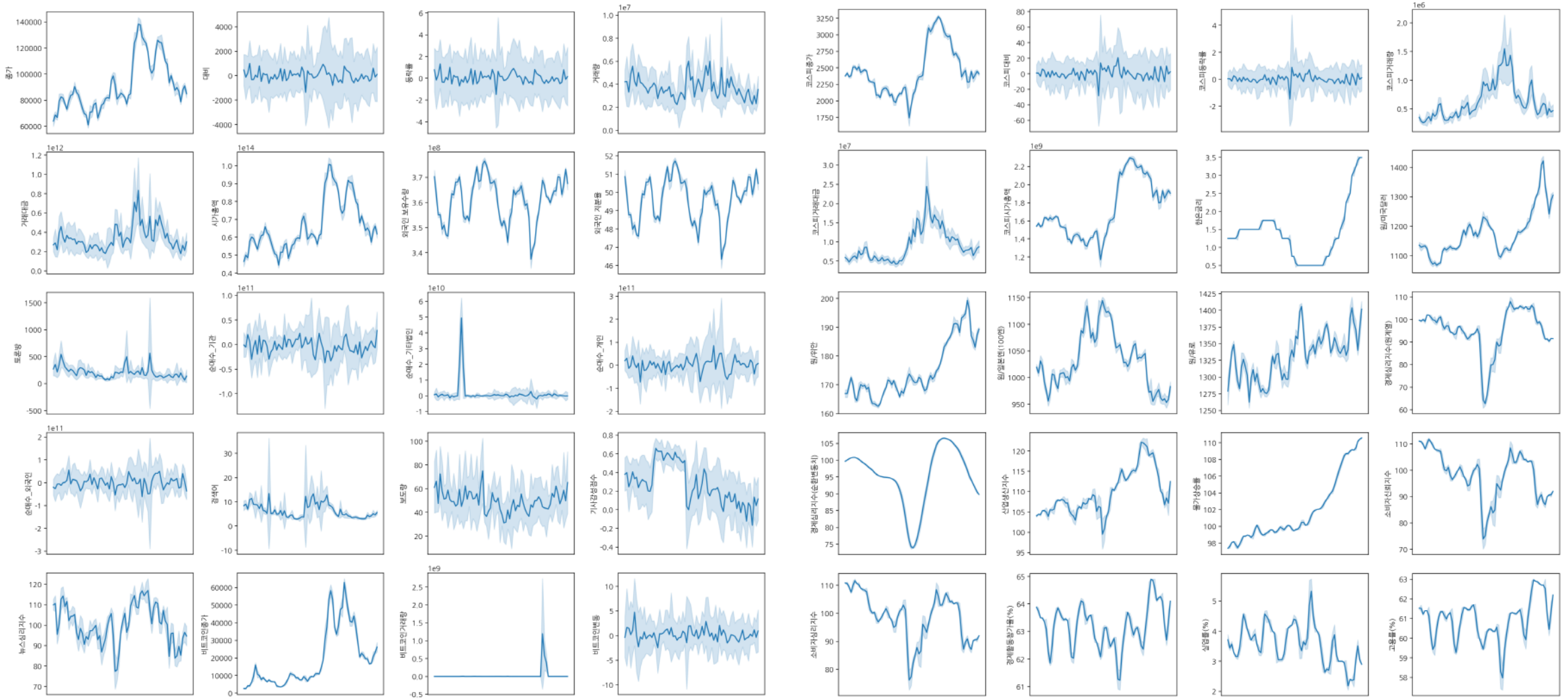
Distribution of X variables



There are no visible outliers observed

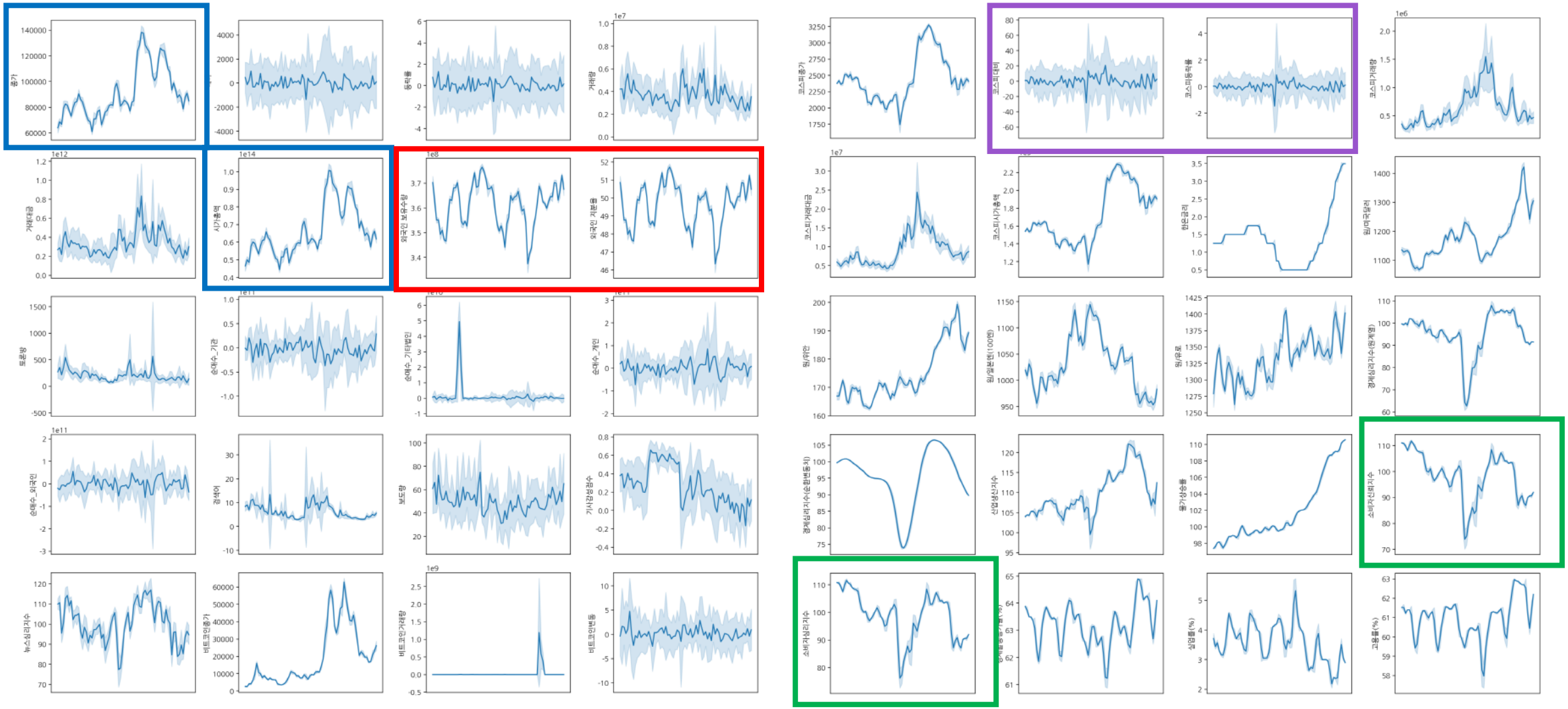
4 EDA : SKHynix

TS plot



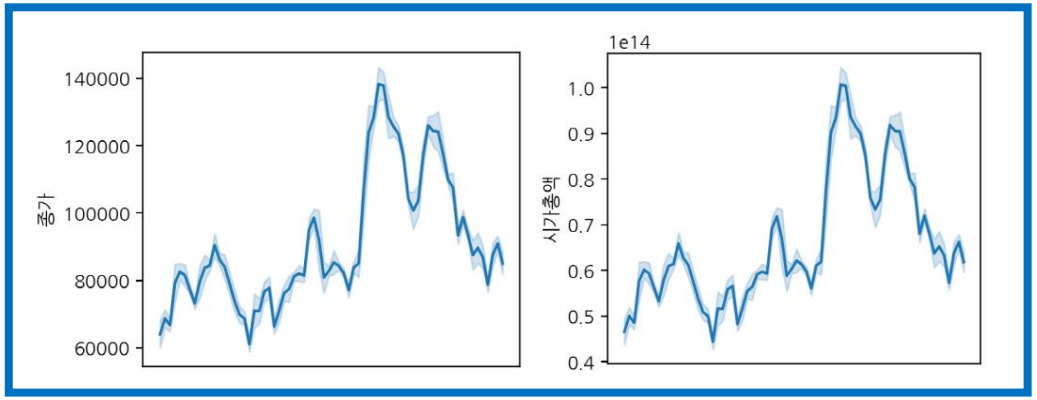
4 EDA : SKHynix

TS plot : plot with high similarity

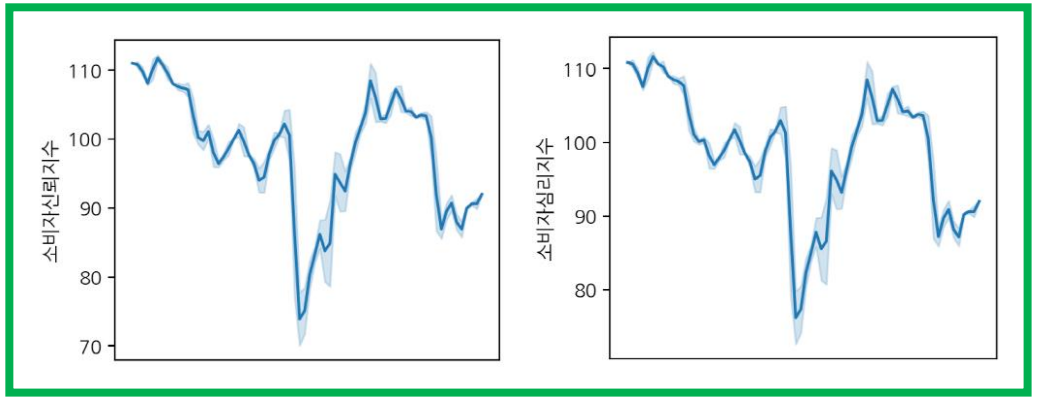


4 EDA : SKHynix

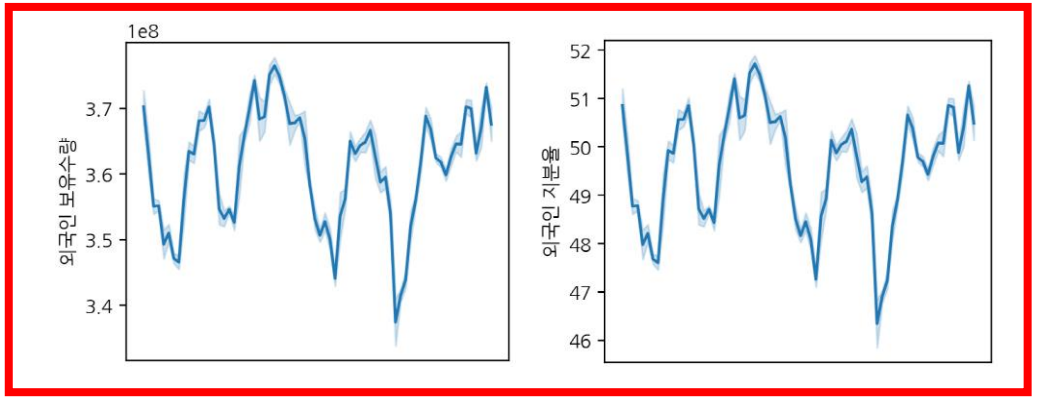
TS plot : plot with high similarity



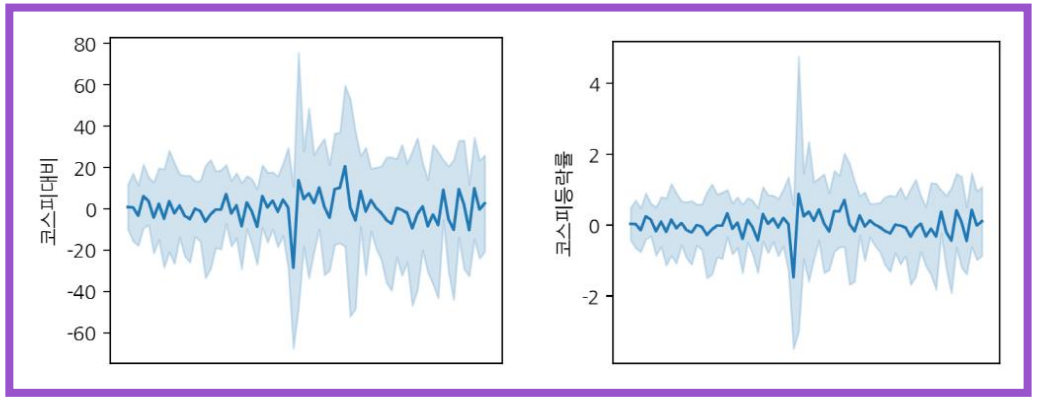
Closing price & market capitalization



Consumer confidence index & consumer sentiment index



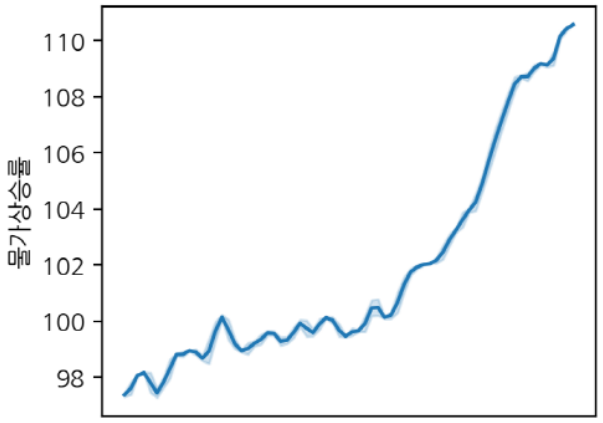
Foreign investor holding quantity & foreign investor ownership percentage



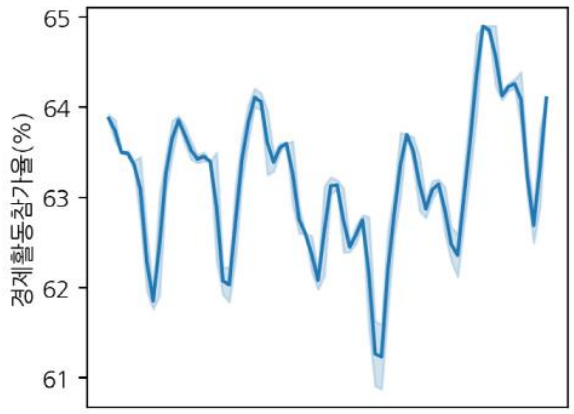
KOSPI change & KOSPI fluctuation rate

4 EDA : SKHynix

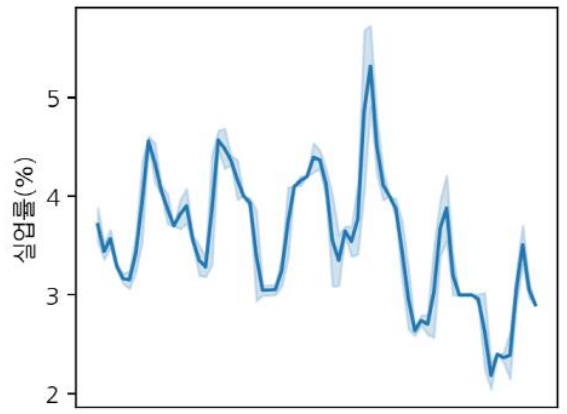
TS plot : Trend & Seasonality



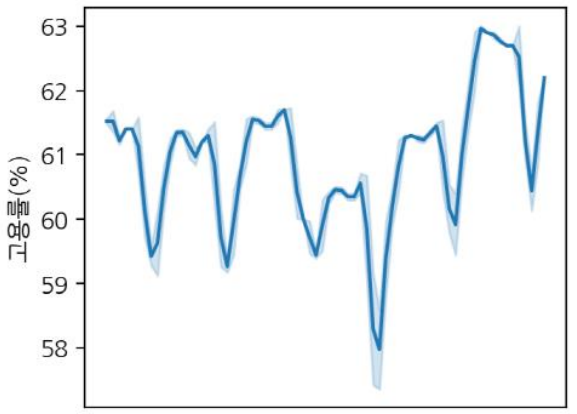
Inflation rate



Labor force participation rate (%)



Unemployment rate (%)



Employment rate (%)

Inflation rate shows an increasing trend

Labor force participation rate(%), Unemployment rate(%),

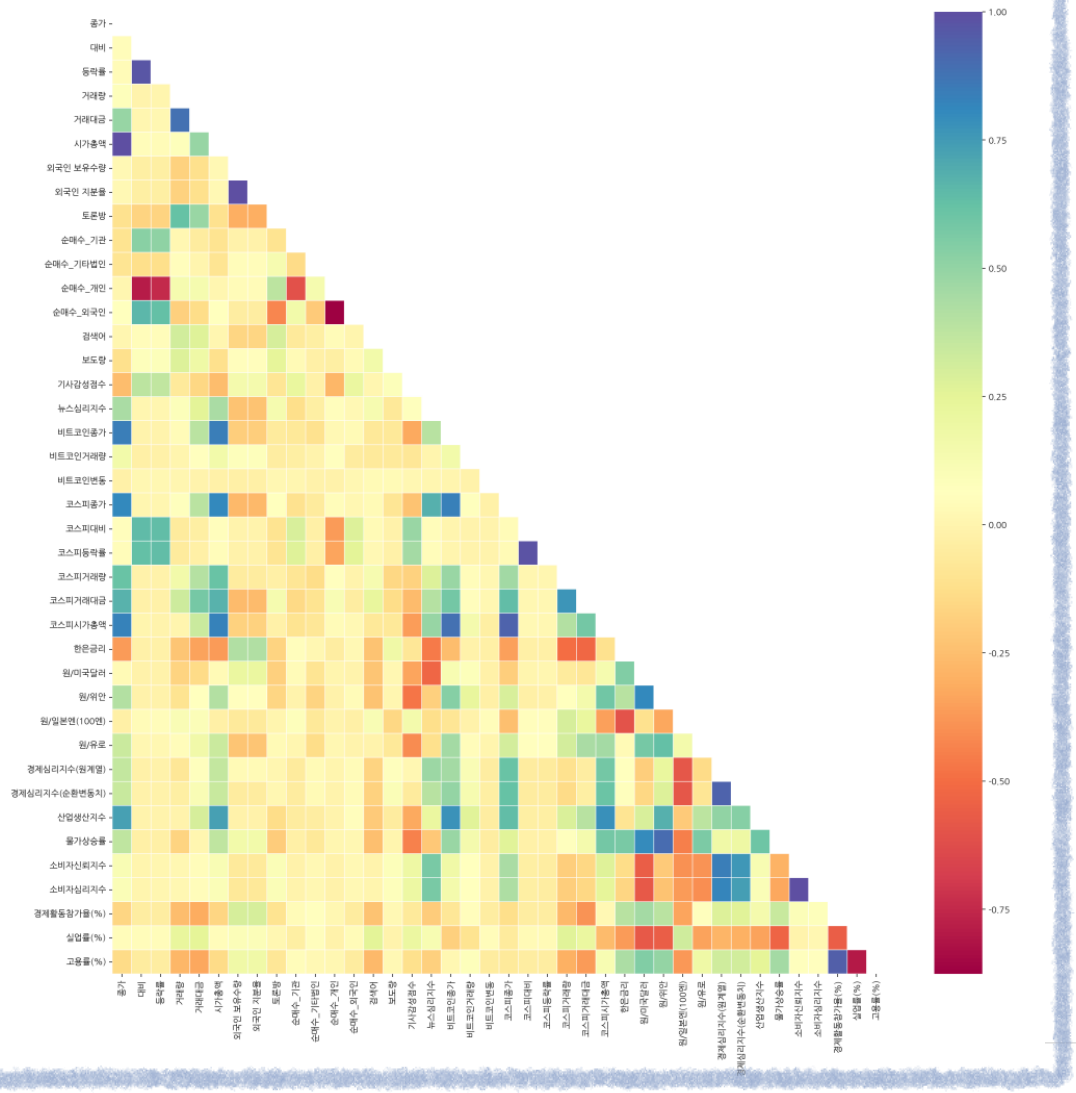
Employment rate(%) ,shows seasonality following periodic patterns

The definitions of trend and seasonality can be found in the Time Series Team Clean-up Week 1 reference!

