Chapter 4: Initial Result

4.1 Exploratory Data Analysis (EDA)

4.1.1 Data processing

Step 1: Read the data of the three datasets

[5]:
```
import pandas as pd

stock_data = pd.read_csv('stock_data.csv')
profit_data = pd.read_csv('profit_data.csv')
china_data1_0 = pd.read_csv('china_data1_0.csv')

stock_data_head = stock_data.head()
stock_data_head
```

[5]:

	日期	股票代码	开盘	收盘	最高	最低	成交量	成交额	振幅	涨跌幅	涨跌额	换手率
0	2011-06-30	2594	22.00	25.45	26.19	22.00	562925	1.349342e+09	23.28	41.39	7.45	87.96
1	2011-07-01	2594	25.77	28.00	28.00	25.00	335613	8.986792e+08	11.79	10.02	2.55	52.44
2	2011-07-04	2594	28.88	30.80	30.80	28.58	219646	6.535179e+08	7.93	10.00	2.80	34.32
3	2011-07-05	2594	32.78	33.05	33.88	31.89	381081	1.257046e+09	6.46	7.31	2.25	59.54
4	2011-07-06	2594	32.80	33.40	35.55	32.11	312967	1.056614e+09	10.41	1.06	0.35	48.90

[7]:
```
profit_data_head = profit_data.head()
profit_data_head
```

[7]:

	code	pubDate	statDate	roeAvg	npMargin	gpMargin	netProfit	epsTTM	MBRevenue	totalShare	liqaShare
0	sz.002594	2011-08-23	2011-06-30	0.014288	0.017500	0.157585	3.945320e+08	0.160401	2.183512e+10	2.354100e+09	64000000.0
1	sz.002594	2011-10-29	2011-09-30	0.018285	0.014911	0.158514	5.119430e+08	0.188447	NaN	2.354100e+09	79000000.0
2	sz.002594	2012-03-26	2011-12-31	0.069957	0.032668	0.171793	1.595076e+09	0.588176	4.728200e+10	2.354100e+09	79000000.0
3	sz.002594	2012-04-26	2012-03-31	0.001279	0.004568	0.174024	5.360600e+07	0.486353	NaN	2.354100e+09	79000000.0
4	sz.002594	2012-08-28	2012-06-30	0.000770	0.004612	0.156560	1.041540e+08	0.478115	2.198346e+10	2.354100e+09	79000000.0

[9]:
```
china_data1_0_head = china_data1_0.head()
china_data1_0_head
```

[9]:

	年份	中国	GDP(美元)	占世界%
0	2023	17.79万亿 (17,794,781,986,104)	16.8775%	NaN
1	2022	17.88万亿 (17,881,783,387,000)	17.6654%	NaN
2	2021	17.82万亿 (17,820,459,508,852)	18.2723%	NaN
3	2020	14.69万亿 (14,687,744,162,801)	17.1630%	NaN
4	2019	14.28万亿 (14,279,968,506,271)	16.2373%	NaN

Step 2: Combine the three datasets

```
[11]: # Remove the market prefix from the 'code' column in profit_data
profit_data['股票代码'] = profit_data['code'].str.replace('sz.', '').astype(int)

# Ensure the '日期' column in stock_data and 'statDate' in profit_data are in datetime format
stock_data['日期'] = pd.to_datetime(stock_data['日期'])
profit_data['statDate'] = pd.to_datetime(profit_data['statDate'])

# Merge stock_data and profit_data on the common columns '股票代码' and 'statDate'
merged_data = pd.merge(stock_data, profit_data, how='left', left_on=['股票代码', '日期'], right_on=['股票代码', 'statDate'])

# Convert the '年份' column in china_data1_0 to integer format
china_data1_0['年份'] = china_data1_0['年份'].astype(int)

# Add the china_data1_0 to the merged data, matching the '年份' with the year part of '日期' in merged_data
# First, extract the year from the '日期' column in merged_data
merged_data['年份'] = merged_data['日期'].dt.year

# Now merge the china_data1_0 into the merged_data
final_data = pd.merge(merged_data, china_data1_0, how='left', on='年份')

# Display the first few rows of the final_data
final_data.head()
```

