Database Design and Analytical Implementation for Enhanced Macro Financial Insight

NUS Faculty of Science

QF5214 Project Report

Group 7

Name	Student No.
Su Guanting	**A0253559U**
Ma Yuchen	**A0253817Y**
Hu Ning	**A0253783W**
Wang Tianhui	**A0274561B**
Fang Dongke	**A0253700R**
Li Jiming	**A0253609B**
Jiang Dongli	**A0253553E**

1 Introduction

This project explores the robust and intricate systems employed in the acquisition, management, and analysis of macroeconomic and financial data. The focus is on several key financial assets such as stock indices, futures, options, and government bonds, which serve as critical barometers and indicators of China's economic health and fiscal stability. By leveraging comprehensive data sources and advanced technological tools, this research aims to provide deeper insights into the dynamic interplay between financial markets and macroeconomic indicators. Utilizing databases like the Wind database and Ricequant, coupled with innovative data processing techniques such as web scraping for sentiment analysis, the study ensures a holistic approach to understanding the complexities of the financial landscape. The methodologies adopted for data integration and processing, such as API utilization and the application of machine learning tools, set the stage for rigorous financial analysis and decision-making.

2 Data Source

To capture a snapshot of China's macroeconomic conditions during a specific period, the selection of data sources must be both targeted and as comprehensive as possible in terms of coverage. Thus, the data utilised in this study are focused on several core assets closely linked to the macroeconomic situation: stock indices, futures, options, and government bonds. The trends in stock indices are commonly viewed as a barometer of a nation's overall capital market condition. Futures and options, on the other hand, reflect the market's current confidence and expectations for the future. Government bonds, used as tools for implementing monetary policy, serve as indicators of a nation's economic and fiscal health.

The data sources are primarily categorised into two types: financial asset data and financial news data. Each asset category possesses two facets of data: a static base information table and a dynamic price time series table. For instance, the information table for options includes various horizontal dimensions of all historical options, such as ID, listing date, expiration date, exchange, option type, and strike price. The time series table records historical trading data, including Open, High, Low, Close, Volume (OHLCV), as well as turnover rate and open interest. Textual data covers four channels: global economy, commodities, foreign exchange, and Chinese A-shares, predominantly sourced from financial market news websites such as financialnews.com.cn.

The acquisition of financial asset data is primarily facilitated through API interfaces of financial databases, such as the Wind database and the Ricequant database, which have already integrated data from most Chinese exchanges, such as Shanghai Stock Exchange, Shenzhen Stock Exchange, Dalian Commodity Exchange, and Guangzhou Futures Exchange etc. Sentiment data, on the other hand, is obtained by scraping headlines from financial news websites using web crawling programs.

The original data volume exceeds 80 GB, primarily because the raw financial asset time series data and textual data are recorded at minute-level granularity. Due to constraints in database resources and budget limitations, as well as for ease of presentation, the database showcased in this study only displays price data aggregated at the daily level and textual data for specific time periods. See Figure 1 - Figure 3.

listed_date	exchange	underlying_symbol	symbol	underlying_order_book_id	round_lot	de_listed_date	maturity_date	option_type	exercise_type	type	cc
2015-02-09	XSHG	510050.XSHG	50ETF购3月2200	510050.XSHG	1.0	2015-03-25	2015-03-25	C	E	Option	
2015-02-09	XSHG	510050.XSHG	50ETF购3月2250	510050.XSHG	1.0	2015-03-25	2015-03-25	C	E	Option	
2015-02-09	XSHG	510050.XSHG	50ETF购3月2300	510050.XSHG	1.0	2015-03-25	2015-03-25	C	E	Option	
2015-02-09	XSHG	510050.XSHG	50ETF购3月2350	510050.XSHG	1.0	2015-03-25	2015-03-25	C	E	Option	
2015-02-09	XSHG	510050.XSHG	50ETF购3月2400	510050.XSHG	1.0	2015-03-25	2015-03-25	C	E	Option	
...	...	...	...	...	...	...	...	...	...	...	...