4 EDA : SKHynix

Correlation between X variables

Correlation : X



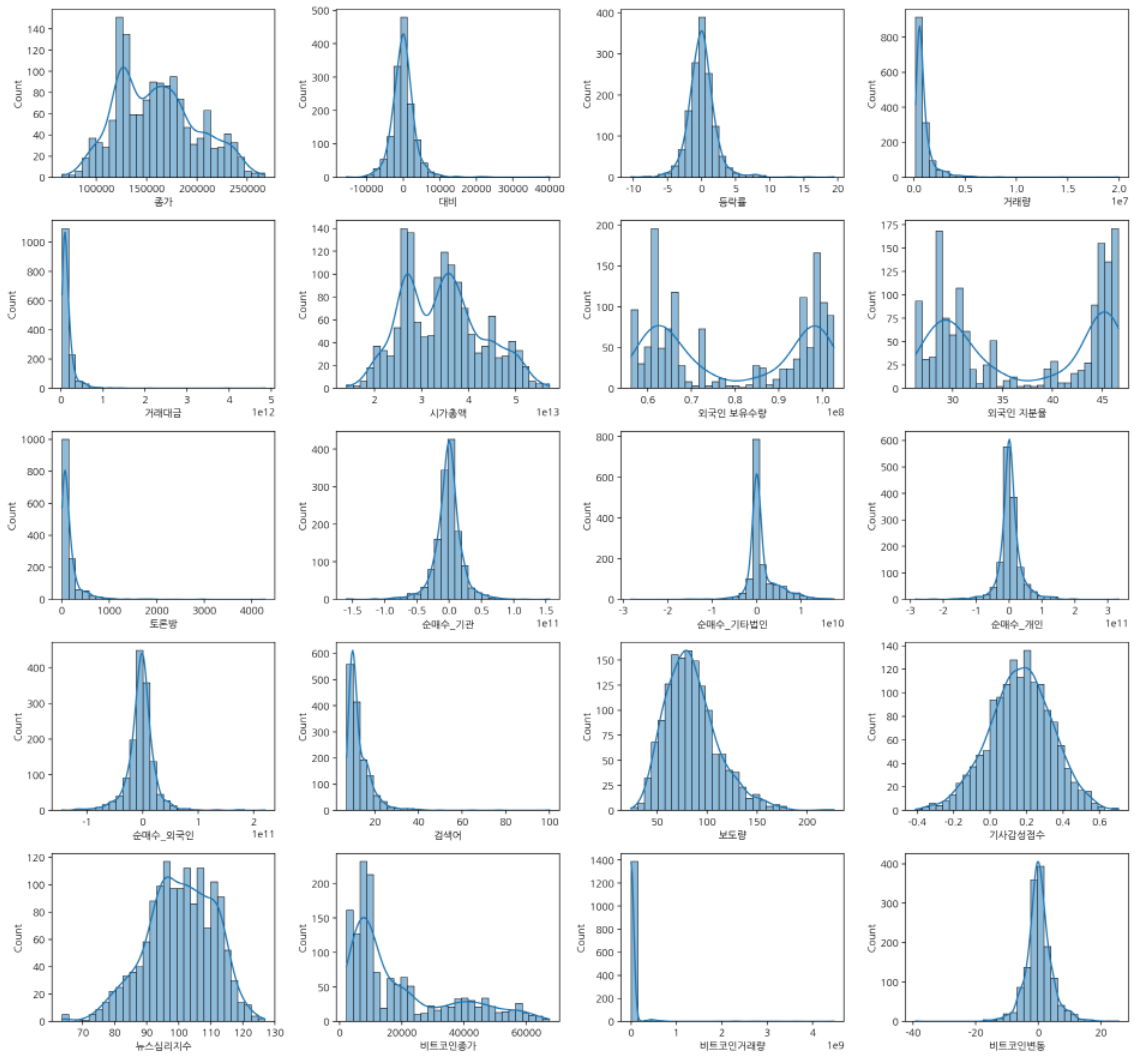
Variables with Correlation > 0.9

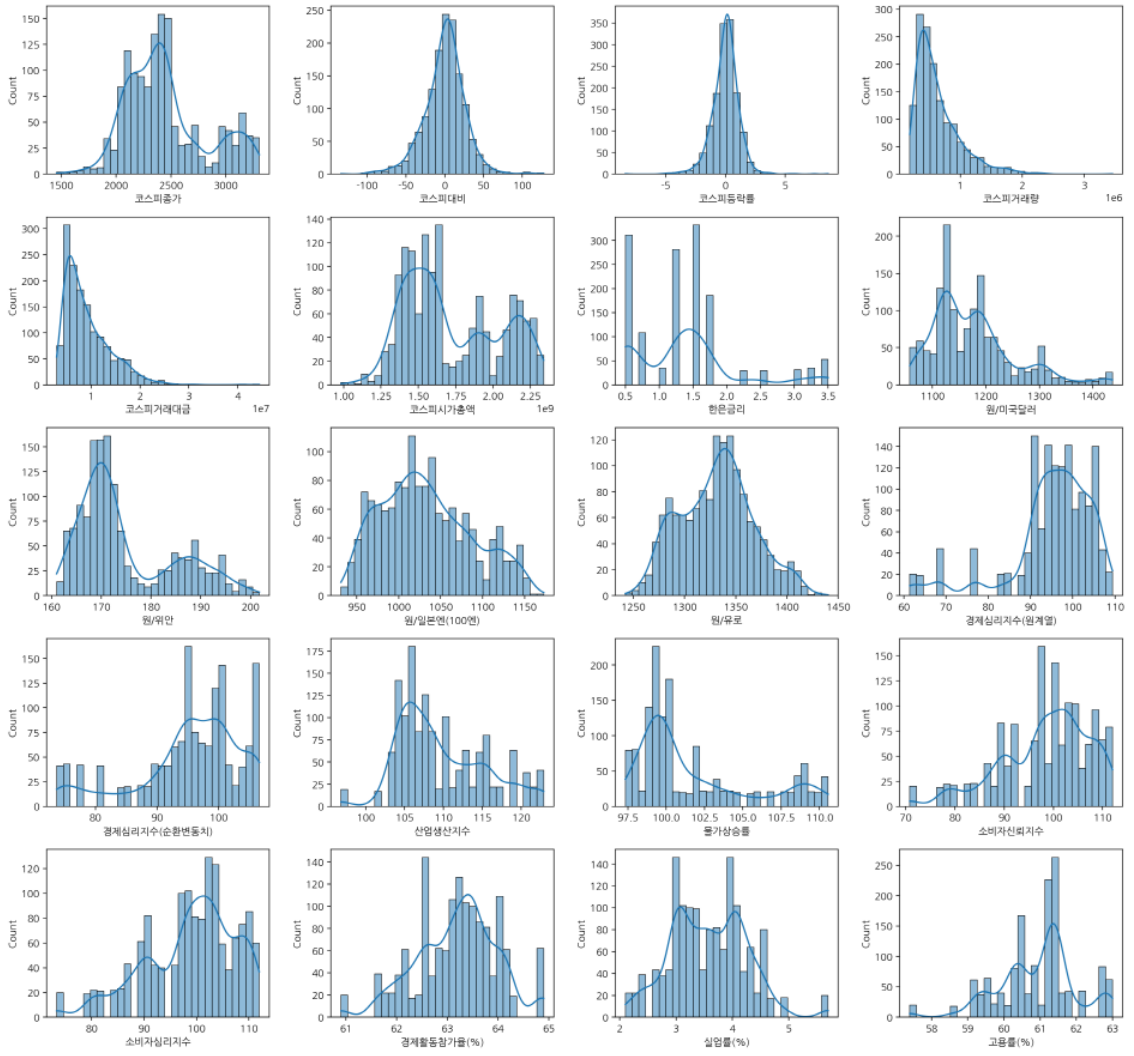
varable1	variable2	correlation
Closing price	Market cap	1.0000
foreign-held shares	Foreign ownership rate	0.9999
Consumer confidence index	Consumer sentiment index	0.9982
KOSPI change	KOSPI fluctuation rate	0.9792
change	fluctuation rate	0.9752
LFPR(%)	Employment rate(%)	0.9418
KOSPIclosing price	KOSPImarket cap	0.9313
Economic Sentiment Index (Raw Series)	Economic Sentiment Index (Seasonally Adjusted Series)	0.9290
KRW/CNY	Inflation rate	0.9014

Identified Variables with high similarity in the TS plot, actually have a high correlation each other