[11]:

	日期	股票代码	开盘	收盘	最高	最低	成交量	成交额	振幅	涨跌幅	...	gpMargin	netProfit	epsTTM	MBRevenue	totalShare	liqaShare
0	2011-06-30	2594	22.00	25.45	26.19	22.00	562925	1.349342e+09	23.28	41.39	...	0.157585	394532000.0	0.160401	2.183512e+10	2.354100e+09	64000000.0
1	2011-07-01	2594	25.77	28.00	28.00	25.00	335613	8.986792e+08	11.79	10.02	...	NaN	NaN	NaN	NaN	NaN	NaN
2	2011-07-04	2594	28.88	30.80	30.80	28.58	219646	6.535179e+08	7.93	10.00	...	NaN	NaN	NaN	NaN	NaN	NaN

Step 3: Process null values and delete unnecessary data

```
final_data_cleaned.head()
```

	TIME	Stock_Price	Turnover rate	GDP(Trillion)	npMargin
0	2011-06-30	25.45	87.96	7.55	0.017500
64	2011-09-30	19.30	6.84	7.55	0.014911
366	2012-12-31	20.35	2.66	8.53	0.004544
543	2013-09-30	39.36	1.87	9.57	0.016119
604	2013-12-31	37.68	1.22	9.57	0.014677

Variable	Describe
TIME	Time Sequence
Stock_price	Stock Price
Turnover rate	Represents high and low emotions
GDP(Trillion)	Gross Domestic Product
Npmargin	Represents the company's profitability

4.1.2 VAR Model

The vector autoregression model, referred to as the VAR model, is a generalization of the AR model and is a commonly used econometric model. It is used to determine the correlation.

Step 1: Define variables

Variable definition:

1. Turnover rate (represents the public sentiment towards this stock): Turnover rate
2. Company net profit margin (represents the company's operating conditions): npMargin
3. China's GDP (represents the value of GDP): GDP(Trillion)
4. Stock price: Stock_Price

Step 2: Stationarity test

The stability test method used is the ADF unit root test. Generally speaking, the P value of the unit root test t statistic should be significant at the 5% significance level. If the P value is lower than 0.05, the time series can be considered stable. If it is higher than 0.05, it is unstable. The unstable time series can be processed by first-order difference to test the stability of its first-order difference terms. If it is still unstable, the second-order difference method can be used to test until it is stable.

Null Hypothesis: LNTURNOVER RATE has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-5.488289	0.0001
Test critical values:	1% level	-3.653730	
	5% level	-2.957110	
	10% level	-2.617434	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: LNNPMARGIN has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.739282	0.0786
Test critical values:	1% level	-3.653730	
	5% level	-2.957110	
	10% level	-2.617434	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: LNGDP TRILLION has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.472742	0.3384
Test critical values:	1% level	-4.273277	
	5% level	-3.557759	
	10% level	-3.212361	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: LNSTOCK PRICE has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.989035	0.5850
Test critical values:	1% level	-4.273277	
	5% level	-3.557759	
	10% level	-3.212361	

*MacKinnon (1996) one-sided p-values.

variable	ADF test value	5% critical value	P-value	Test results
lnTurnover rate	-5.488289	-2.957110	0.0001	Stable
lnnpMargin	-2.739282	-2.957110	0.0786	Unstable
lnGDP(Trillion)	-2.472742	-3.557759	0.3384	Unstable
lnStock_Price	-1.989035	-3.557759	0.5850	Unstable

The table above is the unit root stationarity test results, where d represents the first-order difference of the variable. According to the test results, the P values of lnnpMargin,lnGDP(Trillion), and lnStock_Price in the original variable sequence are all greater than 0.05, which means that at the 5% significance level, these variables are not stationary.After first-order difference processing of all indicators, it was found that the corresponding P values of all variable sequences at the 5% significance level were less than 0.05, and passed the stationarity test.