Figure 1: Static Option Information

	id	date	open	high	low	close	total_turnover	volume	open_interest
0	10000001	2015-02-09	0.1820	0.2029	0.1699	0.1826	4712265.0	2501.0	674.0
1	10000001	2015-02-10	0.1856	0.2144	0.1800	0.2072	3744873.0	1842.0	1087.0
2	10000001	2015-02-11	0.2083	0.2195	0.2028	0.2107	4231181.0	1999.0	1628.0
3	10000001	2015-02-12	0.2141	0.2143	0.1915	0.2109	4804803.0	2334.0	1930.0
4	10000001	2015-02-13	0.2130	0.2459	0.2090	0.2090	4604396.0	2066.0	2215.0
...	...	...	...	...	...	...	...	...	...

Figure 2: Option Time-series Data

	datetime	channel	content
0	2021-03-01 07:30:01	commodity	全球债市跌势暂歇，美股科技股上周五小幅反弹，道指则跌1.5%。美国众议院通过1.9万亿财政刺...
1	2021-03-01 07:33:11	commodity	周一到周五到期规模分别为200亿、100亿、100亿、200亿和200亿元。
2	2021-03-01 09:00:38	commodity	中国央行公开市场进行100亿元7天期逆回购操作，另有200亿元逆回购到期。
3	2021-03-01 09:00:59	commodity	乙二醇期货主力开盘涨超5%，硅铁涨超4%，锰硅、纯碱涨超3%。20号胶跌超3%，沪银、菜粕、...
4	2021-03-01 09:45:00	commodity	中国2月财新制造业PMI 50.9，预期 51.3，前值 51.5。
...	...	...	...

Figure 3: Textual Data

To maintain the accuracy and integrity of the data used in this study, several robust validation processes are implemented. After the database was established, data from March 2024 was simulated as new data to be inserted for testing and validation mechanisms. These tests include checking the data types of fields in the new data, as well as identifying outliers, anomalies, and missing values. This process helps to reduce the likelihood of unexpected errors in the database in the future. Furthermore, to ensure that the data management practices adhere to legal standards concerning data retention and privacy, the tables related to time series in the database are planned to be designed to retain data for a maximum of 20 years' history only.

3 Database Design

In this section, we describe the design and implementation of a financial database that can efficiently store, retrieve, and manage both static and time-series data related to financial products such as options, futures, and indices. This database will ensure data integrity, optimize queries, and support extensive data analysis.

3.1 Database Infrastructure: MySQL on Google Cloud

For the deployment of our financial database, we are leveraging MySQL hosted on Google Cloud. This choice brings several benefits in terms of scalability, security, and integration capabilities, making it a robust solution for handling both static and time-series financial data.

The initial setup involves provisioning a MySQL instance on Google Cloud SQL. This process includes configuring the database size, setting up network options, applying initial security settings, initializing connection and deployment using customized Python script. Figure 4 is the dashboard for our Google Cloud MySQL instance.

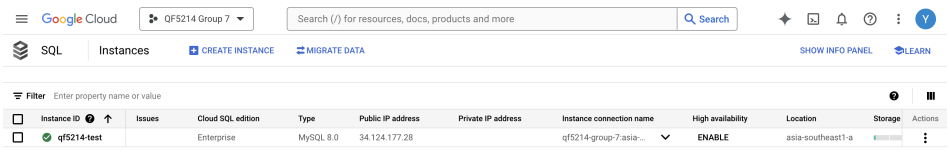

Figure 4: Google Cloud MySQL Database

3.2 Database Relations Design and Normalization

We have collected time series daily price date for different instrument including option, future and index. For each day, we have open, high, low, close price and volume data. And for each instrument, we have its listed date, exchange, underlying, maturity date etc.

Normalization reduces data redundancy and improves data integrity. The process involves dividing large tables into smaller, and more manageable pieces while ensuring relationships between the tables remain intact. We divided the initial time series table into smaller tables, namely exchange information, static information for option, future and index, and price information in time series format. After the normalization, the database is in 3NF format.