4 EDA : Hyundai

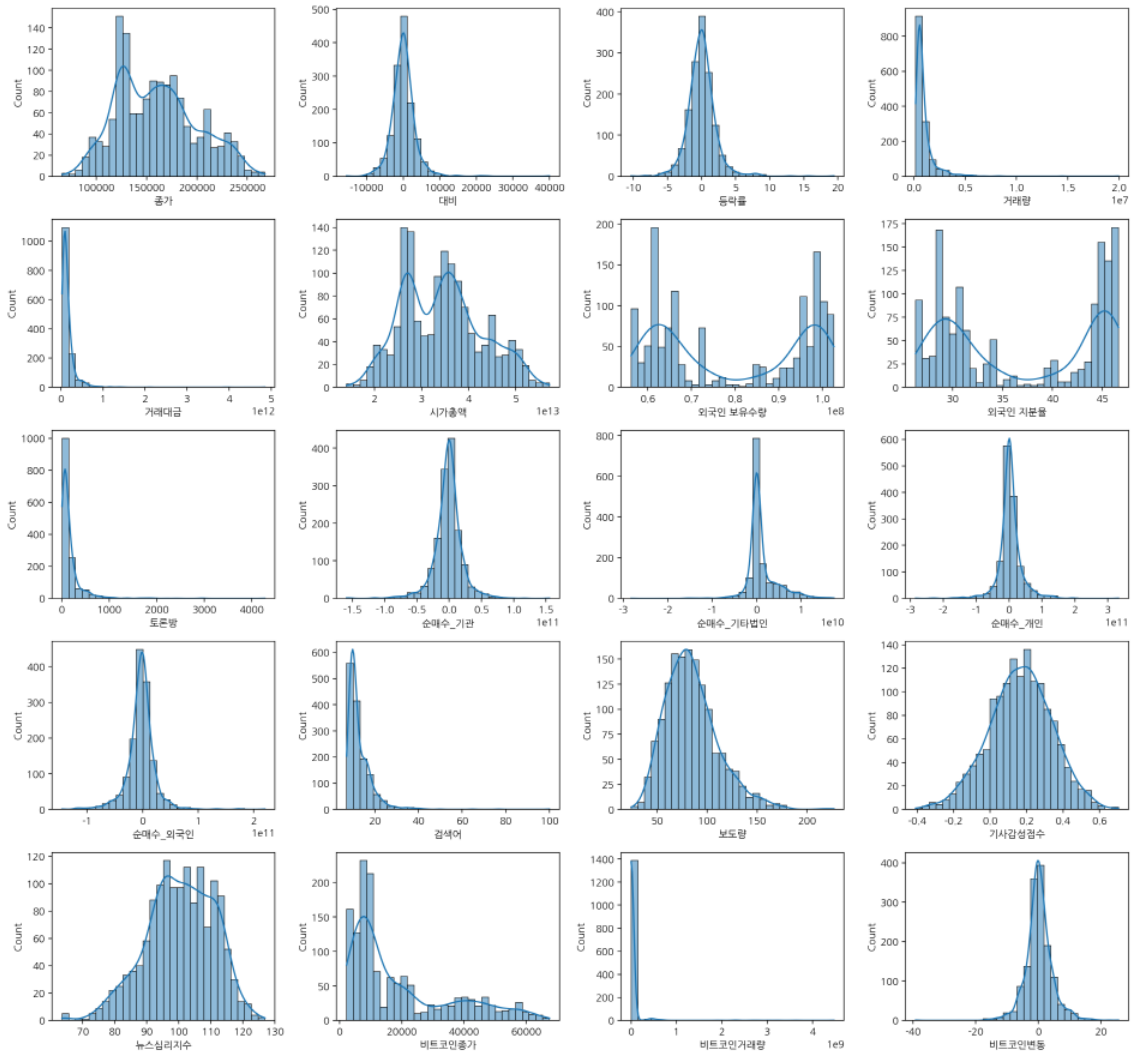
Distribution of X variables





4 EDA : Hyundai

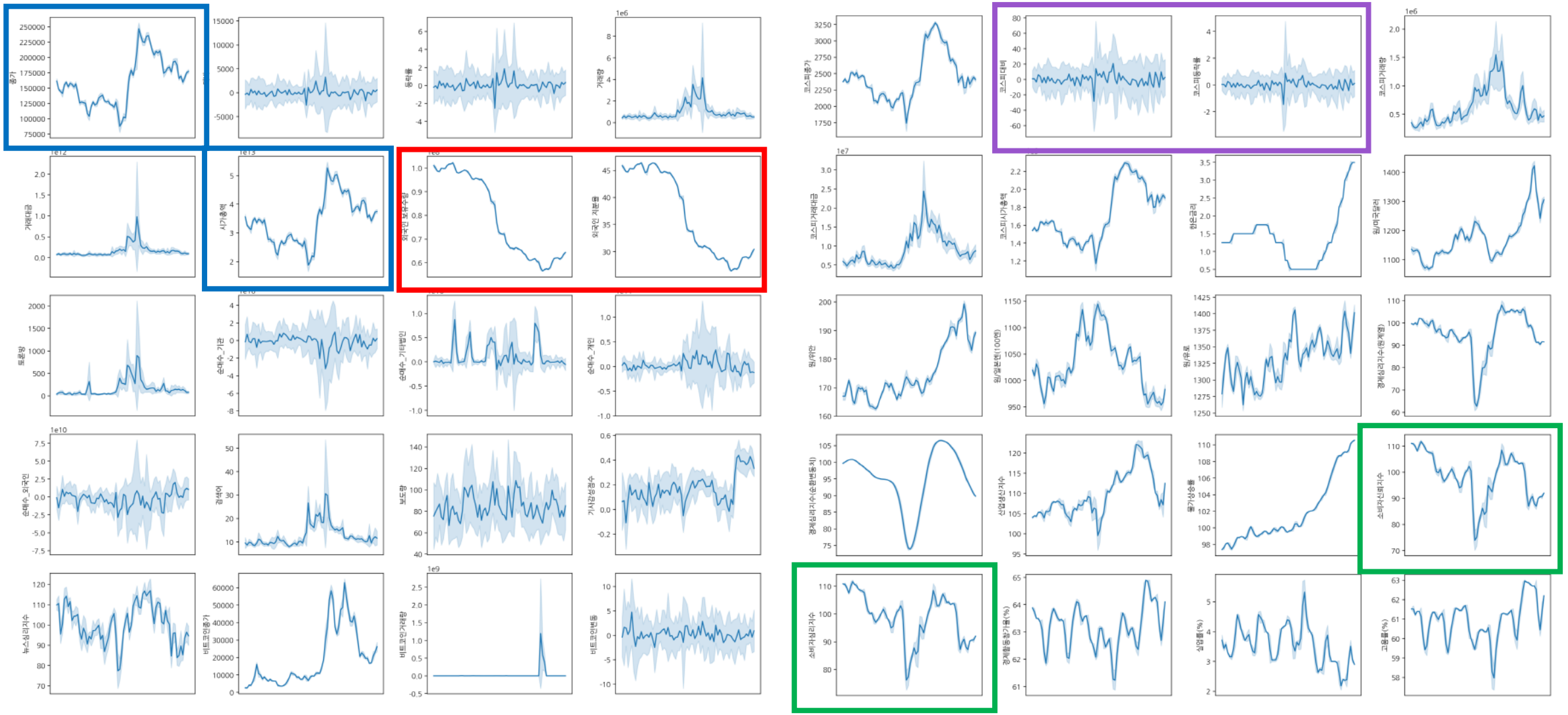
Distribution of X variables



There are no visible outliers observed

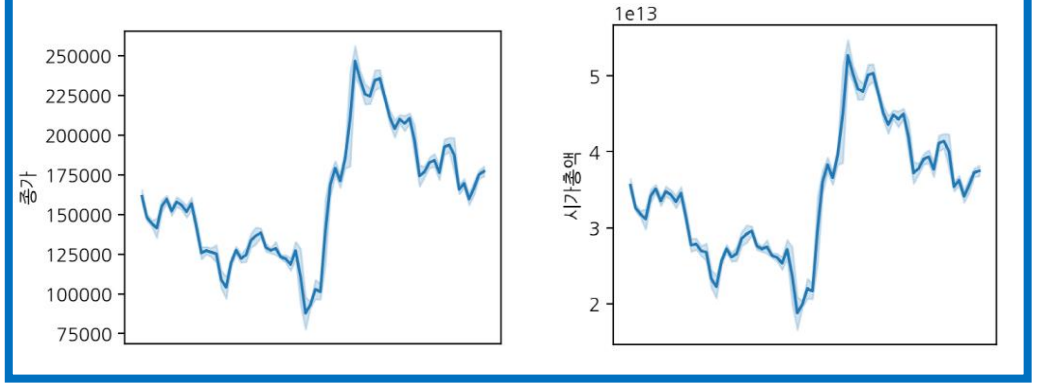
4 EDA : Hyundai

TS plot

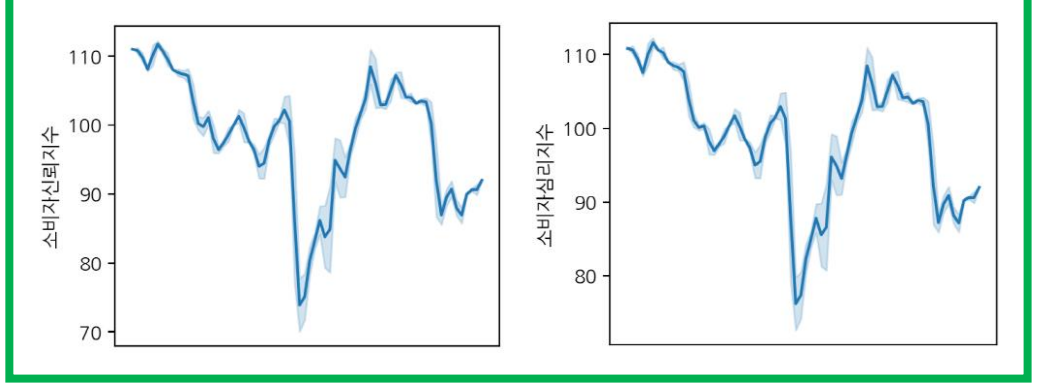


4 EDA : Hyundai

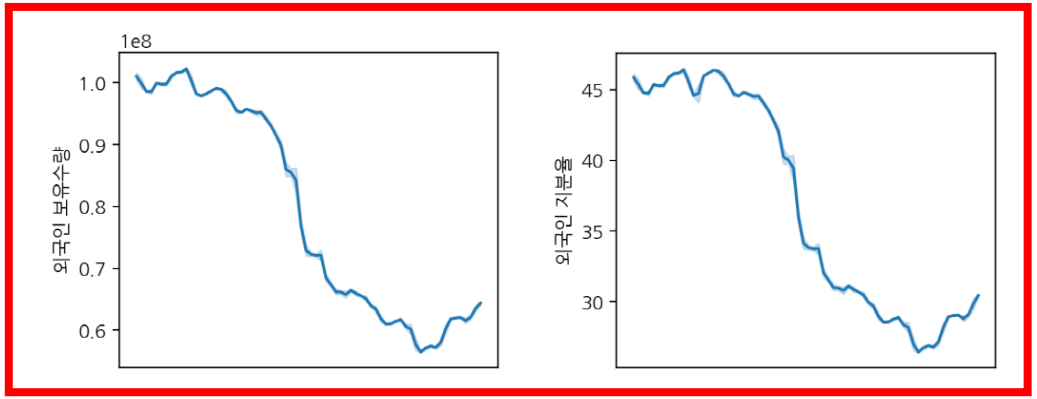
TS plot : plot with high similarity



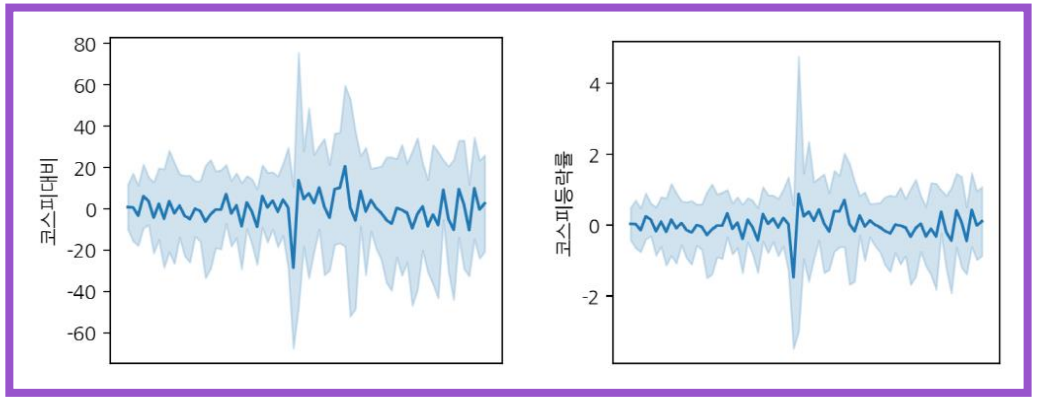
Closing price & market capitalization



Consumer confidence index & consumer sentiment index



Foreign investor holding quantity & foreign investor ownership percentage

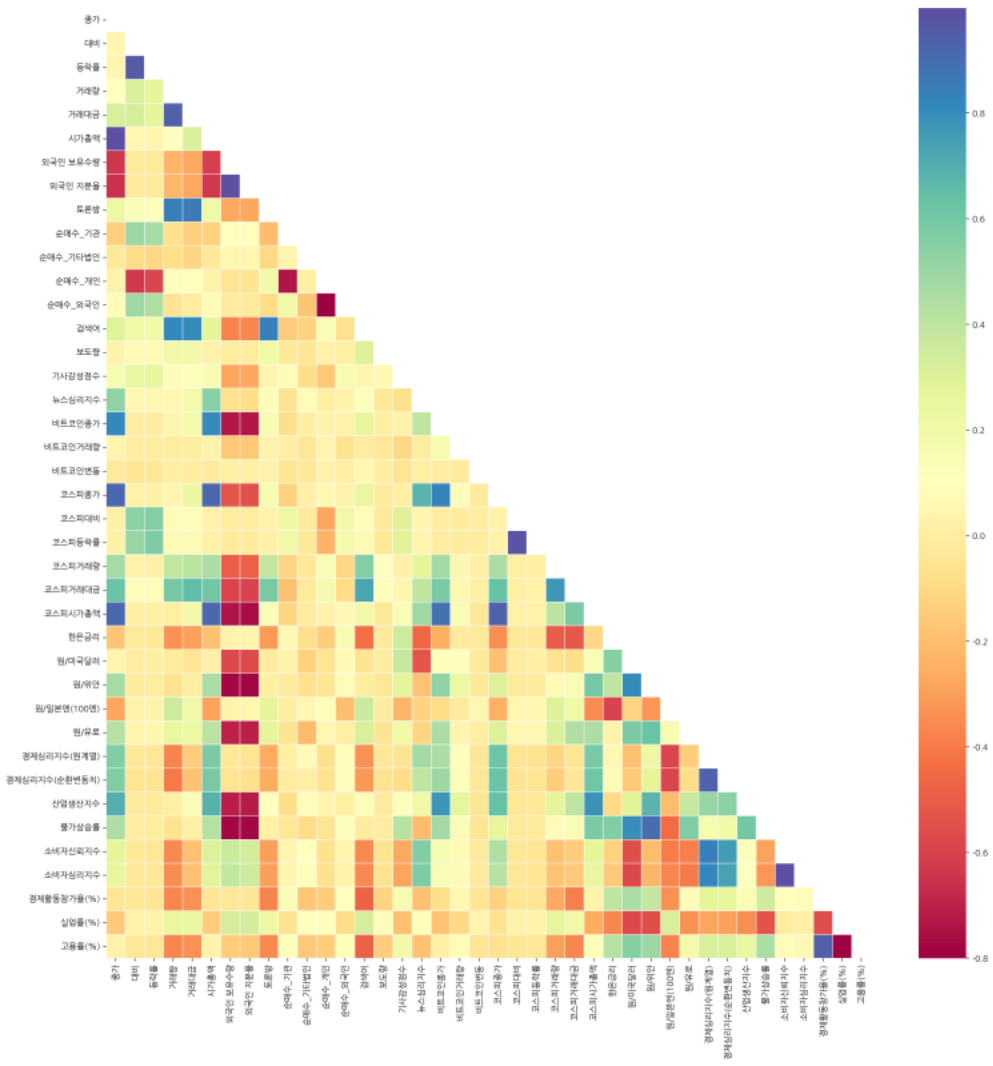


KOSPI change & KOSPI fluctuation rate

4 EDA : Hyundai

Correlation between X variables

Correlation : X



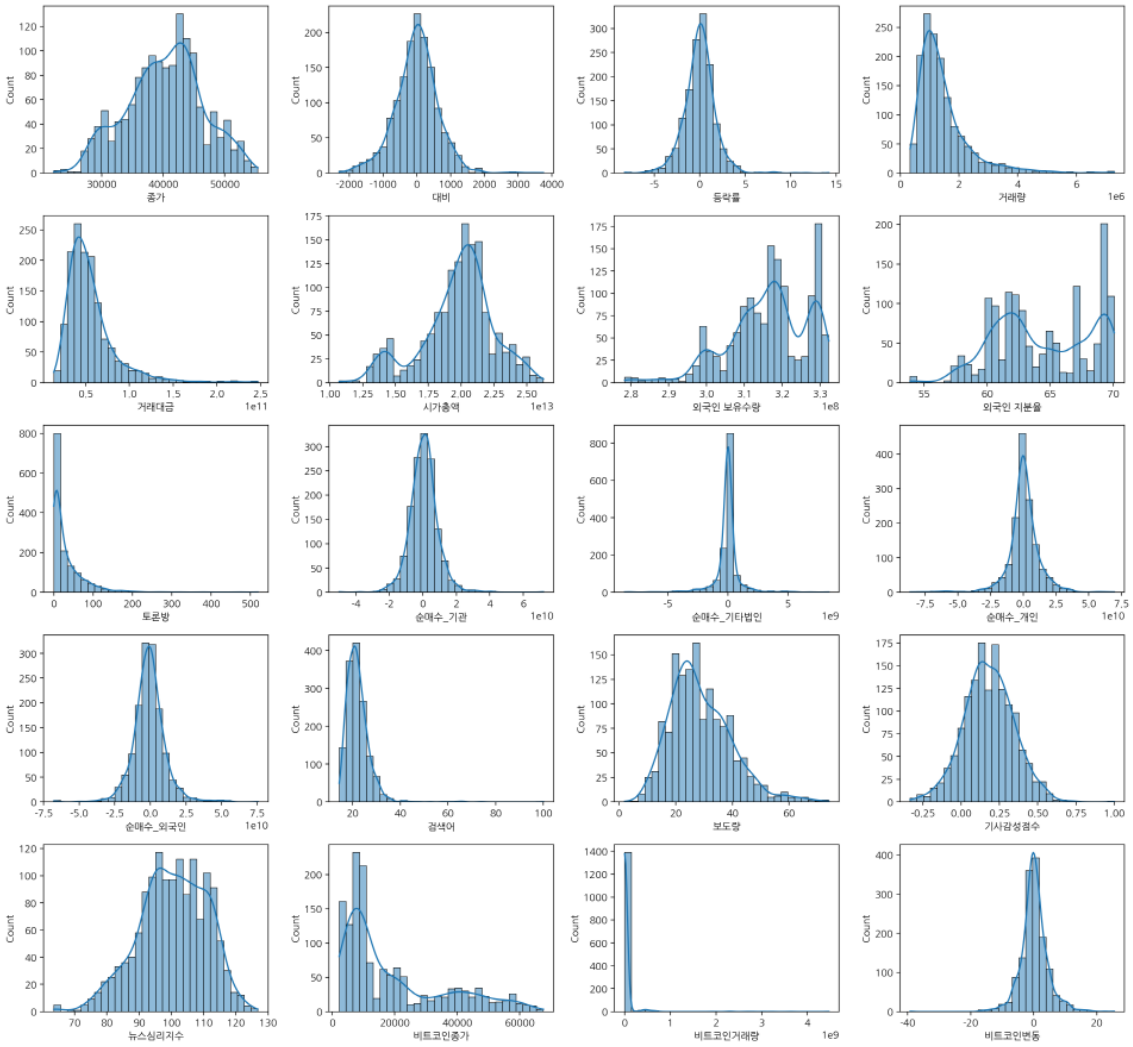
Variables with Correlation > 0.9

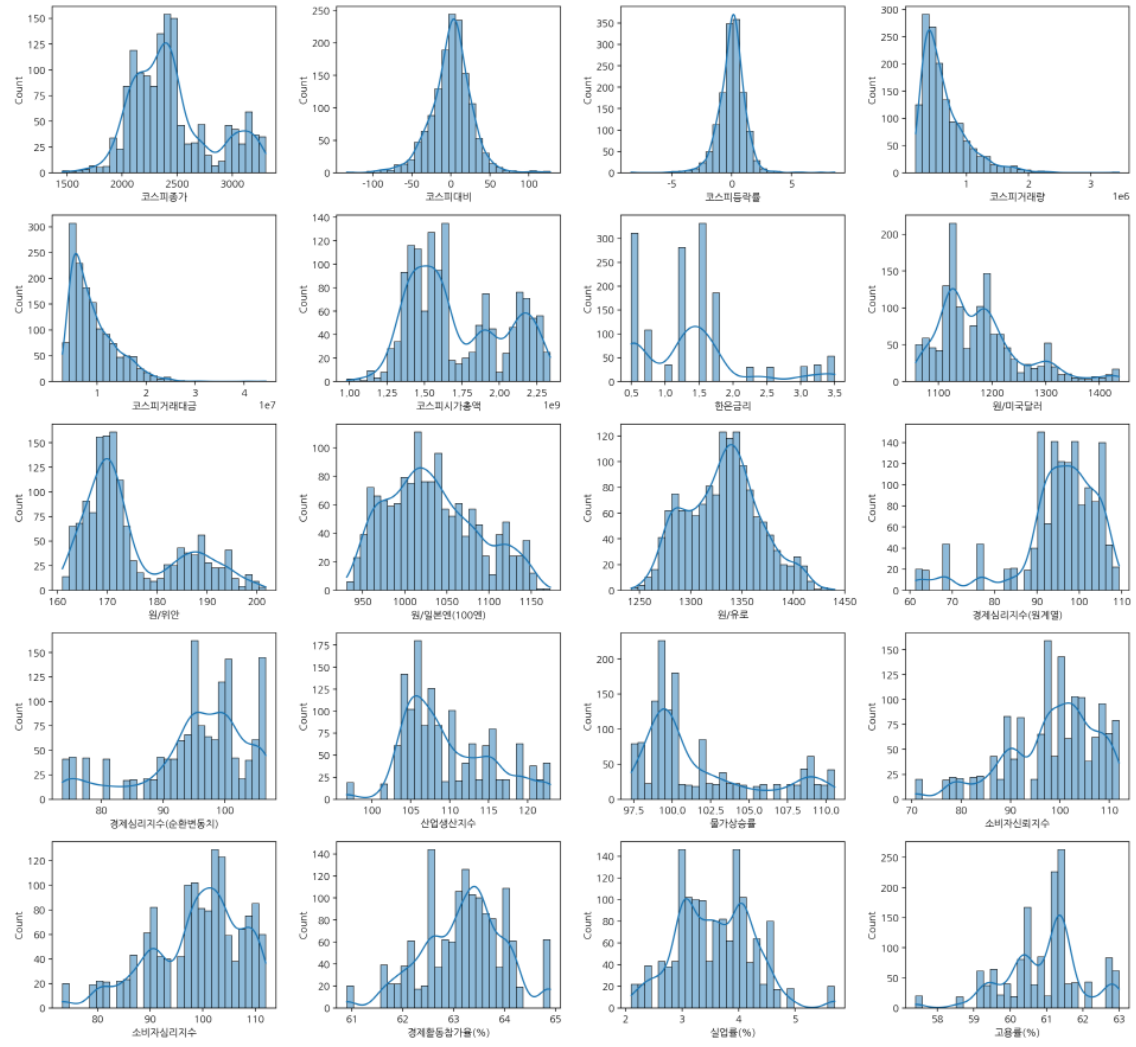
Variable1	variable2	correlation
Closing price	Market cap	0.9988
foreign-held shares	Foreign ownership rate	0.9982
Consumer confidence index	Consumer sentiment index	0.9982
KOSPI change	KOSPI fluctuation rate	0.9792
change	fluctuation rate	0.9541
LFPR(%)	Employment rate(%)	0.9418
Trading volume	Trading amount	0.9363
KOSPI closing price	KOSPI market cap	0.9313
Economic Sentiment Index (Raw Series)	Economic Sentiment Index (Seasonally Adjusted Series)	0.9290
Closing price	KOSPI market cap	0.9211
Market cap	KOSPI closing price	0.9209
Closing price	KOSPI closing price	0.9174
Market cap	KOSPI market cap	0.9162
KRW/CNY	Inflation rate	0.9014

Identified Variables with high similarity in the TS plot,
actually have a high correlation each other