Null Hypothesis: D(LNTURNOVER RATE) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.474323	0.0000
Test critical values:	1% level	-3.661661	
	5% level	-2.960411	
	10% level	-2.619160	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(LNNPMARGIN) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.780305	0.0000
Test critical values:	1% level	-3.661661	
	5% level	-2.960411	
	10% level	-2.619160	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(LNGDP TRILLION) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.345588	0.0001
Test critical values:	1% level	-4.284580	
	5% level	-3.562882	
	10% level	-3.215267	

*MacKinnon (1996) one-sided p-values.

Null Hypothesis: D(LNSTOCK PRICE) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=8)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.416630	0.0000
Test critical values:	1% level	-4.284580	
	5% level	-3.562882	
	10% level	-3.215267	

*MacKinnon (1996) one-sided p-values.

variable	ADF test value	5% critical value	P-value	Test results
dlnTurnover rate	-6.474323	-2.960411	0.0000	Stable
dlnnpMargin	-6.780305	-2.960411	0.0000	Stable
dlnGDP(Trillion)	-6.345588	-3.562882	0.0001	Stable
dlnStock_Price	-6.416630	-3.562882	0.0000	Stable

Step 3: Determine the lag order of the model

Before conducting a cointegration test, it is necessary to reasonably determine the lag order of the model to avoid problems such as too little freedom or autocorrelation caused by a lag order that is too large or too small.

VAR Lag Order Selection Criteria
Endogenous variables: LNSTOCK PRICE LNGDP TRILLION LNNPMARGIN LN...
Exogenous variables: C
Date: 01/02/25 Time: 18:19
Sample: 2011Q2 2023Q2
Included observations: 31

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-63.09299	NA	0.000891	4.328580	4.513610	4.388895
1	26.47936	150.2504*	7.83e-06*	-0.418023*	0.507130*	-0.116446*
2	38.62415	17.23777	1.06e-05	-0.169300	1.495976	0.373538

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Lag	LogL	LR	FPE	AIC	SC	HQ

0	-63.09299	NA	0.000891	4.328580	4.513610	4.388895
1	26.47936	150.2504*	7.83e-06*	-0.418023*	0.507130*	-0.116446*
2	38.62415	17.23777	1.06e-05	-0.169300	1.495976	0.373538

A VAR model is established to study the impact of the three influencing factors on stock price fluctuations. According to the minimum AIC and SC value criteria, the line with the most * in the above figure is the lag order. Therefore, the lag orders determined by LR, FPE, AIC, SC and HQ in the lag order are all 1, so it is necessary to establish a VAR (1) model.

Step 4: Cointegration test

The unit root stationarity test assumes that the variable sequences are all first-order single integrated stationary sequences. The Johansen test can be used for cointegration test to observe whether these variable sequences have a long-term equilibrium relationship.

Date: 01/02/25 Time: 18:19
Sample (adjusted): 2011Q3 2023Q2
Included observations: 32 after adjustments
Trend assumption: Linear deterministic trend
Series: LNSTOCK PRICE LNGDP TRILLION LNNPMARGIN LNTURNOVER...
Lags interval (in first differences): No lags

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.637007	54.03092	47.85613	0.0118
At most 1	0.348164	21.60303	29.79707	0.3211
At most 2	0.161723	7.908236	15.49471	0.4753
At most 3	0.068283	2.263233	3.841466	0.1325