The following is a special case. Both exchange name and trading date exist in each static table. We suspect that this might be redundant dependency. However, for option, there are multiple trading time exist for a single exchange. Hence, this is not a dependency and 3NF is achieved. Our final ER diagram is shown in Figure 5.

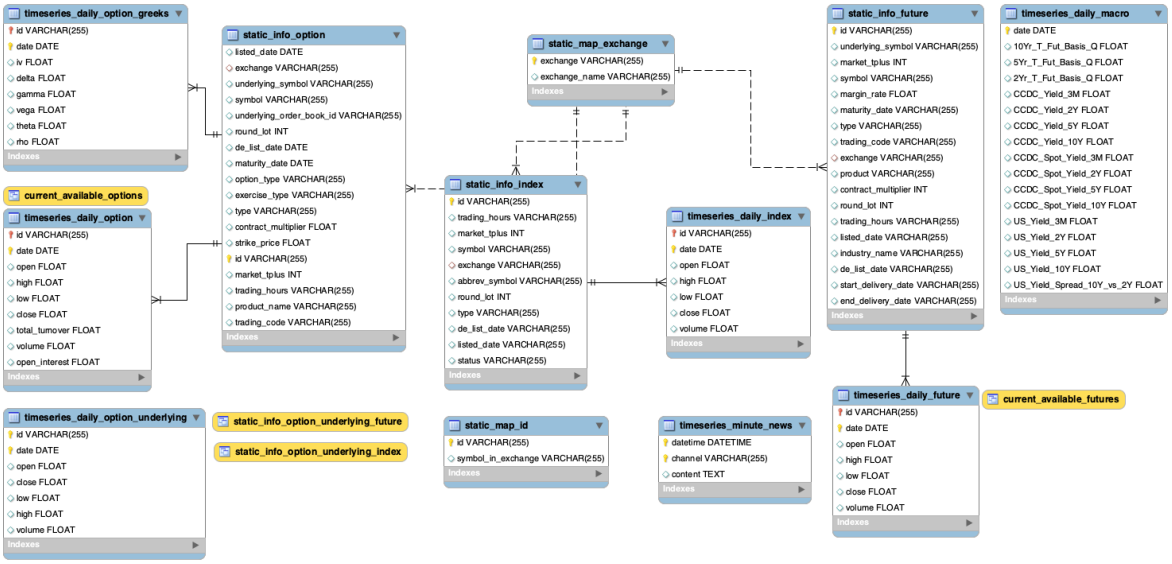

Figure 5: ER Diagram

3.3 Views and Stored Procedures

3.3.1 Creation of Foreign Keys

Proper use of foreign keys is vital to link related information across the database efficiently.

- Exchange Description to Financial Products: A foreign key from the financial products table to the exchange table ensures referential integrity and quick lookups.
- Time Series Data to Static Data: Linking these tables allows for correlating dynamic data points with static attributes of financial products.

3.3.2 Creation of Necessary Views

Views are virtual tables that do not store data themselves but display data stored in other tables based on a SQL query. The following views are created for simplification of complex query, reproducibility data integrity and performance optimization.

- Extract current available options based on expiry time.
- Extract underlying information of options, including commodity, ETF and stock index.

3.3.3 Utilization of stored procedures

We will utilize stored procedures to handle the creation of foreign keys and views, ensuring consistency and security in database modifications. Figure 6 is an example of utilizing a stored procedure to add foreign keys to existing table.

```
CALL AddForeignKey(
    'timeseries_daily_future',  -- Table name
    'fk_id_future',             -- Constraint name
    'id',                               -- Foreign key column
    'static_info_future',               -- Referenced table
    'id'                                -- Referenced column
);
```

Figure 6: Call Stored Procedure with Inputs

3.4 Conclusion and Future Work

This section described a foundational design for a financial database handling static and time-series data. Through careful normalization, strategic data cleaning, and the creation of necessary foreign keys and views, this database design will ensure robust data management and high query performance, critical for further financial analysis and decision-making.