4 EDA : Shinhan

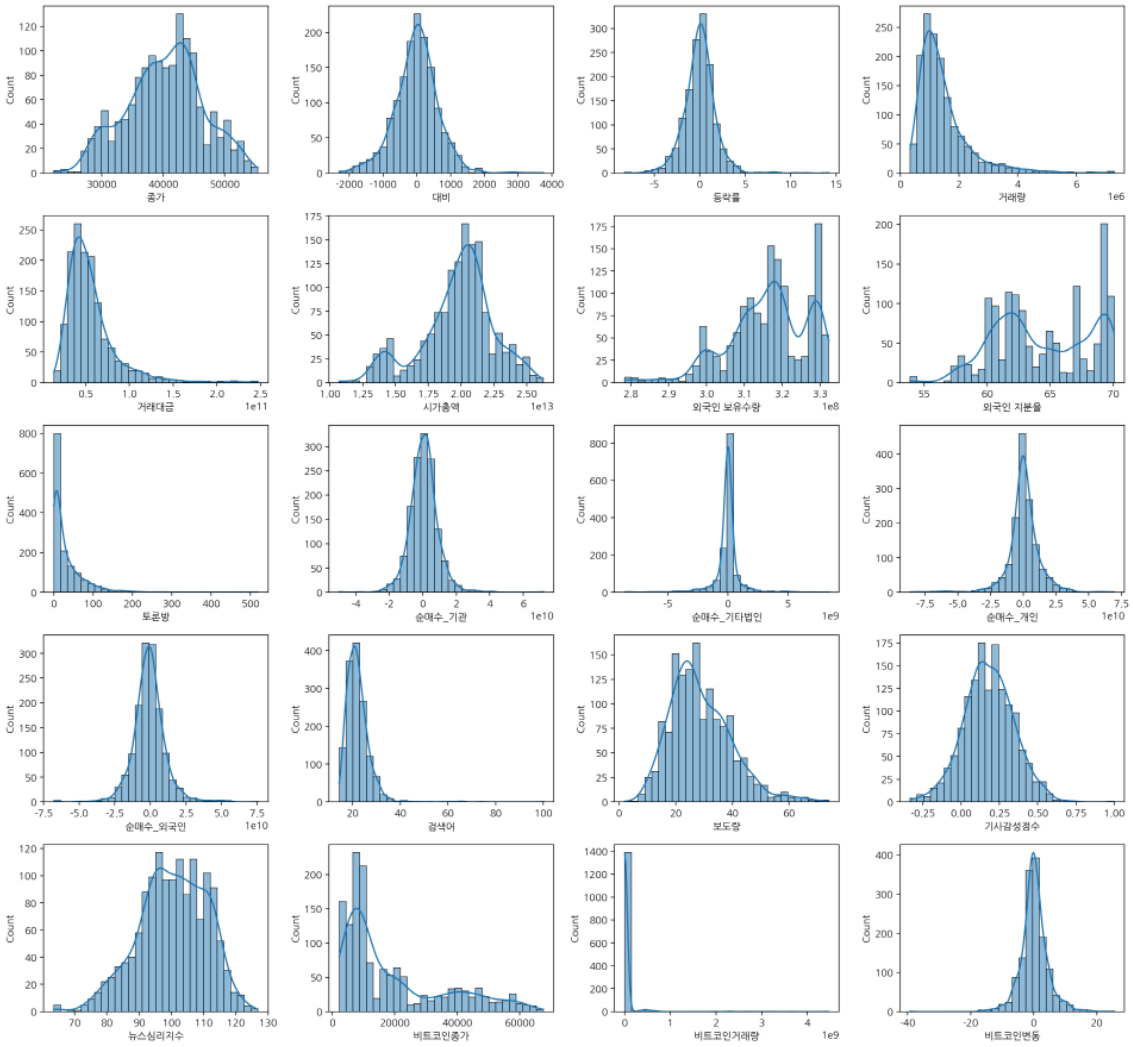
Distribution of X variables





4 EDA : Shinhan

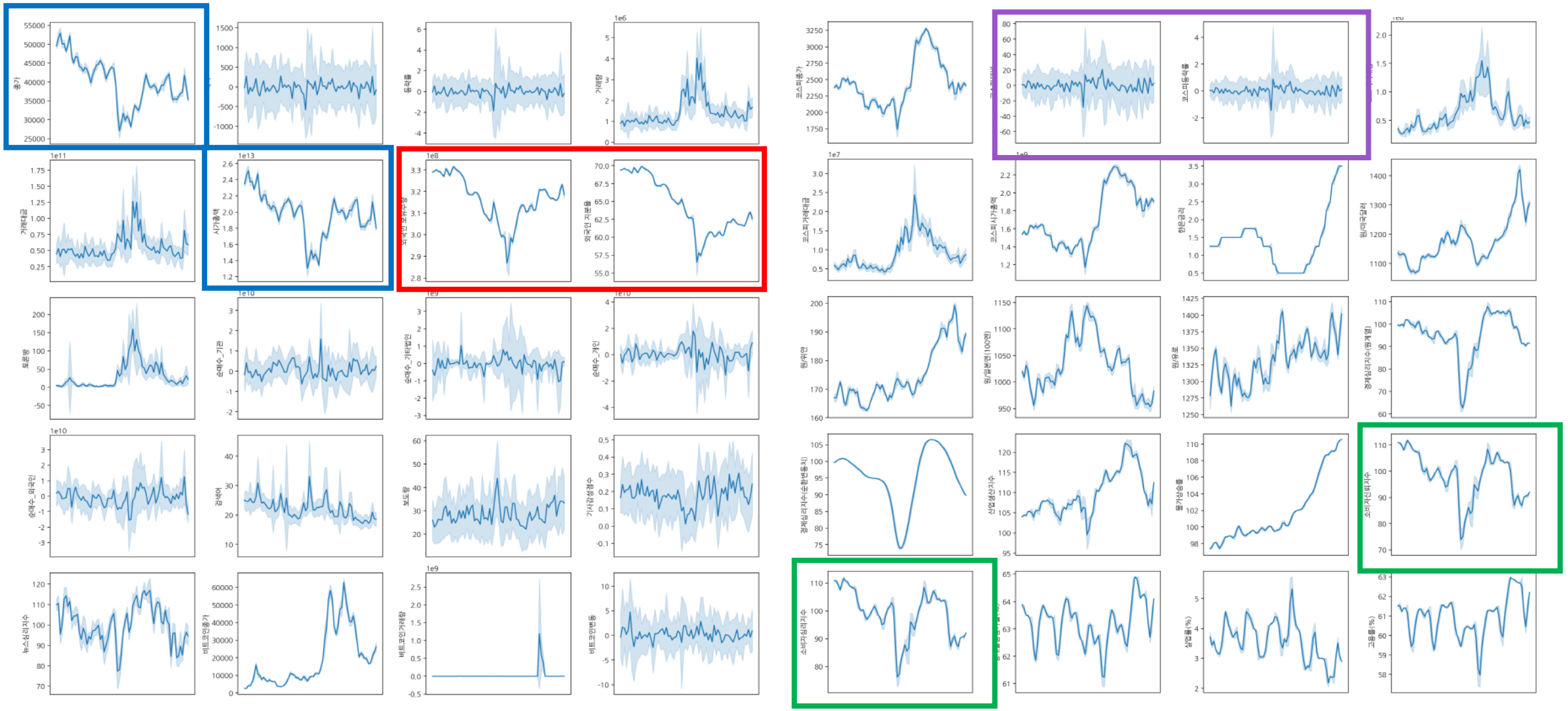
Distribution of X variables



There’s no visible outliers observed

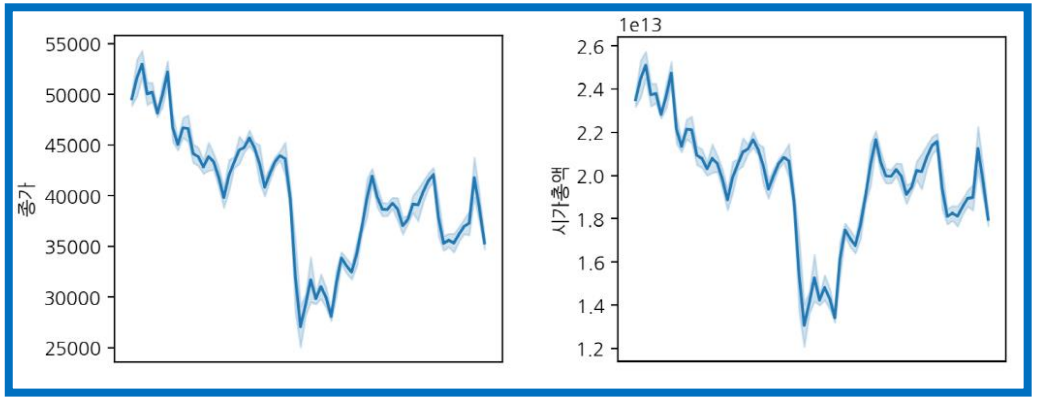
4 EDA : Shinhan

TS plot

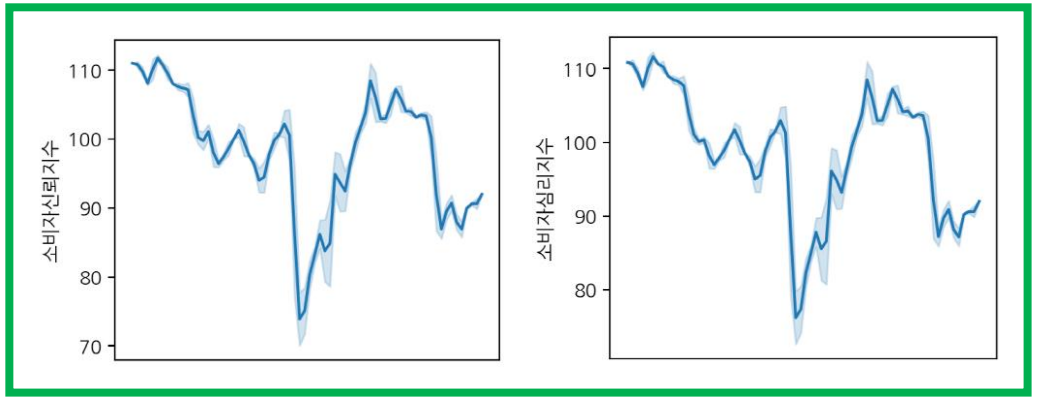


4 EDA : Shinhan

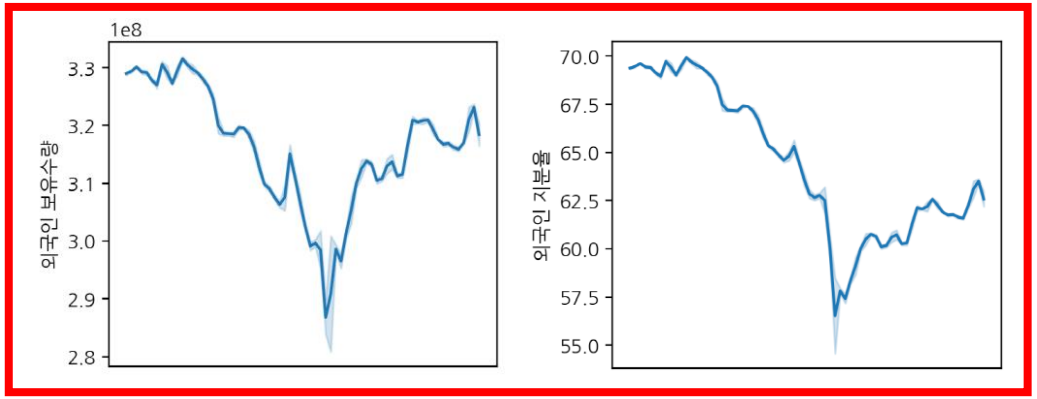
TS plot : plot with high similarity



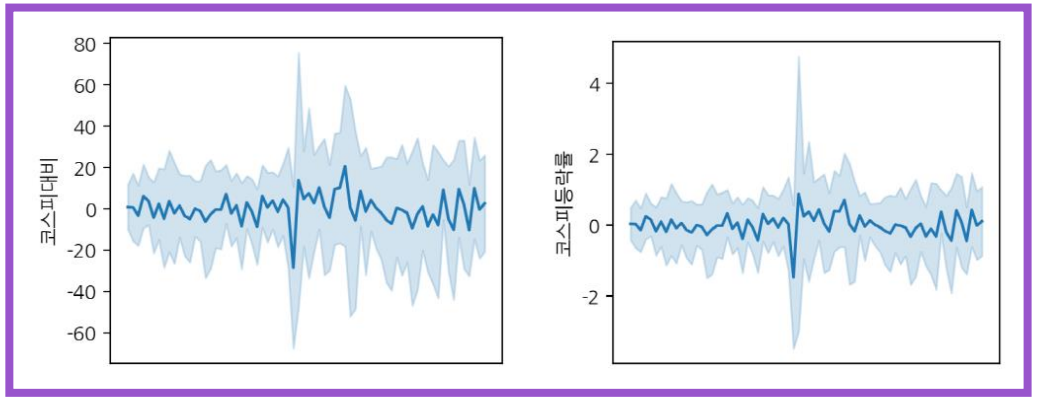
Closing price & market capitalization



Consumer confidence index & consumer sentiment index



Foreign investor holding quantity & foreign investor ownership percentage

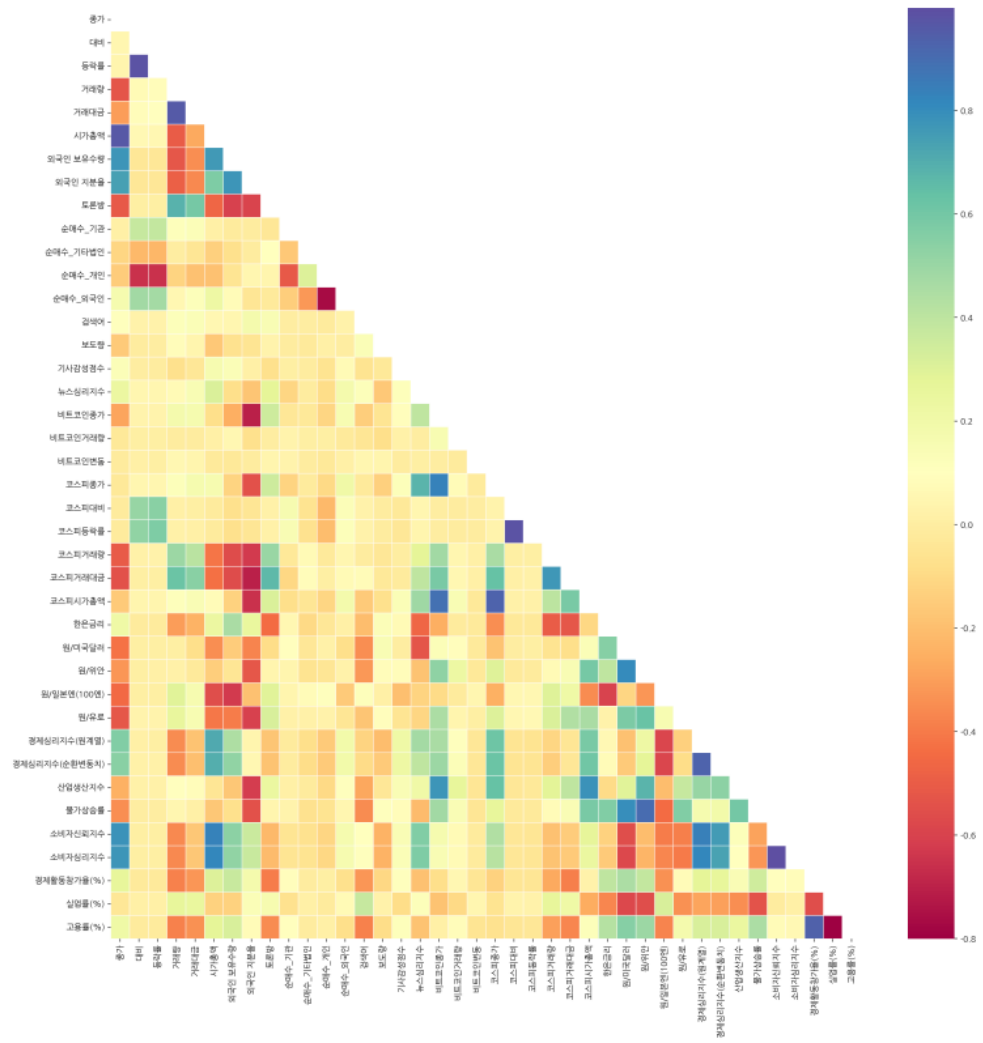


KOSPI change & KOSPI fluctuation rate

4 EDA : Shinhan

Correlation between X variables

Correlation : X



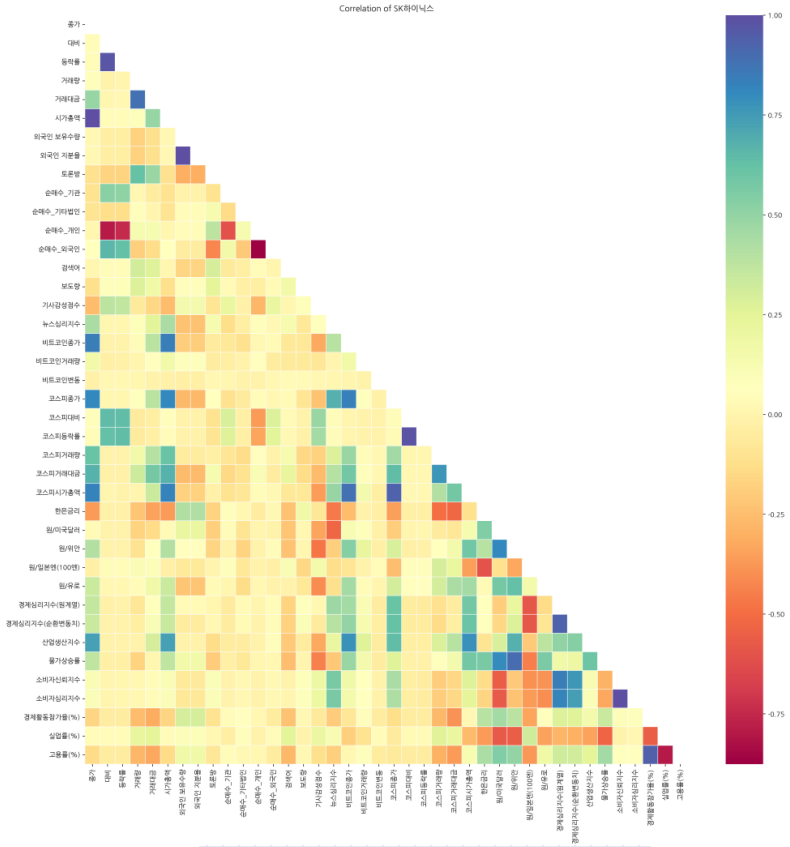
Variables with Correlation > 0.9

variable1	variable2	correlation
Consumer confidence index	Consumer sentiment index	**0.9982**
KOSPI change	KOSPI fluctuation rate	**0.9792**
Change	fluctuation rate	0.9780
Closing price	Market cap	0.9681
Trading volume	Trading amount	0.9582
LFPR(%)	Employment rate(%)	0.9418
KOSPI closing price	KOSPI market cap	0.9313
Economic Sentiment Index (Raw Series)	Economic Sentiment Index (Seasonally Adjusted Series)	0.9290
KRW/CNY	Inflation rate	0.9014

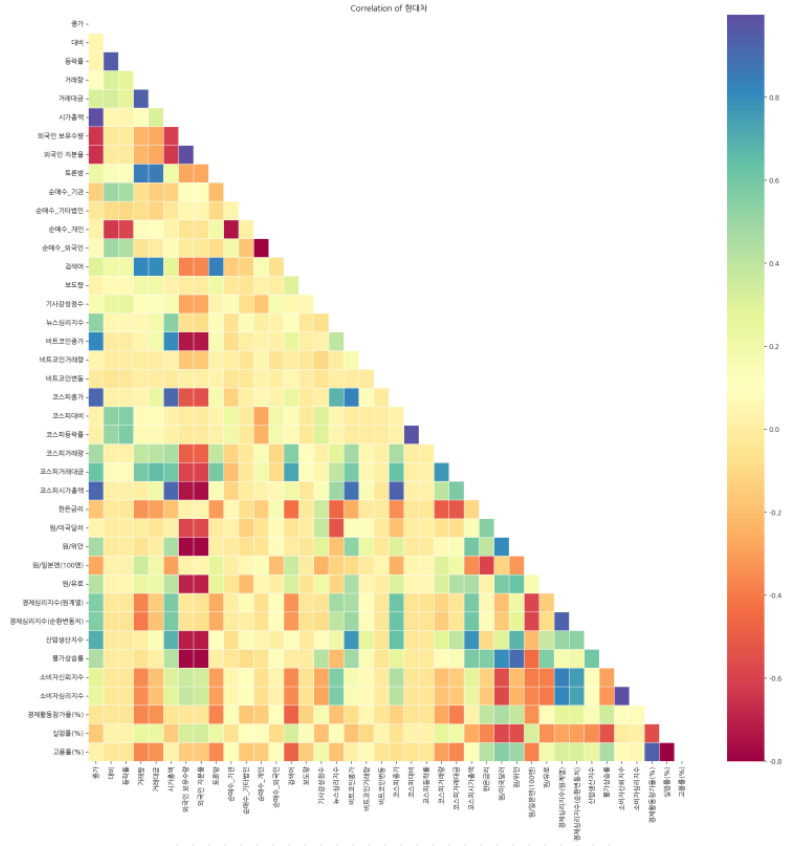
Identified Variables with high similarity in the TS plot, actually have a high correlation each other

4 EDA

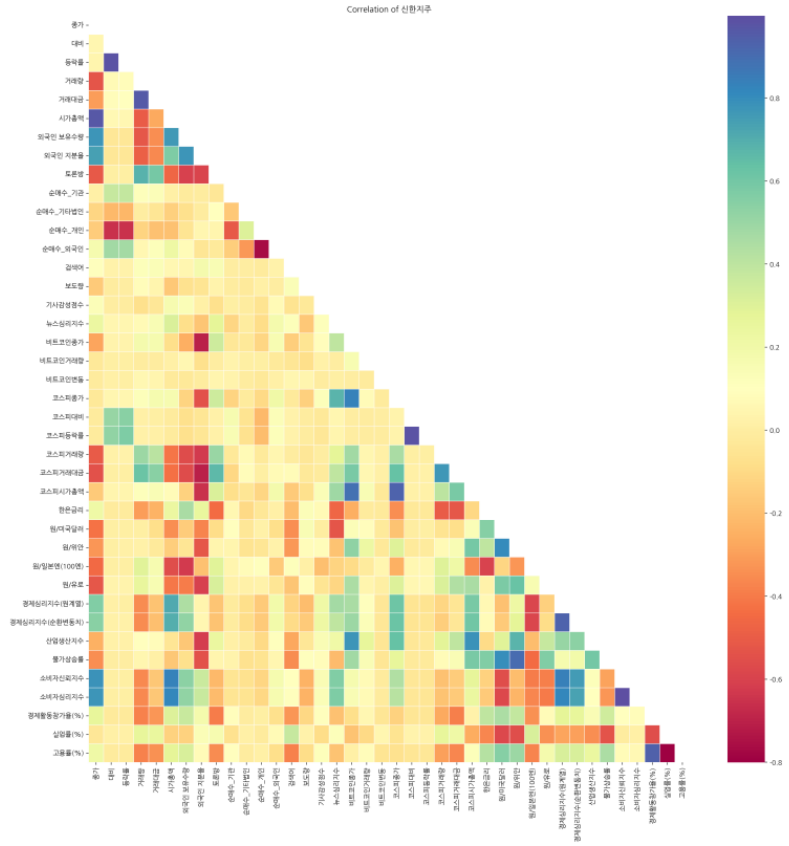
Correlation with Fluctuation rate



SKHynix



Hyundai



Shinhan

4 EDA

Correlation with Fluctuation rate

Let's focus on variables
that have a high correlation with the fluctuation rate!