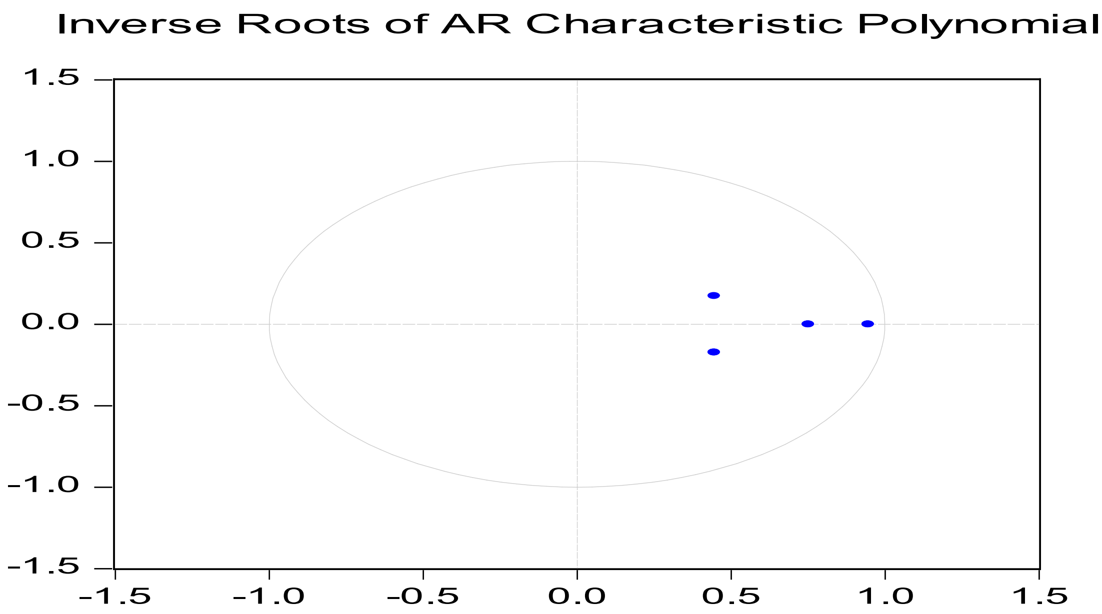
Trace test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

hypothesis	Eigenvalue	Trace Statistics	5% critical value	P-value
None*	0.637007	54.03092	47.85613	0.0118
At most 1	0.348164	21.60303	29.79707	0.3211
At most 2	0.161723	7.908236	15.49471	0.4753
At most 3	0.068283	2.263233	3.841466	0.1325

The above table is the result of cointegration test. The trace statistic test shows that the P value rejects the null hypothesis of no cointegration relationship at the 5% significance level, which means that there is a cointegration relationship. Therefore, there is a long-term equilibrium relationship between the time series. The original sequence lnTurnover rate, lnnpMargin, lnGDP(Trillion), and lnStock_Price can be used to build a model.

Step 5: AR characteristic root test

The AR eigenvalue test is a method used to verify the stability of a vector autoregression (VAR) model. When determining the stability of a model, we focus on whether the modulus of the AR unit root is less than 1. If the modulus of all AR unit roots is less than 1, it means that they are all located within the so-called "unit circle", which is visually represented by a graph.



The figure above shows the results of the AR characteristic root test, from which we can observe that all the test points are within the unit circle. This finding confirms that the constructed VAR model is stable, thus indicating that the two VAR models are effective. In addition, this shows that there is a long-term and stable equilibrium relationship between the model variables, which is an important prerequisite for VAR analysis.

Step 6: Granger causality test

The cointegration relationship test shows that there is a long-term stable relationship between the variables. However, the Granger causality behind this stable relationship needs further testing.

Pairwise Granger Causality Tests
Date: 01/02/25 Time: 18:21
Sample: 2011Q2 2023Q2
Lags: 1