However, due to time limitations, several key developments are planned to address future growth:

- Auto Update Function: To automate the process of extracting data from various external sources and importing it into our time series tables. This function will ensure our database is continuously updated with the latest data without manual input.
- Consider Different Time Zones: To improve the database's ability to handle data from multiple time zones. This will involve adding a new column for time zone information (e.g., UTC+8) to relevant time-series tables, allowing for accurate data representation and usability across different geographical locations.
- Improve Normalization: Refine our database schema to better represent the complexities of financial markets, such as exchange trading hours and various trading phases. This will include enhancements to the 'static_exchange_map' table and adjustments to how options are linked to multiple trading phases.

4 Data Integration and Processing

4.1 Data Cleaning

Since financial data obtained from APIs are typically pre-aggregated and cleaned by the platform, they only require basic checks. For static information tables, this involves checking each column for data type, uniqueness, and missing values. For time series tables, it is essential to verify that the start and end times are reasonable and to analyze the distribution, quantiles, and outliers of the price data.

For sentiment data acquired through web scraping, it is necessary to ensure that the content of the article titles is indeed related to financial economics. This can be achieved by setting up a basic list of keywords such as ['asset', 'future', 'option', 'bond', 'price', 'economy', 'market'] to filter out news titles that do not contain any of these keywords. Additionally, words like "rise" and "fall" can be used to categorize the sentiment of the news titles, further refining the relevance and emotional tone of the collected data.

4.2 ETL

The ETL (Extract, Transform, Load) process is implemented through the following steps: First, data required from data platforms and news websites is fetched using APIs and web scraping tools, and downloaded as pickle files. Next, this data is cleaned and organized in Python. Subsequently, the cleaned dataframe is written into a Google Cloud database using the mysql-python package. Finally, the necessary data is accessed from the database for subsequent analysis. This sequence ensures efficient and effective handling of data from collection through to utilization for analysis purposes.

4.3 Handling NULL Values

Use SQL commands to convert NaN and empty values to NULL, improving the integrity of queries and analytics. In Figure 7, empty value was detected from exchange value, and it was replaced to NULL to increase the database integrity.

id	trading_hours	market_tpl...	symbol	exchange	abbrev_symbol	round_lot	type	de_list
704843.INDX	09:31-11:30,13:01-15:00	0	MSCI中国A股国际		MSCIZGAGGJ	1	INDX	N/A
716567.INDX	09:31-11:30,13:01-15:00	0	MSCI中国A股国际通(人民币)		MSCIZGAGGJTRMB	1	INDX	N/A
718708.INDX	09:31-11:30,13:01-15:00	0	MSCI中国A股		MSCIZGAG	1	INDX	N/A
718711.INDX	09:31-11:30,13:01-15:00	0	MSCI中国A股人民币		MSCIZGAGRMB	1	INDX	N/A
801001.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-申万50		SW50	1	INDX	N/A
801003.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-申万A指		SWAZ	1	INDX	N/A
801004.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-申万股改		SWGG	1	INDX	N/A
801005.INDX	09:31-11:30,13:01-15:00	0	申万创业指数		SWCY	1	INDX	N/A
801010.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-农林牧渔		SWNLMY	1	INDX	N/A
801011.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-林业II		SWLYE	1	INDX	N/A
801012.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-农产品加工		SWNCPJG	1	INDX	N/A
801013.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-农业综合II		SWNYZHE	1	INDX	N/A
801014.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-饲料		SWSLII	1	INDX	N/A
801015.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-渔业		SWYY	1	INDX	N/A
801016.INDX	09:31-11:30,13:01-15:00	0	申银万国指数-种植业		SWZZY	1	INDX	N/A

```
UPDATE static_info_index
SET exchange = NULL
WHERE exchange = '';
```

Figure 7: Handle NULL values

5 Implementation

In recent years, the demand for automated market reports has surged, driven by the need for timely insights into financial markets. To address this, our team has developed a robust system leveraging Python's capabilities, which is "Automated Market Report Generation" .