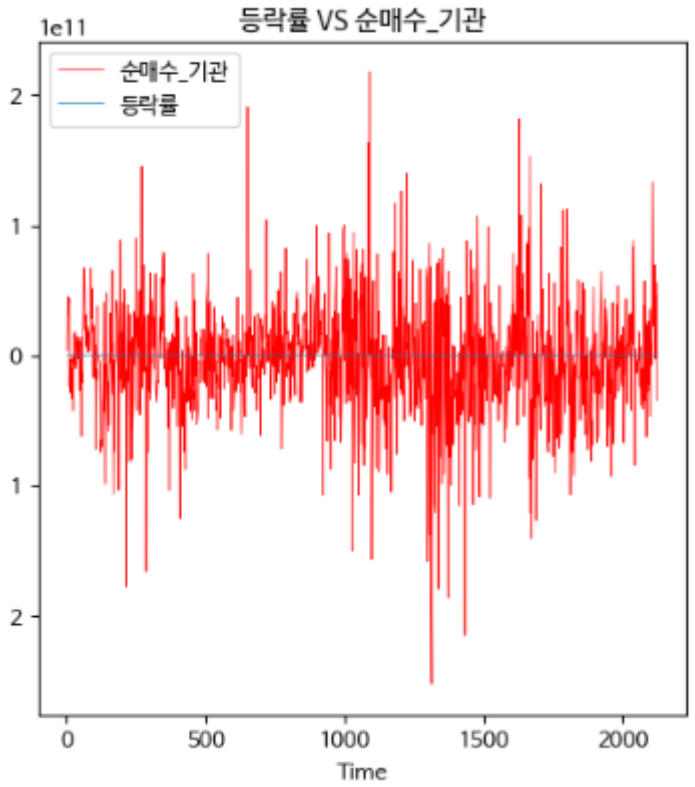
SKHynix

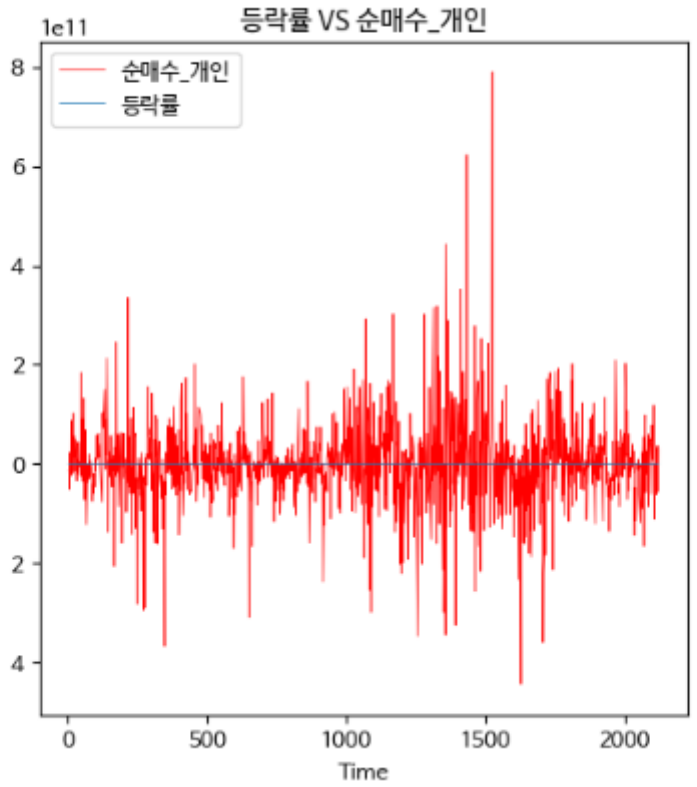
Hyundai

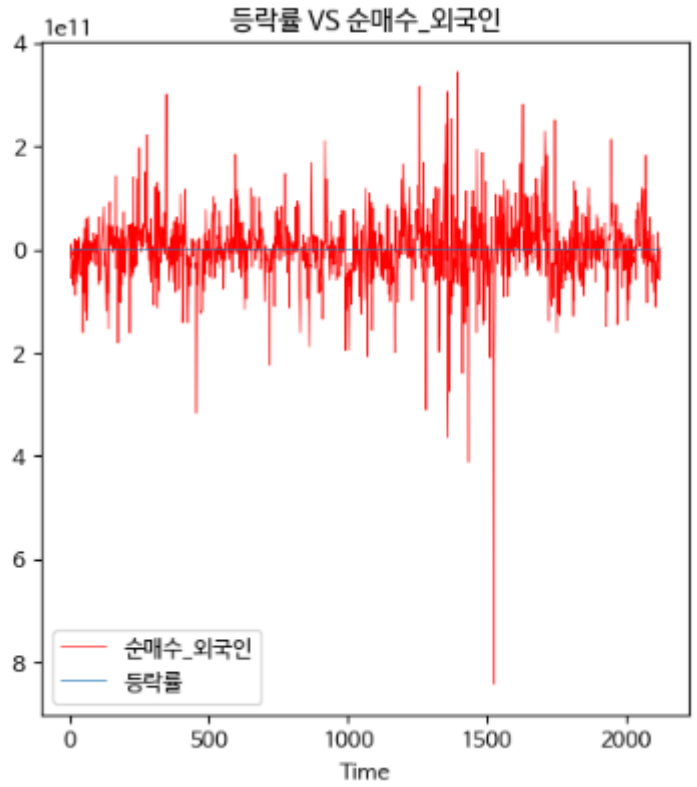
Shinhan

4 EDA

Net purchases by institutions, individuals, foreign



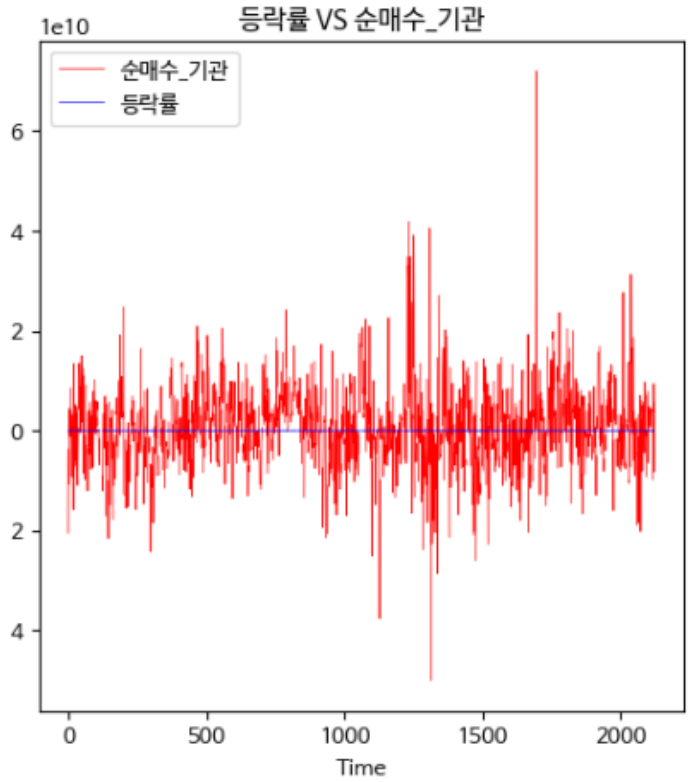


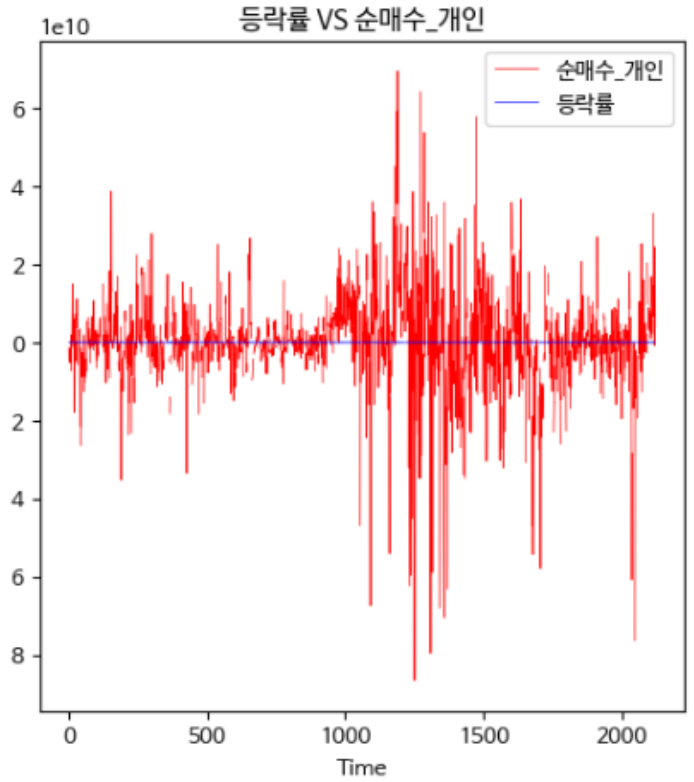


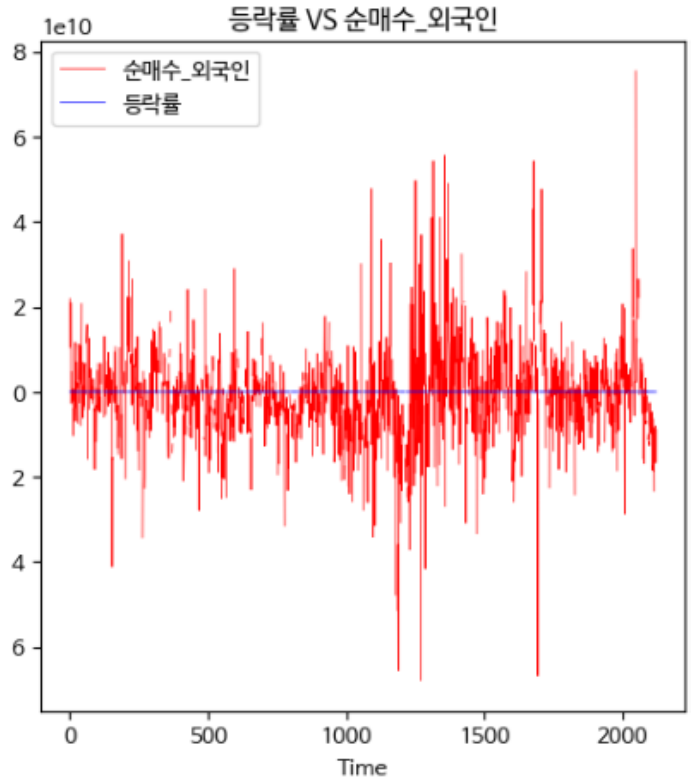
SKHynix

4 EDA

Net purchases by institutions, individuals, foreign



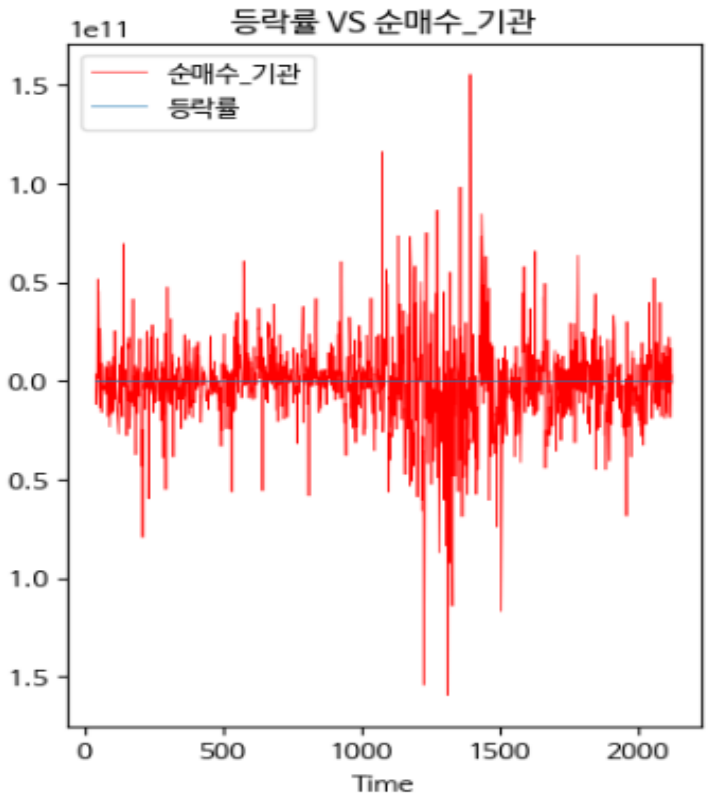


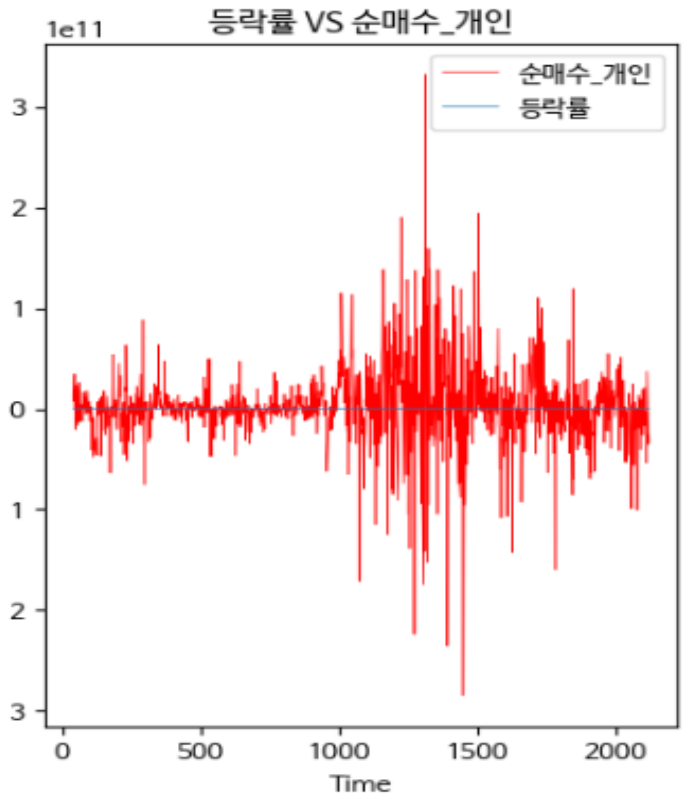


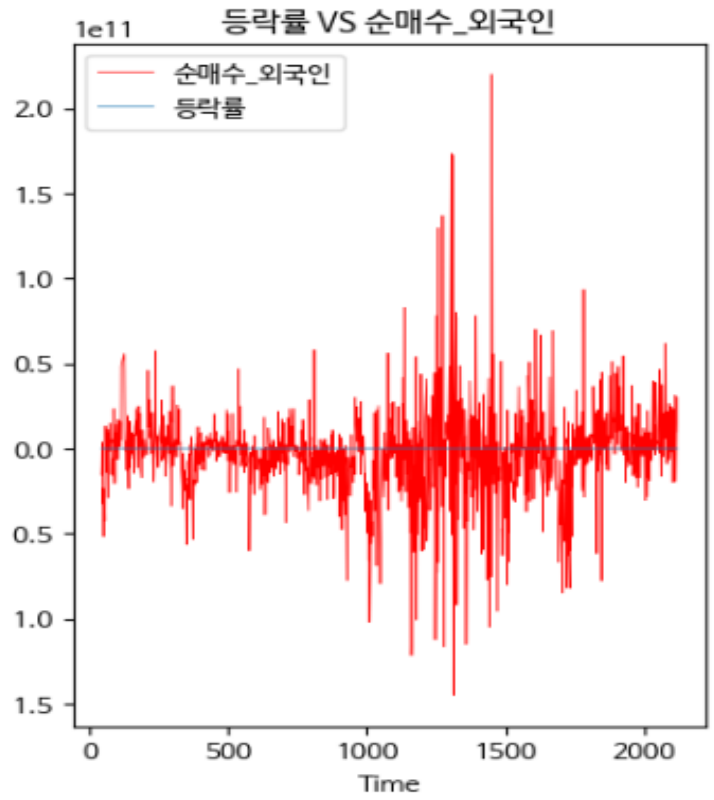
Hyundai

4 EDA

Net purchases by institutions, individuals, foreign



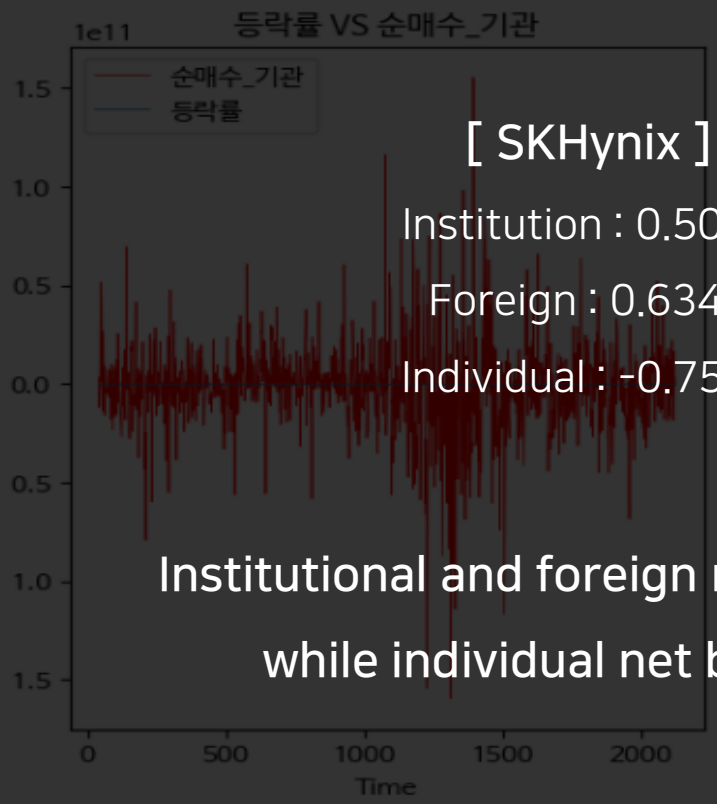




Shinhan

4 EDA

Net purchases by institutions, individuals, foreign



[SKHynix]