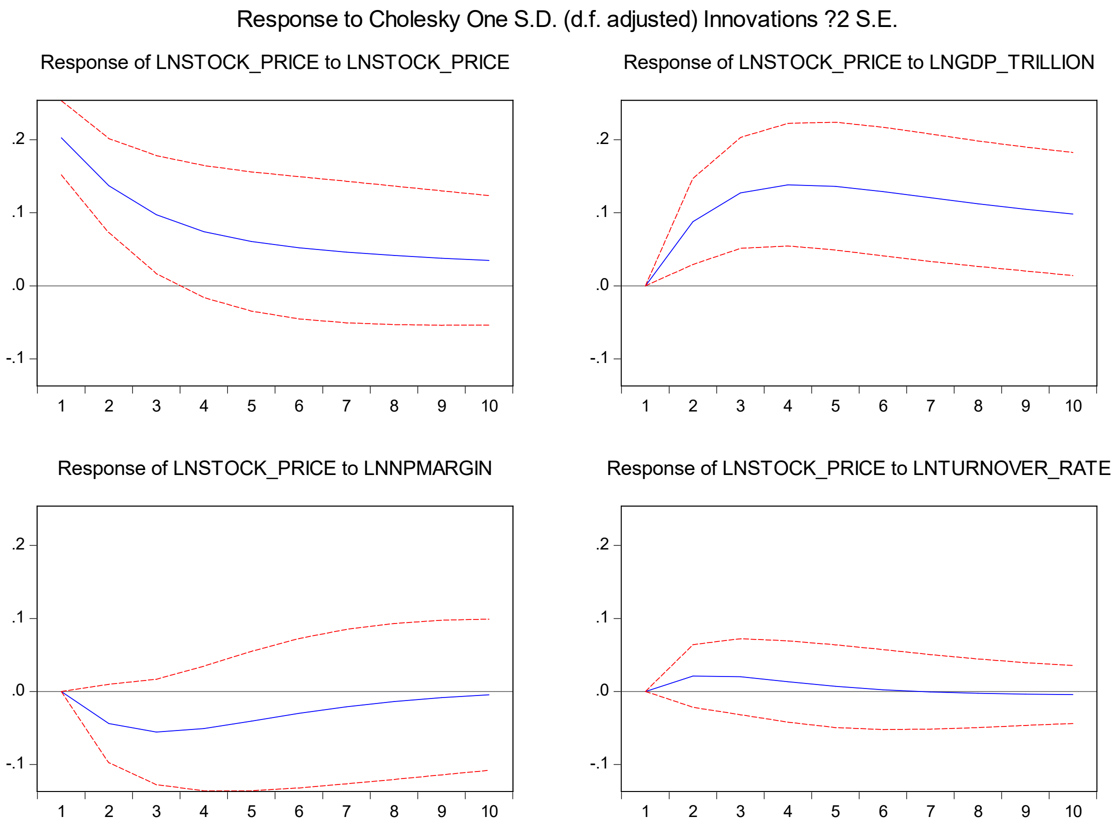
Null Hypothesis:	Obs	F-Statistic	Prob.
LNGDP TRILLION does not Granger Cause LNSTOCK PRICE	32	7.65775	0.0097
LNSTOCK PRICE does not Granger Cause LNGDP TRILLION		0.04856	0.8271
LNNPMARGIN does not Granger Cause LNSTOCK PRICE	32	2.73783	0.1088
LNSTOCK PRICE does not Granger Cause LNNPMARGIN		1.11490	0.2997
LNTURNOVER RATE does not Granger Cause LNSTOCK PRICE	32	0.00651	0.9362
LNSTOCK PRICE does not Granger Cause LNTURNOVER RATE		0.00904	0.9249

Null hypothesis	Statistics	P-value
Turnover rate is not the Granger cause of stock price	0.00651	0.9362
The company's net profit margin is not the Granger cause of the stock price	2.73783	0.1088
China's GDP is not a Granger cause for stock prices	7.65775	0.0097

The Granger causality test shows that at a 5% significance level, the P values of the null hypothesis that turnover rate and company net profit margin are not the Granger cause of stock price are 0.9362 and 0.1088 respectively, and the P values are both greater than 0.05, indicating that the test results cannot reject these two null hypotheses at a 5% significance level. The P value of the null hypothesis that China's GDP is not the Granger cause of stock price is 0.0097, and the P value is less than 0.05, which means that China's GDP is the Granger cause of stock price. From the perspective of Granger causality, China's GDP has a significant impact on the fluctuation of stock price, but the Granger causality test is a test of the statistical time sequence, which does not mean that there is a real causal relationship. The specific impact relationship should be determined by the pulse response and variance decomposition of the VAR model.