Institution : 0.5056

Foreign : 0.6349

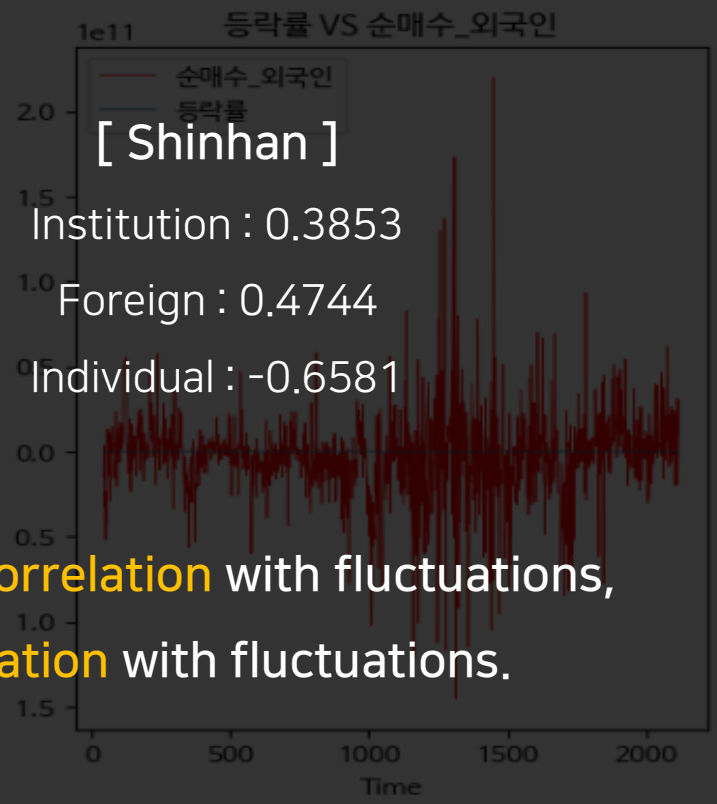
Individual : -0.7541

[Hyundai]

Institution : 0.4678

Foreign : 0.4444

Individual : -0.5798



[Shinhan]

Institution : 0.3853

Foreign : 0.4744

Individual : -0.6581

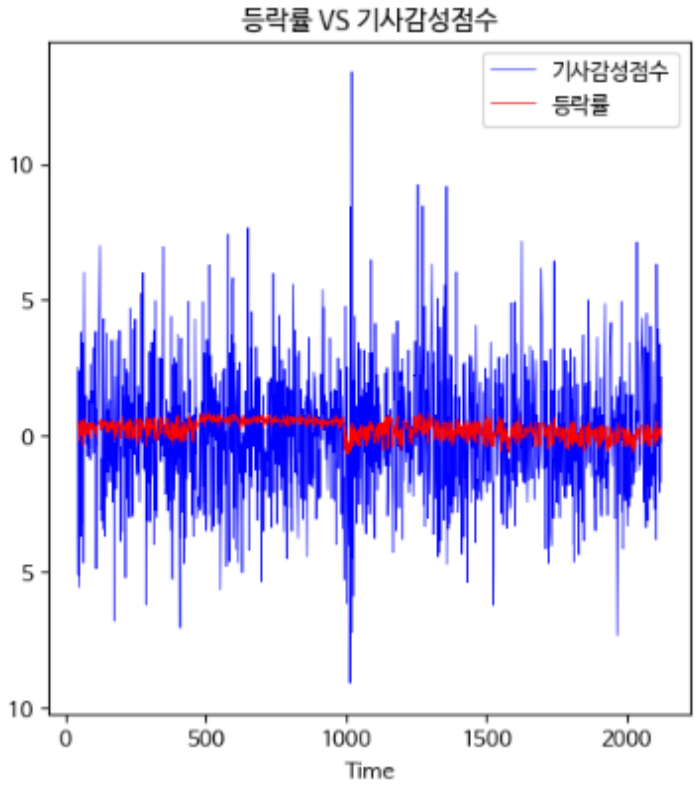
Institutional and foreign net buying volume have a positive correlation with fluctuations,

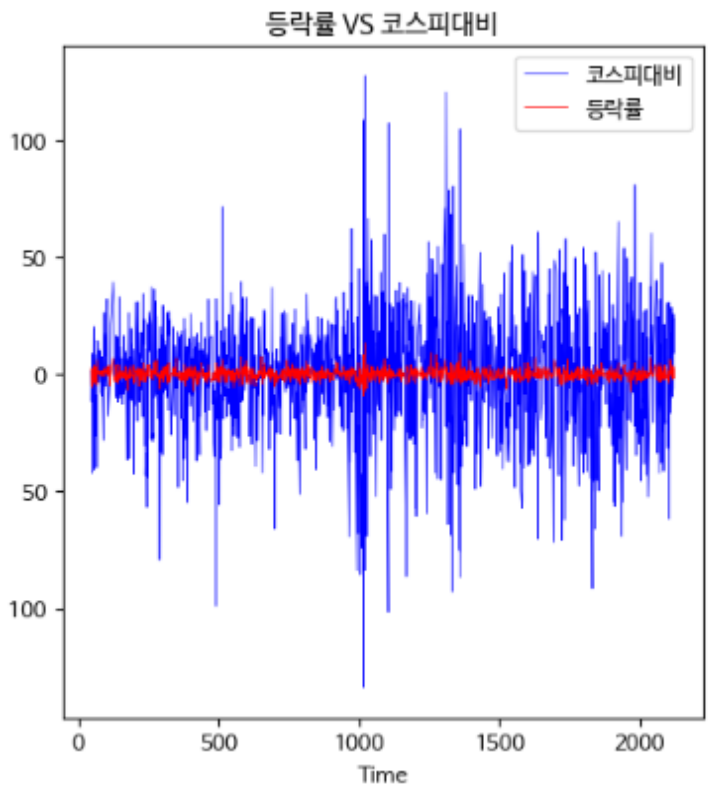
while individual net buying volume has a negative correlation with fluctuations.

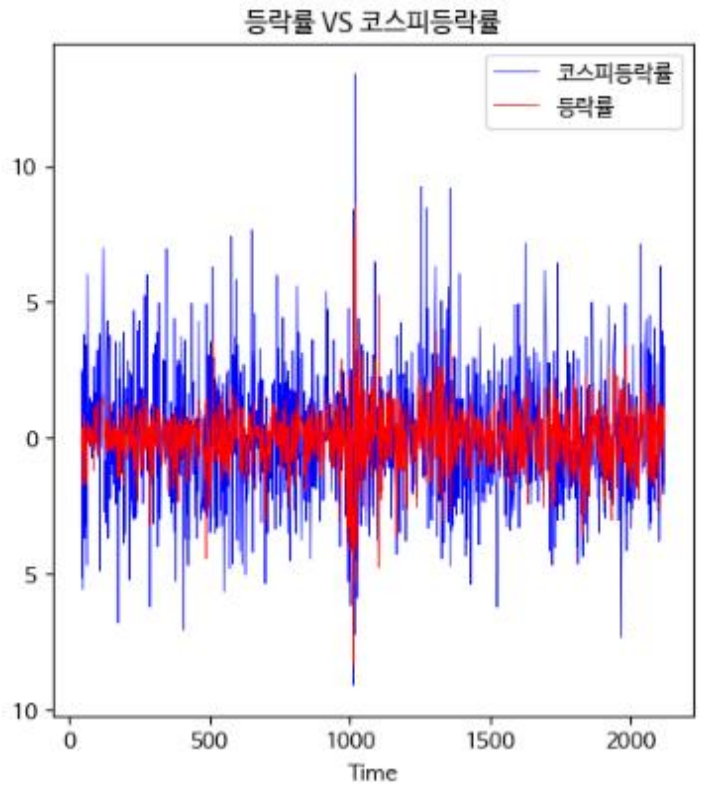
Shinhan

4 EDA

Sentiment score, KOSPI change, KOSPI fluctuation rate



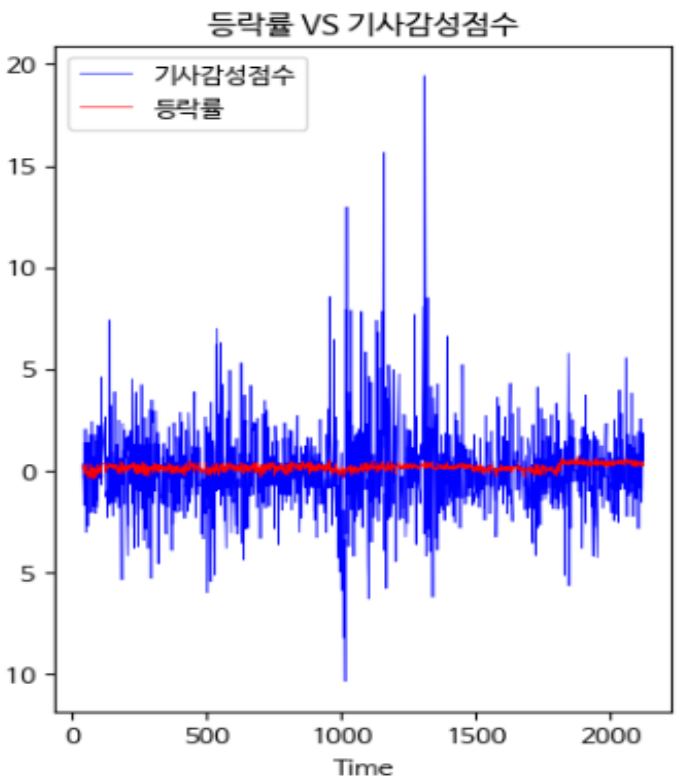


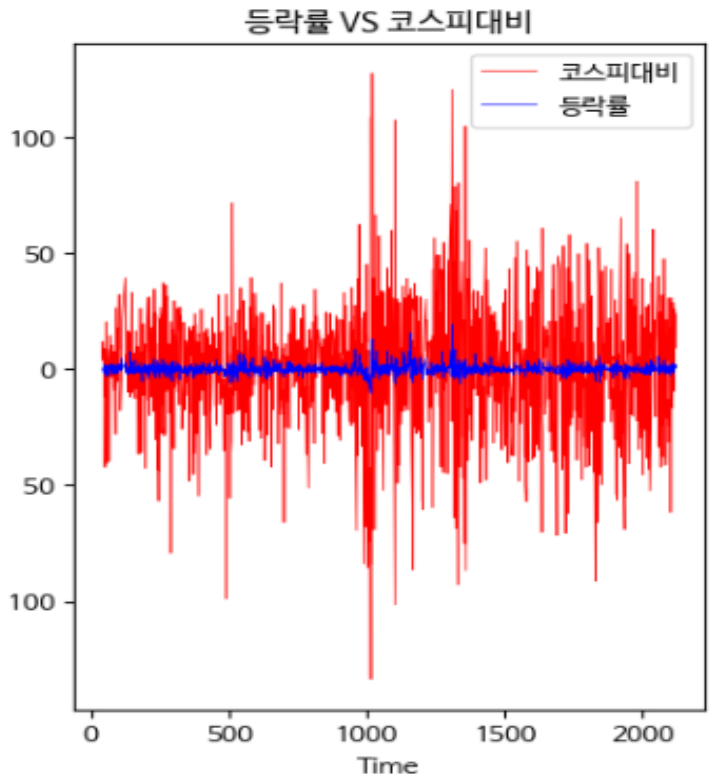


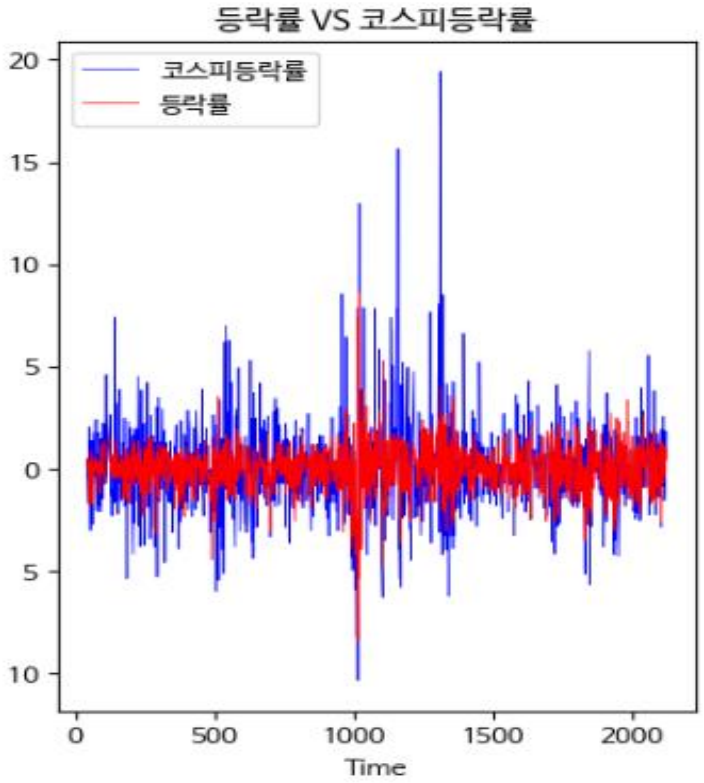
SKHynix

4 EDA

Sentiment score, KOSPI change, KOSPI fluctuation rate



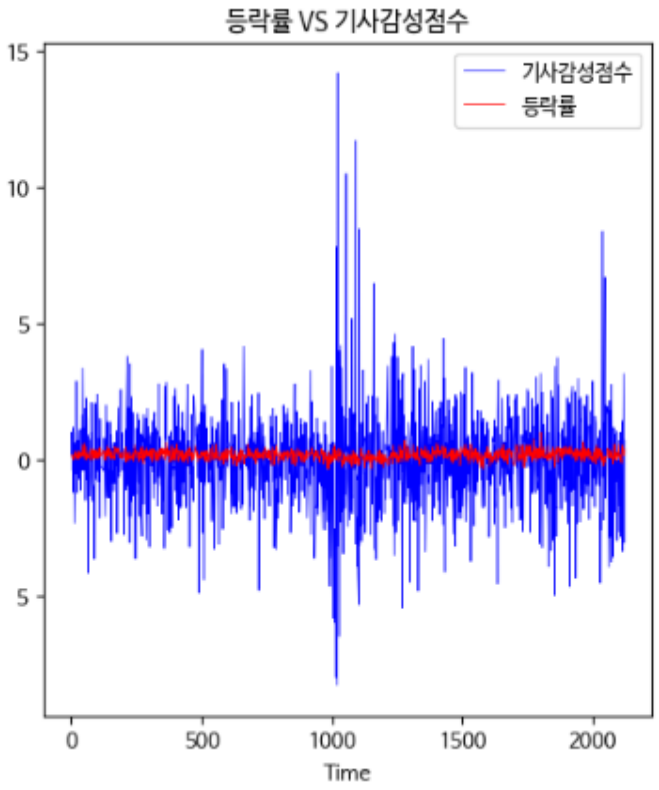


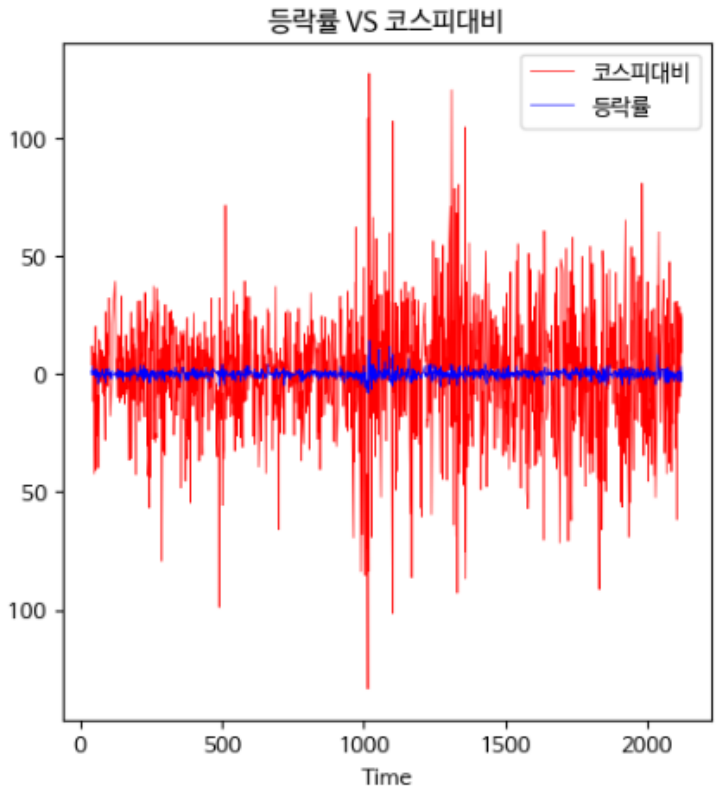


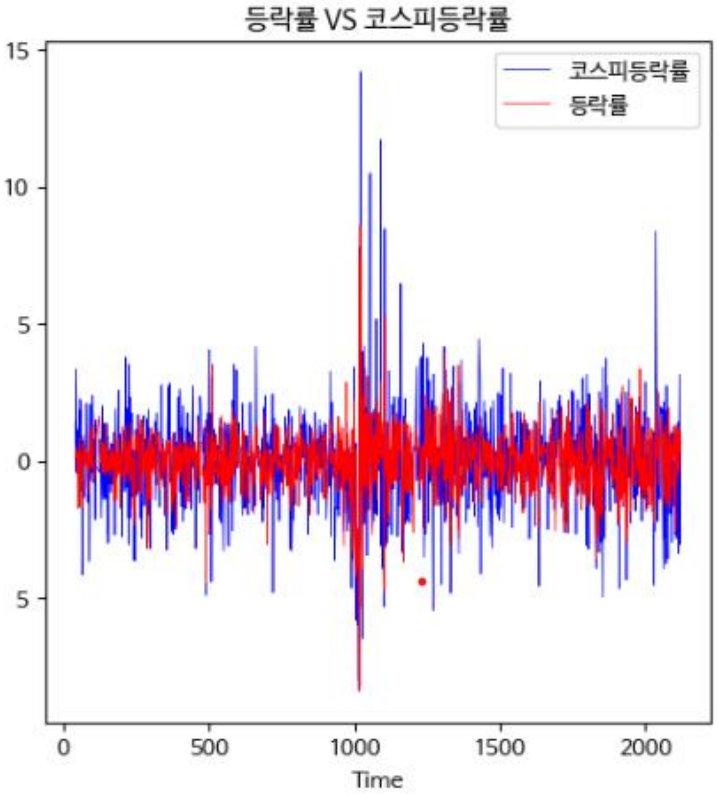
Hyundai

4 EDA

KOSPI change, KOSPI fluctuation rate



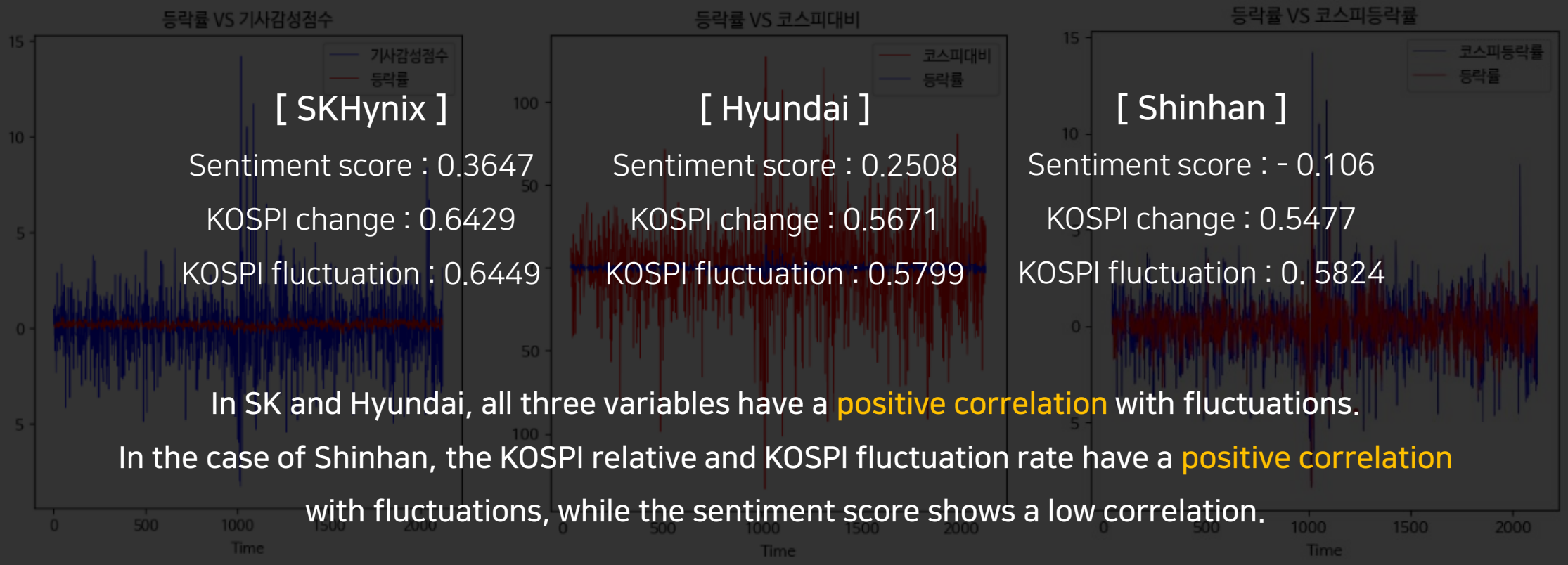




Shinhan

4 EDA

KOSPI change, KOSPI fluctuation rate



[SKHynix]

Sentiment score : 0.3647

KOSPI change : 0.6429

KOSPI fluctuation : 0.6449

[Hyundai]

Sentiment score : 0.2508

KOSPI change : 0.5671

KOSPI fluctuation : 0.5799

[Shinhan]

Sentiment score : - 0.106

KOSPI change : 0.5477

KOSPI fluctuation : 0. 5824

In SK and Hyundai, all three variables have a positive correlation with fluctuations.