Step 7: Impulse response

The impulse response function analysis method has obvious intuitive advantages. By observing its impulse diagram, we can clearly see the influence of different variables. The impulse response results in the figure below show the impulse response diagrams between China's GDP, corporate net profit margin, turnover rate and stock price fluctuations, where the horizontal axis represents the lag period of the impact, the vertical axis represents the degree of impact on the variable, the solid line represents the impulse response function, and the dotted line represents the positive and negative standard deviations of the impact at a 95% confidence level.



From the impulse response diagram in the figure above, it can be seen that when a positive shock is given to the stock price, China's GDP will have a greater impact on the fluctuation of the stock price, and it is a positive shock from the first to the tenth period, reaching the maximum value of the positive shock in the fourth period, and the impact gradually converges in the long-term development. This shows that China's economic development level will greatly affect the fluctuation of the stock price, and will have a positive impact on the stock price. The company's net profit margin will also have a certain impact on the fluctuation of the stock price, but the impact from the first to the tenth period is negative, reaching the maximum value of the negative shock in the third period, and the impact between the seventh and tenth periods gradually approaches the horizontal axis. This means that the company's net profit margin will have a negative impact on the fluctuation of the stock price. The turnover rate will have a relatively small impact on the fluctuation of the stock price, and will have a positive impact from the first to the fifth period, reaching the maximum value of the positive shock in the second period, and the impact from the sixth to the tenth period is not obvious. This shows that the public's sentiment towards this stock will also have a certain degree of positive impact on the fluctuation of the stock price, and the rise of the public's sentiment towards the stock will drive the stock price to rise.

Step 8: Variance Decomposition

In order to further analyze the impact relationship between China's GDP, corporate net profit margin, turnover rate and stock price fluctuations, a model was established using the variance decomposition method to analyze the contribution of the impact.

Period	S.E.	LNSTOC...	LNGDP ...	LNNPMA...	LNTURN...
1	0.202681	100.0000	0.000000	0.000000	0.000000
2	0.264501	85.56577	11.04669	2.752281	0.635261
3	0.314699	69.99067	24.10792	5.042350	0.859058
4	0.355563	59.17305	34.01390	5.995267	0.817783
5	0.387738	52.19708	40.94219	6.139189	0.721547
6	0.413006	47.58975	45.83528	5.935484	0.639485
7	0.433171	44.39221	49.40089	5.625380	0.581524
8	0.449618	42.05672	52.08737	5.313128	0.542782
9	0.463311	40.27431	54.17263	5.035888	0.517171
10	0.474904	38.86631	55.83166	4.801850	0.500184

Cholesky Ordering: LNSTOCK PRICE LNGDP TRILLION LNNPMARGIN LNTURNOVER RATE

Period	S.E.	lnStock_Price	lnGDP(Trillion)	lnnpMargin	lnTurnover rate
1	0.202681	100.0000	0.000000	0.000000	0.000000
2	0.264501	85.56577	11.04669	2.752281	0.635261
3	0.314699	69.99067	24.10792	5.042350	0.859058
4	0.355563	59.17305	34.01390	5.995267	0.817783
5	0.387738	52.19708	40.94219	6.139189	0.721547
6	0.413006	47.58975	45.83528	5.935484	0.639485
7	0.433171	44.39221	49.40089	5.625380	0.581524
8	0.449618	42.05672	52.08737	5.313128	0.542782
9	0.463311	40.27431	54.17263	5.035888	0.517171
10	0.474904	38.86631	55.83166	4.801850	0.500184

The variance decomposition of stock price fluctuations shows that in the short term, stock price fluctuations are most affected by their own changes, and in the long term, they are most affected by China's GDP. Specifically, the impact of China's GDP on stock price fluctuations was 0 in the first period, and it increased rapidly to about 11% in the second period. After that, with the continuous rise of China's economic level in the long term, the impact on stock prices has greatly increased, and the contribution has stabilized at about 55% in the long term. The impact of the company's net profit margin on stock price fluctuations has generally increased first and then decreased, and the overall impact has not changed much. The contribution has stabilized at about 5% in the long term. In contrast, the turnover rate contributes less to stock price fluctuations, and the impact does not exceed 1% in both the long term and the short term. Overall, China's GDP, company net profit margin, and turnover rate will have a certain degree of impact on stock price fluctuations, among which China's GDP has the greatest impact on stock price fluctuations.