In the case of Shinhan, the KOSPI relative and KOSPI fluctuation rate have a positive correlation with fluctuations, while the sentiment score shows a low correlation.

Shinhan

4 EDA

correlation

Correlation Coefficient in Time series data

Even if correlation between two variables is high, due to the attribute of time series data, the correlation could be high if both variables increase over time

An indicator that can reflect this attribute is needed!

4 EDA

correlation

Correlation Coefficient in Time series data

Even if correlation between two variables is high, due to the attribute of time series data, the correlation could be high if both variables increase over time

An indicator that can reflect this attribute is needed!

VAR (Vector Auto Regression)

4 EDA

About VAR

VAR model

A model that extends univariate autoregressive models to

multivariate autoregressive models by considering interactions between variables

▶ In time series data, it is possible to identify interactions between variables!

Available analysis methods through VAR model…

1. Causality test (Granger's Causality test)
2. Impulse response function
3. Forecasting error variance decomposition

4 EDA

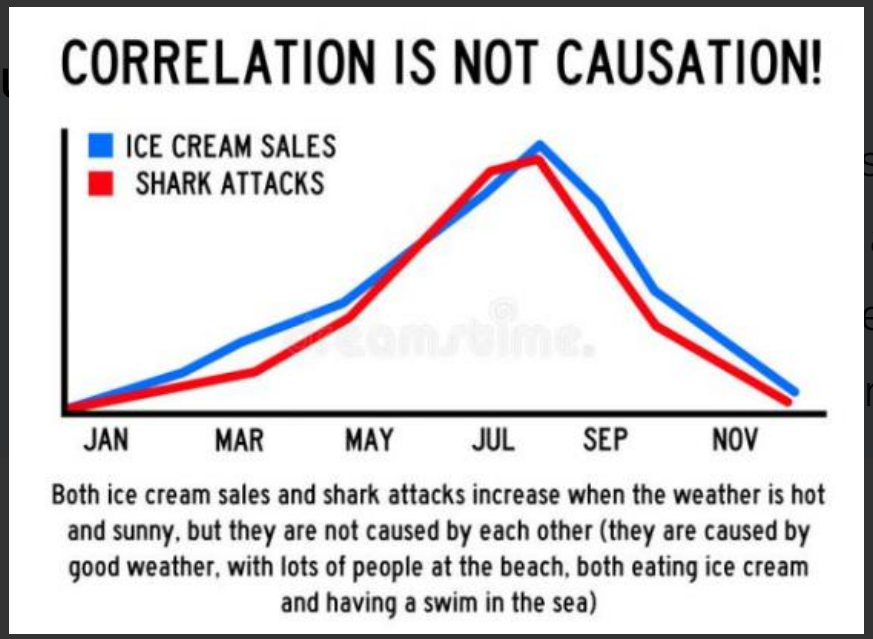
About VAR

Causality test

In general, in the regression analysis assume that the relationship between independent variable and dependent variable has already been determined by economic theory

However using VAR model, it is possible to analyze the functional relationship between variables without relying on economic theory

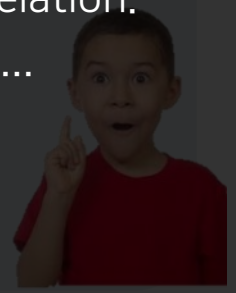
Granger's Causality test



According to graph, it seems that

Ice cream sales and shark attacks have high correlation.

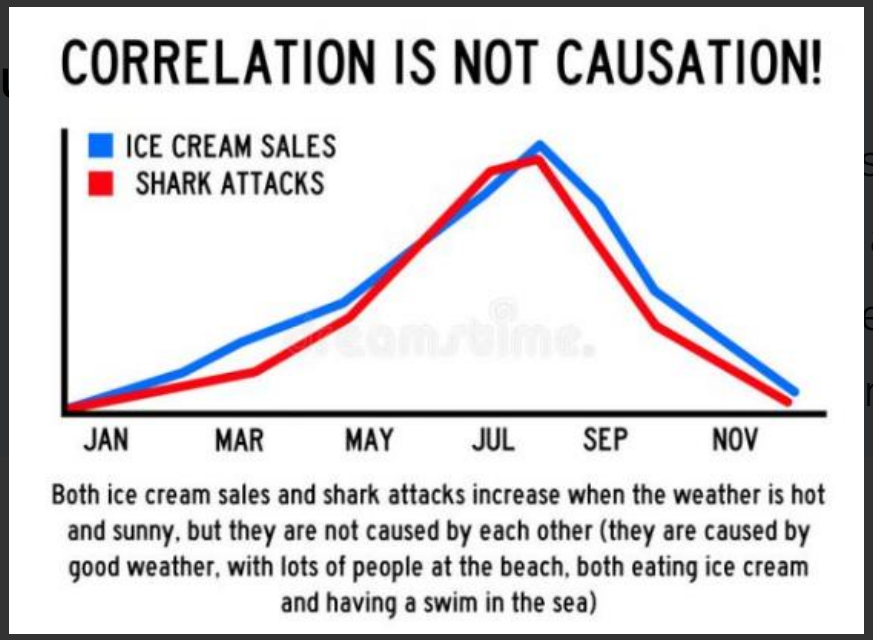
HOWEVER!!! They are completely unrelated…



Both events depend on the passage of time

And this creates a spurious relationship

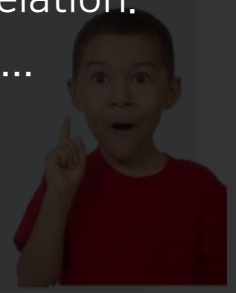
Granger's Causality test



According to graph, it seems that

Ice cream sales and shark attacks have high correlation.

HOWEVER!!! They are completely unrelated…



Both events depend on the passage of time

And this creates a spurious relationship

In such cases, it is possible to determine

if the two variables truly have correlation

by checking the causality between them

4 EDA

About VAR

VAR model

$$y_{1,t} = c_1 + \emptyset_{11,1}\, y_{1,t-1} + \emptyset_{12,1}\, y_{2,t-1} + \emptyset_{13,1}\, y_{3,t-1} + e_{1,t}$$
$$y_{2,t} = c_2 + \emptyset_{21,1}\, y_{1,t-1} + \emptyset_{22,1}\, y_{2,t-1} + \emptyset_{23,1}\, y_{3,t-1} + e_{2,t}$$
$$y_{3,t} = c_3 + \emptyset_{31,1}\, y_{1,t-1} + \emptyset_{32,1}\, y_{3,t-1} + \emptyset_{33,1}\, y_{3,t-1} + e_{3,t}$$

As shown in the equation above, the equation is expressed using two or more variables, and the Lag is determined using criteria such as AIC

4 EDA

About VAR

VAR model

$$y_{1,t} = c_1 + \emptyset_{11,1}\, y_{1,t-1} + \emptyset_{12,1}\, y_{2,t-1} + \emptyset_{13,1}\, y_{3,t-1} + e_{1,t}$$
$$y_{2,t} = c_2 + \emptyset_{21,1}\, y_{1,t-1} + \emptyset_{22,1}\, y_{2,t-1} + \emptyset_{23,1}\, y_{3,t-1} + e_{2,t}$$
$$y_{3,t} = c_3 + \emptyset_{31,1}\, y_{1,t-1} + \emptyset_{32,1}\, y_{3,t-1} + \emptyset_{33,1}\, y_{3,t-1} + e_{3,t}$$

As shown in the equation above, the equation is expressed using two or more variables, and the Lag is determined using criteria such as AIC

Before constructing the model, data to be used must satisfy the stationarity requirements in advance

4 EDA

About VAR

VAR model

$$y_{1,t} = c_1 + \emptyset_{11,1}\, y_{1,t-1} + \emptyset_{12,1}\, y_{2,t-1} + \emptyset_{13,1}\, y_{3,t-1} + e_{1,t}$$
$$y_{2,t} = c_2 + \emptyset_{21,1}\, y_{1,t-1} + \emptyset_{22,1}\, y_{2,t-1} + \emptyset_{23,1}\, y_{3,t-1} + e_{2,t}$$
$$y_{3,t} = c_3 + \emptyset_{31,1}\, y_{1,t-1} + \emptyset_{32,1}\, y_{3,t-1} + \emptyset_{33,1}\, y_{3,t-1} + e_{3,t}$$

As shown in the equation above, the equation is expressed using two or more variables, and the Lag is determined using criteria such as AIC

Using ADF Test, check the requirement based on p-value
if it does not satisfy it, normalized it through differencing

4 EDA

Result of ADF test

```
변수 이름 :  순매수_기관
상관계수 :  0.5055750025453257
ADF test statistic: -6.67301520460964
p-value: 4.537630517828522e-09
----------------------------------------
변수 이름 :  순매수_개인
상관계수 :  -0.7541191493885748
ADF test statistic: -20.906740001753008
p-value: 0.0
----------------------------------------
변수 이름 :  순매수_외국인
상관계수 :  0.6349264835779028
ADF test statistic: -20.22520834445593
p-value: 0.0
----------------------------------------
변수 이름 :  기사감성점수
상관계수 :  0.36375403088613123
ADF test statistic: -2.6733736847059912
p-value: 0.0787601969813459
----------------------------------------
변수 이름 :  코스피대비
상관계수 :  0.6428993515640895
ADF test statistic: -24.12643025336355
p-value: 0.0
----------------------------------------
변수 이름 :  코스피등락률
상관계수 :  0.6449025349106924
ADF test statistic: -23.292657037744313
p-value: 0.0
----------------------------------------
```

SKHynix

```
변수 이름 :  순매수_기관
상관계수 :  0.3853419868587516
ADF test statistic: -7.652220636689801
p-value: 1.7768688762323596e-11
----------------------------------------
변수 이름 :  순매수_기타법인
상관계수 :  -0.22749164499120803
ADF test statistic: -7.362603205596465
p-value: 9.414468192222495e-11
----------------------------------------
변수 이름 :  순매수_개인
상관계수 :  -0.6581470324244414
ADF test statistic: -6.328809952034797
p-value: 2.9408527070877686e-08
----------------------------------------
변수 이름 :  순매수_외국인
상관계수 :  0.4744294748167927
ADF test statistic: -5.6017988653858115
p-value: 1.2598721239693619e-06
----------------------------------------
변수 이름 :  코스피대비
상관계수 :  0.5417531175543868
ADF test statistic: -24.642926495602705
p-value: 0.0
----------------------------------------
변수 이름 :  코스피등락률
상관계수 :  0.5753231808034169
ADF test statistic: -23.851179656232862
p-value: 0.0
----------------------------------------
```

Hyundai

```
변수 이름 :  순매수_기관
상관계수 :  0.46789029815513855
ADF test statistic: -7.444208842143568
p-value: 5.896649517240632e-11
----------------------------------------
변수 이름 :  순매수_개인
상관계수 :  -0.5798550526446591
ADF test statistic: -6.516519865220665
p-value: 1.0679684761116576e-08
----------------------------------------
변수 이름 :  순매수_외국인
상관계수 :  0.44442857615510684
ADF test statistic: -5.243526078477658
p-value: 7.160526374121345 6e-06
----------------------------------------
변수 이름 :  기사감성점수
상관계수 :  0.25087665212638605
ADF test statistic: -3.7998438510739008
p-value: 0.0029089130050871264
----------------------------------------
변수 이름 :  코스피대비
상관계수 :  0.5671142635317672
ADF test statistic: -24.154868487137247
p-value: 0.0
----------------------------------------
변수 이름 :  코스피등락률
상관계수 :  0.5799177679929071
ADF test statistic: -23.32015038309087
p-value: 0.0
----------------------------------------
```

Shinhan

Result of stationarity test

Since the P-values for SK/Hyundai/Shinhan are all significant, run causality test without differencing

4 EDA

Result of Causality test

SK Hynix

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0006	0.2176
KOSPI Fluctuation(%)	0.0002	0.03
Net_purchase(institution)	0.0281	0.0001
Net_purchase(individual)	0.0104	0.000***
Net_purchase(foreign)	0.2009	0.0465
Sentiment score	0.84736	0.000***
SK Hynix(lag:5)		

4 EDA

Result of Causality test

SK Hynix

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0006	0.2176
KOSPI Fluctuation(%)	0.0002	0.03
Net_purchase(institution)	0.0281	0.0001
Net_purchase(individual)	0.0104	0.000***
Net_purchase(foreign)	0.2009	0.0465
Sentiment score	0.84736	0.000***
SK Hynix(lag:5)		

KOSPI change, net purchase(institution),
Net purchase(individual) **mutually influenced** by fluctuation(%).

4 EDA

Result of Causality test

SK Hynix

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0006	0.2176
KOSPI Fluctuation(%)	0.0002	0.03
Net_purchase(institution)	0.0281	0.0001
Net_purchase(individual)	0.0104	0.000***
Net_purchase(foreign)	0.2009	0.0465
Sentiment score	0.84736	0.000***
SK Hynix(lag:5)		

KOSPI change have influence on Fluctuation

4 EDA

Result of Causality test

SK Hynix

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0006	0.2176
KOSPI Fluctuation(%)	0.0002	0.03
Net_purchase(institution)	0.0281	0.0001
Net_purchase(individual)	0.0104	0.000***
Net_purchase(foreign)	0.2009	0.0465
Sentiment score	0.84736	0.000***
SK Hynix(lag:5)		

fluctuation(%) has influence on net purchase(foreign) and sentiment score

4 EDA

Result of Causality test

Hyundai

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0457	0.0128
KOSPI Fluctuation(%)	0.0124	0.0215
Net_purchase(institution)	0.0424	0.00***
Net_purchase(individual)	0.0096	0.00***
Net_purchase(foreign)	0.0342	0.00***
Sentiment score	0.0646	0.0049
Hyundai(lag:8)		

4 EDA

Result of Causality test

Hyundai

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0457	0.0128
KOSPI Fluctuation(%)	0.0124	0.0215
Net_purchase(institution)	0.0424	0.00***
Net_purchase(individual)	0.0096	0.00***
Net_purchase(foreign)	0.0342	0.00***
Sentiment score	0.0646	0.0049
Hyundai(lag:8)		

Every variable has mutual influence on fluctuation(%)

4 EDA

Result of Causality test

Shinhan

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0515	0.022
KOSPI Fluctuation(%)	0.0359	0.0118
Net_purchase(institution)	0.0554	0.1155
Net_purchase(individual)	0.0209	0.0795
Net_purchase(foreign)	0.2787	0.0465
Sentiment score	0.2466	0.3881
Shinhan(lag:4)		

4 EDA

Result of Causality test

Shinhan

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0515	0.022
KOSPI Fluctuation(%)	0.0359	0.0118
Net_purchase(institution)	0.0554	0.1155
Net_purchase(individual)	0.0209	0.0795
Net_purchase(foreign)	0.2787	0.0465
Sentiment score	0.2466	0.3881
Shinhan(lag:4)		

KOSPI change, KOSPI Fluctuation(%),net purchase(individual)
mutually influenced by fluctuation(%).

4 EDA

Result of Causality test

Shinhan

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0515	0.022
KOSPI Fluctuation(%)	0.0359	0.0118
Net_purchase(institution)	0.0554	0.1155
Net_purchase(individual)	0.0209	0.0795
Net_purchase(foreign)	0.2787	0.0465
Sentiment score	0.2466	0.3881
Shinhan(lag:4)		

net purchase(institution) has influence on fluctuation(%)

4 EDA

Result of Causality test

Shinhan

H0	X has no impact on fluctuation rate	fluctuation rate has no impact on X
variable	X->Fluctuation (%)	Fluctuation (%)->X
KOSPI change	0.0515	0.022
KOSPI Fluctuation(%)	0.0359	0.0118
Net_purchase(institution)	0.0554	0.1155
Net_purchase(individual)	0.0209	0.0795
Net_purchase(foreign)	0.2787	0.0465
Sentiment score	0.2466	0.3881
Shinhan(lag:4)		

fluctuation(%) has influence on net purchase(foreign),
sentiment score has no influence on fluctuation(%)

4 EDA

Correlation between categorical and continuous

Labeling is needed based on the fluctuation rate
for buy/sell recommendation.
How should the criteria be determined?

Professor Baek Chang-ryong

4 EDA

2023년 5월 1일 (월) 오전 10:21

김민 학생께,

안녕하세요.
여의도의 많은 증권맨들과 퀀트들을 생각해보시면 참 어려운 문제를 다루고 계시네요...
정답이 없는 문제이고, 만약 그런 공식을 알고 있다면.. 어찌보면 증권을 통해서는
이윤을 남길수 없을지도 몰라요. 여하튼 아래 질문에 대해서 답해드리자면 trial-and-error를 통해서
하나씩 선택해갈수 밖에 없는 문제입니다. 그래서 학습의 기본적인 기준,
즉 validation-train-test 를 통해서 모형을 선택하고, 이를 기준으로 최종 모형을 선택해보는 수밖에 없어요.
예를 들면 4)번의 경우 validation+train set을 통해서 tuning parameter 인 window size, 즉 4가 가장
optimal 한지를 결정하고, 이러한 튜닝모수 선택을 통해서 test set에서의 모형의 유용성을 살펴볼수 밖에 없어요.
Deep learning 은 data mining의 확장입니다. 다양한 모형을 다양한 방법으로 해보세요..

창룡.

Partial reply from the professor & summary…

1. 라벨링한 Y변수에 대해 X변수의 polyserial 상관계
하게 나옵니다. 라벨링한 Y변수 대신 원래의 등락률
에 대해 모델링을 진행해도 괜찮을까요? (혹시 라벨

예, 그렇게 선택해도 됩니다만 모형의 성능은 장담못합니다.