4.1.3 Arima Model

1.Convert the TIME column to date format.

•[3]:
```
import pandas as pd
new_file_path = 'merged_data.csv'
new_df = pd.read_csv(new_file_path)
new_df.head()
```

[3]:

	TIME	Turnover rate	npMargin	GDP(Trillion)	Stock_Price
0	2011/6/30	87.96	0.017500	7.55	25.45
1	2011/9/30	6.84	0.014911	7.55	19.30
2	2012/12/31	2.66	0.004544	8.53	20.35
3	2013/9/30	1.87	0.016119	9.57	39.36
4	2013/12/31	1.22	0.014677	9.57	37.68

2. Calculate lnTurnover rate as a new column.

•[7]:
```
new_df['TIME'] = pd.to_datetime(new_df['TIME'])
new_df['lnTurnover rate'] = np.log(new_df['Turnover rate'])
new_df.head()
```

[7]:

	TIME	Turnover rate	npMargin	GDP(Trillion)	Stock_Price	lnTurnover rate
0	2011-06-30	87.96	0.017500	7.55	25.45	4.476882
1	2011-09-30	6.84	0.014911	7.55	19.30	1.922788
2	2012-12-31	2.66	0.004544	8.53	20.35	0.978326
3	2013-09-30	1.87	0.016119	9.57	39.36	0.625938
4	2013-12-31	1.22	0.014677	9.57	37.68	0.198851

3. Adjust each feature according to the new impact.

Now that we have the lnTurnover rate column, I will adjust each feature according to the new impact. The new impacts are:

GDP(Trillion): 0.55

npMargin: 0.04

lnTurnover rate: 0.005

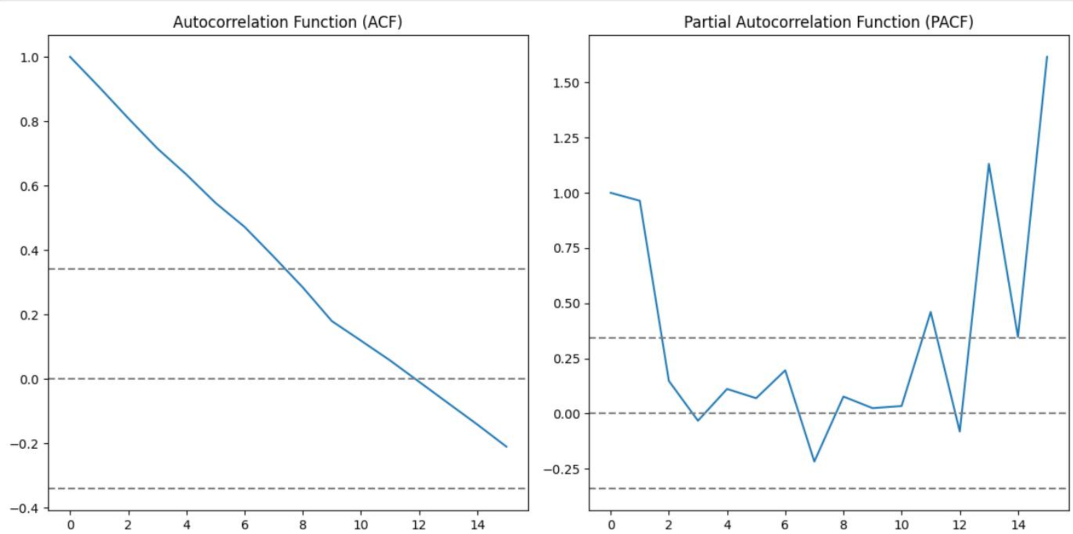
I will create a new DataFrame with these adjusted features and use it to train the ARIMA model. Next, let's adjust the features and prepare the data.

```
new_impact_degrees = {
    'GDP(Trillion)': 0.55,
    'npMargin': 0.04,
    'lnTurnover rate': 0.005
}

new_adjusted_features_df = new_df[['TIME']].copy()
for feature, impact in new_impact_degrees.items():
    new_adjusted_features_df[f'adj_{feature}'] = new_df[feature] * impact


new_adjusted_features_df.head()
```