Professor Baek

Three-line summary…

- ✓ Questions without answers
- ✓ Unable to guarantee model performance
- ✓ Continuously reducing errors through trial and error

4 EDA

Labeling for Y variable

Polyserial Correlation

Used to measure the correlation between categorical variable and continuous variables, especially when there are more than 3 categories for the categorical variables

Wouldn't it be possible to discretize while preserving information if we choose the labeling criterion as the point where the correlations between each X variables and Y(fluctuation rate) are maximally similar to the correlations between X variables and labeled Y(buy/sell/hold)?

4 EDA

Labeling for Y variable

```
SK['day1_label'] = SK['등락률'].apply(lambda x: 'maintain' if abs(x) < 2 else 'buy' if x > 2 else 'sell')
SK['day3_label'] = SK['3일 등락률'].apply(lambda x: 'maintain' if abs(x) < 1 else 'buy' if x > 1 else 'sell')
```

```
> day1_df
      거래량  거래대금  시가총액 외국인.보유수량 외국인.지분율    토론방
1 0.2933809 0.2969062 0.1189235   -0.04856931  -0.0476637 0.1149921
  순매수_기관 순매수_기타법인 순매수_개인 순매수_외국인    검색어     보도량
1   0.2711209    -0.06113408 -0.3617469    0.2945925 0.2374334 0.05880711
  기사감성점수 뉴스심리지수 비트코인종가 비트코인거래량 비트코인변동
1    0.2562016   0.2082934  0.03304519  -0.03694125 -0.03178145
  코스피종가 코스피대비 코스피등락률 코스피거래량 코스피거래대금
1  0.1047094  0.4076674   0.4183276  0.09723681   0.1580599
  코스피시가총액   한은금리 원.미국달러    원.위안 원.일본엔.100엔.
1    0.0653132 -0.1005474 -0.07881063 -0.07522479    0.04143524
  원.유로 경제심리지수.원계열. 경제심리지수.순환변동치. 산업생산지수
1 0.06015472          -0.02575634          -0.06410021  -0.05504424
  물가상승률 소비자신뢰지수 소비자심리지수 경제활동참가율...  실업률...
1 -0.04262999  -0.02129617  -0.02085335   -0.08558252 0.08079371
  고용률...
1 -0.09311136
```

Process labeling with arbitrary values

Get a polyserial correlation,

and compare with original correlation

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

```
SK['day1_label'] = SK['등락률'].apply(lambda x: 'maintain' if abs(x) < 3 else 'buy' if x > 3 else 'sell')
SK['day3_label'] = SK['3일 등락률'].apply(lambda x: 'maintain' if abs(x) < 5 else 'buy' if x > 5 else 'sell')
```

```
> day1_df
       거래량     거래대금    시가총액 외국인.보유수량 외국인.지분율      토론방
1 -0.08897863 -0.05680121 0.05549611     -0.04462259    -0.0446082 -0.2027338
  순매수_기관 순매수_기타법인 순매수_개인 순매수_외국인        검색어
1   0.4698782       -0.174011   -0.771628     0.6616383 -0.004588057
      보도량 기사감성점수 뉴스심리지수 비트코인종가 비트코인거래량
1 0.09955342    0.4030009   0.03191558   0.02981671    -0.01917175
  비트코인변동 코스피종가 코스피대비 코스피등락률 코스피거래량 코스피거래대금
1   0.08908453 0.04030898  0.6240779    0.6339948  -0.00255103    -0.04798235
  코스피시가총액     한은금리  원.미국달러     원.위안 원.일본엔.100엔.
1     0.02660973 -0.03269142 -0.009822261 0.00101824        0.02116102
     원.유로 경제심리지수.원계열. 경제심리지수.순환변동치. 산업생산지수
1 0.03048799           0.01822798                0.01374958   0.02519877
  물가상승률 소비자신뢰지수 소비자심리지수 경제활동참가율...   실업률...
1 -0.0168883     0.03635738     0.03872439        0.01914623 -0.04144621
  고용률...
1 0.0337053
```

```
> day3_df
       거래량     거래대금    시가총액 외국인.보유수량 외국인.지분율      토론방
1 -0.194283 -0.1116863 0.04696376     0.07506627    0.07519493 -0.2419422
  순매수_기관 순매수_기타법인 순매수_개인 순매수_외국인        검색어
1   0.1390669       -0.10591   -0.3571065     0.3559347 -0.03050187
      보도량 기사감성점수 뉴스심리지수 비트코인종가 비트코인거래량
1 -0.04214468    0.2465913   0.09385573  -0.03063687    -0.02234154
  비트코인변동 코스피종가 코스피대비 코스피등락률 코스피거래량 코스피거래대금
1   0.00106814 0.01140719  0.2966115    0.2928902  -0.02985773    -0.03624546
  코스피시가총액     한은금리  원.미국달러     원.위안 원.일본엔.100엔.
1    -0.01574311 -0.02591017 -0.08403711 -0.07177428       0.008146044
     원.유로 경제심리지수.원계열. 경제심리지수.순환변동치. 산업생산지수
1 -0.1010702           0.01889489                0.004044409  -0.02760391
  물가상승률 소비자신뢰지수 소비자심리지수 경제활동참가율...   실업률...
1 -0.06785786     0.05334882     0.05612054       -0.01928165 0.07331032
    고용률...
1 -0.03871836
```

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

```
SK['day1_label'] = SK['등락률'].apply(lambda x: 'maintain' if abs(x) < 3 else 'buy' if x > 3 else 'sell')
SK['day3_label'] = SK['3일 등락률'].apply(lambda x: 'maintain' if abs(x) < 5 else 'buy' if x > 5 else 'sell')
```

SK Hynix (daily fluctuation rate)			
net_purchase_insti	0.5056	Sentiment score	0.3647
net_purchase_foreign	0.6349	KOSPI change	0.6429
net_purchase_indi	-0.7541	KOSPI fluctuation	0.6449

If the threshold for 1-day fluctuations deviates from 3%,

and for 3-day fluctuations deviates from 5%,

the new correlation will be either smaller or larger than the original one.

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

```
hd['day1_label'] = hd['등락률'].apply(lambda x: 'maintain' if abs(x) < 3 else 'buy' if x > 3 else 'sell')
hd['day3_label'] = hd['3일 등락률'].apply(lambda x: 'maintain' if abs(x) < 5 else 'buy' if x > 5 else 'sell')
```

```
> day1_df
       거래량     거래대금    시가총액 외국인.보유수량 외국인.지분율      토론방
1 -0.08897863 -0.05680121 0.05549611     -0.04462259    -0.0446082 -0.2027338
  순매수_기관 순매수_기타법인 순매수_개인 순매수_외국인        검색어
1   0.4698782       -0.174011   -0.771628     0.6616383 -0.004588057
      보도량 기사감성점수 뉴스심리지수 비트코인종가 비트코인거래량
1 0.09955342    0.4030009   0.03191558   0.02981671    -0.01917175
  비트코인변동 코스피종가 코스피대비 코스피등락률 코스피거래량 코스피거래대금
1   0.08908453 0.04030898  0.6240779    0.6339948  -0.00255103    -0.04798235
  코스피시가총액     한은금리  원.미국달러     원.위안 원.일본엔.100엔.
1     0.02660973 -0.03269142 -0.009822261 0.00101824       0.02116102
     원.유로 경제심리지수.원계열. 경제심리지수.순환변동치. 산업생산지수
1 0.03048799           0.01822798           0.01374958   0.02519877
  물가상승률 소비자신뢰지수 소비자심리지수 경제활동참가율...   실업률...
1 -0.0168883     0.03635738     0.03872439       0.01914623 -0.04144621
  고용률...
1 0.0337053
```

```
> day3_df
       거래량     거래대금    시가총액 외국인.보유수량 외국인.지분율      토론방
1 -0.194283 -0.1116863 0.04696376      0.07506627    0.07519493 -0.2419422
  순매수_기관 순매수_기타법인 순매수_개인 순매수_외국인        검색어
1   0.1390669       -0.10591   -0.3571065     0.3559347 -0.03050187
      보도량 기사감성점수 뉴스심리지수 비트코인종가 비트코인거래량
1 -0.04214468    0.2465913   0.09385573  -0.03063687    -0.02234154
  비트코인변동 코스피종가 코스피대비 코스피등락률 코스피거래량 코스피거래대금
1   0.00106814 0.01140719  0.2966115    0.2928902  -0.02985773    -0.03624546
  코스피시가총액     한은금리  원.미국달러      원.위안 원.일본엔.100엔.
1    -0.01574311 -0.02591017 -0.08403711 -0.07177428       0.008146044
     원.유로 경제심리지수.원계열. 경제심리지수.순환변동치. 산업생산지수
1 -0.1010702           0.01889489           0.004044409  -0.02760391
  물가상승률 소비자신뢰지수 소비자심리지수 경제활동참가율...   실업률...
1 -0.06785786     0.05334882     0.05612054       -0.01928165 0.07331032
  고용률...
1 -0.03871836
```

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

```
hd['day1_label'] = hd['등락률'].apply(lambda x: 'maintain' if abs(x) < 3 else 'buy' if x > 3 else 'sell')
hd['day3_label'] = hd['3일 등락률'].apply(lambda x: 'maintain' if abs(x) < 5 else 'buy' if x > 5 else 'sell')
```

Hyundai (daily fluctuation rate)			
net_purchase_insti	0.4678	Sentiment score	0.2508
net_purchase_foreign	0.4444	KOSPI change	0.5671
net_purchase_indi	-0.5798	KOSPI fluctuation	0.5799

If the threshold for 1-day fluctuations deviates from 3%,

and for 3-day fluctuations deviates from 5%,

the new correlation will be either smaller or larger than the original one.

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

```
shinhan['day1_label'] = shinhan['등락률'].apply(lambda x: 'maintain' if abs(x) < 3 else 'buy' if x > 3 else 'sell')
shinhan['day3_label'] = shinhan['3일 등락률'].apply(lambda x: 'maintain' if abs(x) < 5 else 'buy' if x > 5 else 'sell')
```

```
> day1_df
      거래량    거래대금      시가총액 외국인.보유수량 외국인.지분율      토론방
1 0.08388242 0.06423691 5.394795e-05      -0.1401925   -0.09924193 0.05994354
  순매수_기관 순매수_기타법인 순매수_개인 순매수_외국인      검색어
1   0.3859562      -0.1714454  -0.7124404     0.5396143 0.05462349
       보도량 기사감성점수 뉴스심리지수 비트코인종가 비트코인거래량
1 -0.03356901  0.002533326   0.06891142   0.04728857     0.01583635
  비트코인변동 코스피종가 코스피대비 코스피등락률 코스피거래량 코스피거래대금
1   0.01588969 0.06763642  0.6310515    0.6888132   0.05035021   -0.003208666
  코스피시가총액    한은금리 원.미국달러      원.위안 원.일본엔.100엔.
1     0.03964872 -0.1077085 -0.03307308 -0.02693606       0.05132923
      원.유로 경제심리지수.원계열. 경제심리지수.순환변동치. 산업생산지수
1 -0.01398052          -0.009911811               0.00464619   0.03076334
   물가상승률 소비자신뢰지수 소비자심리지수 경제활동참가율...   실업률...
1 -0.05151612    -0.01074382   -0.009353009       -0.06395089 0.05643626
   고용률...
1 -0.0652138
```

```
> day3_df
      거래량    거래대금    시가총액 외국인.보유수량 외국인.지분율      토론방
1 0.1523056 0.1906871 0.09944962     -0.08881154   -0.09596716 0.04692263
  순매수_기관 순매수_기타법인 순매수_개인 순매수_외국인      검색어
1   0.3018382      -0.1904523  -0.4778125      0.337664 0.07653775
       보도량 기사감성점수 뉴스심리지수 비트코인종가 비트코인거래량
1 -0.005603635   0.04107263    0.1580147   0.07414601     -0.0405551
  비트코인변동 코스피종가 코스피대비 코스피등락률 코스피거래량 코스피거래대금
1  -0.02256129  0.1254517  0.3694312    0.4227053   0.08341551     0.04741355
  코스피시가총액    한은금리 원.미국달러      원.위안 원.일본엔.100엔.
1        0.10412 -0.04281097 -0.02131133 -0.01757232      -0.03150734
      원.유로 경제심리지수.원계열. 경제심리지수.순환변동치. 산업생산지수
1 0.006260784          0.01630768               0.02733413   0.02173927
   물가상승률 소비자신뢰지수 소비자심리지수 경제활동참가율...   실업률...
1 0.007481365    0.004841717    0.001021169       -0.01278217 0.01268625
   고용률...
1 -0.01276686
```

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

```
shinhan['day1_label'] = shinhan['등락률'].apply(lambda x: 'maintain' if abs(x) < 3 else 'buy' if x > 3 else 'sell')
shinhan['day3_label'] = shinhan['3일 등락률'].apply(lambda x: 'maintain' if abs(x) < 5 else 'buy' if x > 5 else 'sell')
```

신한지주 (1일)			
net_purchase_insti	0.3853	Sentiment score	-
net_purchase_foreign	0.4744	KOSPI change	0.5477
net_purchase_indi	-0.6581	KOSPI fluctuation	0.5824

If the threshold for 1-day fluctuations deviates from 3%,

and for 3-day fluctuations deviates from 5%,

the new correlation will be either smaller or larger than the original one.

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

SK Hynix (daily fluctuation rate)			
net_purchase_insti	0.5056	Sentiment score	0.3647
net_purchase_foreign	0.6349	KOSPI change	0.6429
net_purchase_indi	-0.7541	KOSPI fluctuation	0.6449

Hyundai (daily fluctuation rate)			
net_purchase_insti	0.4678	Sentiment score	0.2508
net_purchase_foreign	0.4444	KOSPI change	0.5671
net_purchase_indi	-0.5798	KOSPI fluctuation	0.5799

Shinhan (daily fluctuation rate)			
net_purchase_insti	0.3853	Sentiment score	-
net_purchase_foreign	0.4744	KOSPI change	0.5477
net_purchase_indi	-0.6581	KOSPI fluctuation	0.5824

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

SK Hynix (daily fluctuation rate)			
net_purchase_insti	0.5056	Sentiment score	0.3647
net_purchase_foreign	0.6349	KOSPI change	0.6429
net_purchase_indi	-0.7541	KOSPI fluctuation	0.6449

Hyundai (daily fluctuation rate)			
net_purchase_insti	0.4678	Sentiment score	0.2508
net_purchase_foreign	0.4444	KOSPI change	0.5671
net_purchase_indi	-0.5798	KOSPI fluctuation	0.5799

Shinhan (daily fluctuation rate)			
net_purchase_insti	0.3853	Sentiment score	-
net_purchase_foreign	0.4744	KOSPI change	0.5477
net_purchase_indi	-0.6581	KOSPI fluctuation	0.5824

Interpretation

Variables with positive correlation indicates that as the variable increases, the chances of BUY increases, and the SELL decreases

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

SK Hynix (daily fluctuation rate)			
net_purchase_insti	0.5056	Sentiment score	0.3647
net_purchase_foreign	0.6349	KOSPI change	0.6429
net_purchase_indi	-0.7541	KOSPI fluctuation	0.6449

Hyundai (daily fluctuation rate)			
net_purchase_insti	0.4678	Sentiment score	0.2508
net_purchase_foreign	0.4444	KOSPI change	0.5671
net_purchase_indi	-0.5798	KOSPI fluctuation	0.5799

Shinhan (daily fluctuation rate)			
net_purchase_insti	0.3853	Sentiment score	-
net_purchase_foreign	0.4744	KOSPI change	0.5477
net_purchase_indi	-0.6581	KOSPI fluctuation	0.5824

Interpretation

Variables with negative correlation indicates that as the variable increases, the chances of BUY decreases, and the SELL increases

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

Result

The thresholds for 1-day and 3-day fluctuation are the same for all three stocks

If examined at a highly precise decimal level, there may be slight variations…

SK Hynix (daily fluctuation rate)			
net_purchase_insti	0.5056	Sentiment score	0.3647
net_purchase_foreign	0.6349	KOSPI change	0.6425
net_purchase_indi	-0.7541	KOSPI fluctuation	0.6449

Hyundai (daily fluctuation rate)			
net_purchase_insti	0.4678	Sentiment score	0.2508
net_purchase_foreign	0.4444	KOSPI change	0.5671
net_purchase_indi	-0.5798	KOSPI fluctuation	0.5799

Shinhan (daily fluctuation rate)			
net_purchase_insti	0.3853	Sentiment score	-
net_purchase_foreign	0.4744	KOSPI change	0.5477
net_purchase_indi	-0.6581	KOSPI fluctuation	0.5824

If those reasons can be identified and the model performance is quite valid, then it is possible to create a model that can be commonly applied to different stocks

Interpretation

Variables with negative correlation indicates that as the variable increases, the chances of BUY decreases, and the SELL increases

4 EDA

Labeling for Y variable;Compare Correlation of Y and labeled Y

so Why all thresholds are same???

SK Hynix	
mean	0.057786
std	2.301890
min	-9.080000
25%	-1.390000
50%	0.000000
75%	1.440000
max	13.400000

Hyundai	
mean	0.030651
std	2.104999
min	-10.340000
25%	-1.110000
50%	0.000000
75%	0.970000
max	19.420000

Shinhan	
mean	-0.011036
std	1.767651
min	-8.260000
25%	-0.920000
50%	0.000000
75%	0.910000
max	14.200000

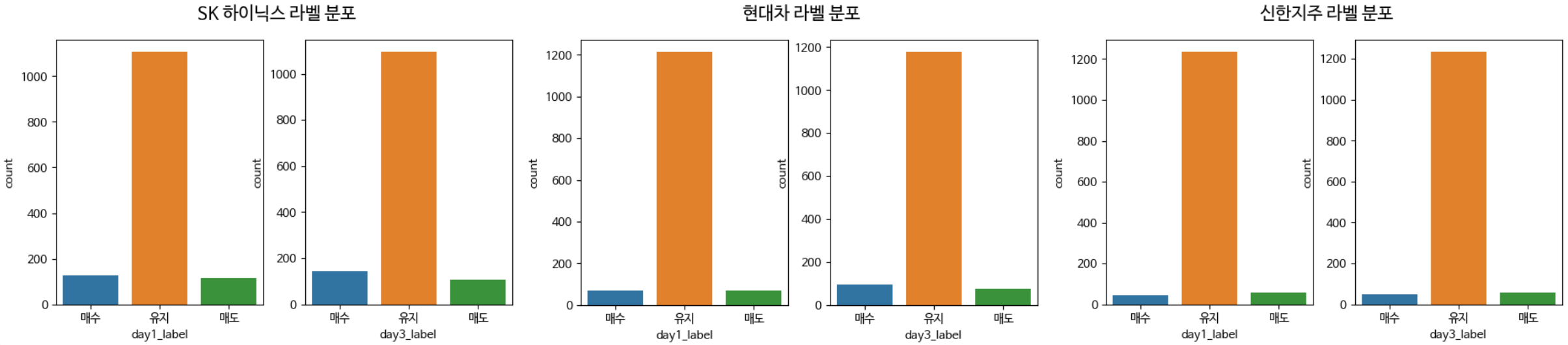
There is not much variation in the descriptive statistics among three stocks,

And guess this may be the reason for identical thresholds

4 EDA

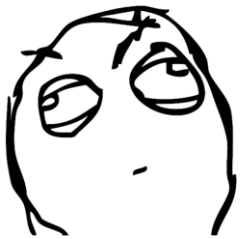
Class imbalance in labeled Y

Each stock's distribution of labeled Y



Class imbalance is very severe

See you in week 4



The quagmire of variable selection …

And modeling with class imbalance…

To be continued