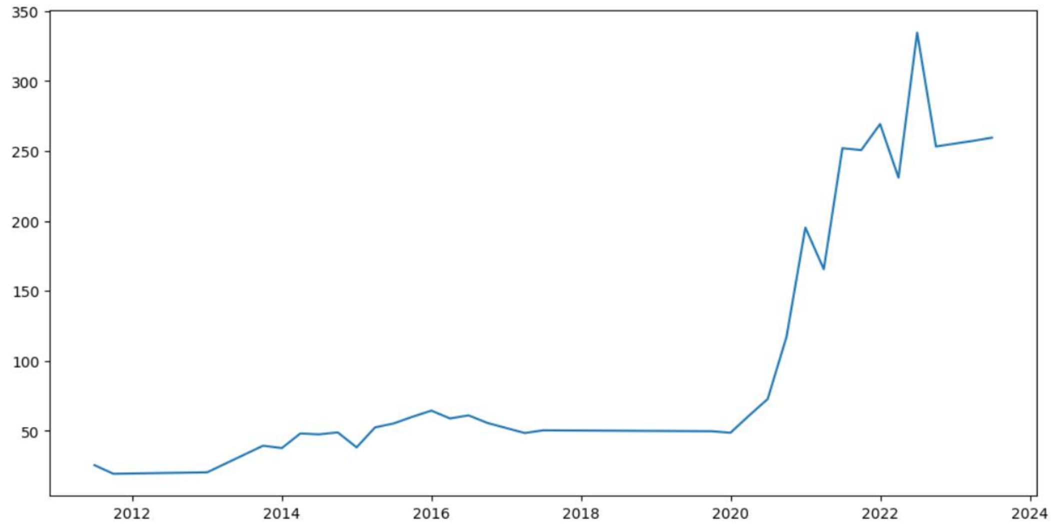
	TIME	adj_GDP(Trillion)	adj_npMargin	adj_lnTurnover rate
0	2011-06-30	4.1525	0.000700	0.022384
1	2011-09-30	4.1525	0.000596	0.009614
2	2012-12-31	4.6915	0.000182	0.004892
3	2013-09-30	5.2635	0.000645	0.003130
4	2013-12-31	5.2635	0.000587	0.000994

4. Forecast using the ARIMA model.

Stock price prediction using ARIMA model. Before applying the ARIMA model, the parameters (p, d, q) of the ARIMA model need to be determined. This is usually done through ACF (autocorrelation function) and PACF (partial autocorrelation function) plots. Since the adjusted features are being used to predict stock prices, a new time series is first created, which is the weighted sum of the adjusted features, and then the parameters of the ARIMA model will be determined based on this time series.



From the ACF plot, you can see that the lag for the first significant crossing of the confidence interval (blue dashed line) is 1, which is probably the p-value in the ARIMA model. From the PACF plot, you see that the lag for the first significant crossing of the confidence interval is also 1, which may be the q value in the ARIMA model. As for the d value, we need to determine the difference order of the time series to make it stationary. Typically, we can check stationarity by observing a time series plot or using the ADF (Augmented Dickey-Fuller) test.



5.Visualize the forecast results.

Now we have a visualization of the actual and predicted stock prices. As can be seen from the chart, the predicted stock price (red line) tries to follow the trend of the actual stock price (blue line).