5.1 Data Retrieval and Processing

We connect Python to databases using SQL queries, extracting essential data on macroeconomic indicators, stock market performance, bond market dynamics, and options data. Using NumPy and Pandas, we process retrieved data, computing key metrics reflecting macroeconomic conditions and market trends.

5.2 Dynamic Visualization

We utilise Pyecharts to facilitate dynamic data visualisation, generating interactive line charts that vividly portray market trends. These visualisations are exported as HTML and PNG files, ensuring accessibility and flexibility in viewing market dynamics.

5.3 Automated Report Generation

Automation is achieved through specialised packages like Reportlab. We script report generation, assembling processed data and insightful visualisations into PDF reports. Scheduled tasks ensure timely dissemination, providing users with regular statistical results.

5.4 Delivering Data Products

The outcome is a sophisticated data product: automated market reports. These reports serve users, offering comprehensive insights into macroeconomic trends, stock market movements, and bond market dynamics. They exemplify automation's potential in enhancing financial analysis and decision-making.

5.5 Future work

Looking ahead, we envision designing a more elegant and intuitive web UI for streamlined PDF output. Additionally, integrating ChatGPT-based viewpoint generation will add a layer of sophisticated analysis, enhancing the depth and relevance of insights provided in our reports.

6 Data Analysis

6.1 Description of the automated report

We automatically retrieve data from the database and generate plots for specific indicators selected by us. These indicators cover macroeconomic, fixed income, stock, futures, options, and other markets. For each major category of indicators, we can automatically generate weekly results from the database. This result can effectively assist us in market analysis and forecasting. Subsequent work can focus on further analysis of the preliminary reports, selecting intuitive content for in-depth analysis.

6.2 Example 2019/1-2024/4

6.2.1 Fixed income market analysis

For the study of fixed income, particularly in bonds, we utilise Chinese government bond yields to maturity (TYM), Chinese government bond futures settlement prices, and US Treasury bond yields to maturity (YTM) from our database for analysis.

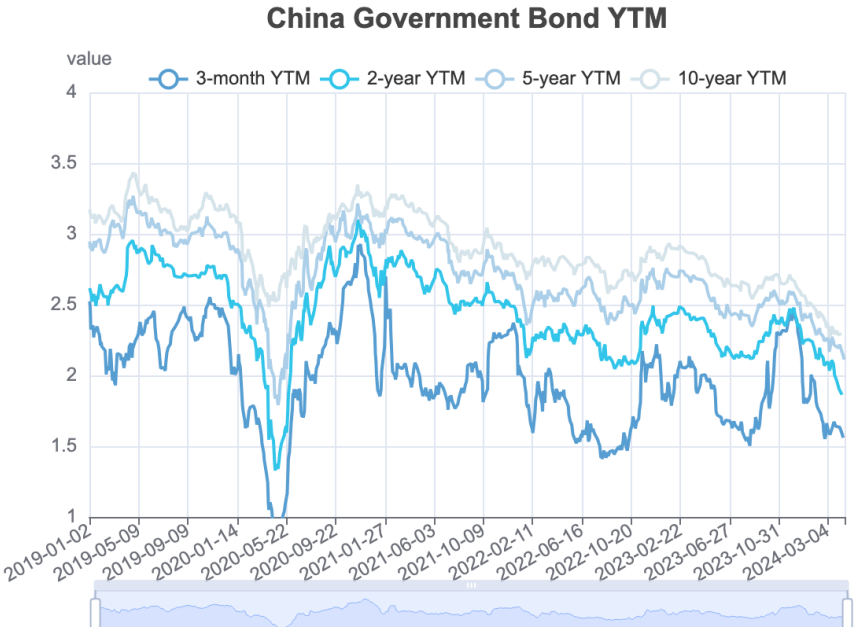

Figure 8: China Government Bond YTM

Firstly, we observe significant fluctuations in the overall yield curve of Chinese government bonds. These fluctuations are influenced by factors such as China's macroeconomic environment, central bank monetary policies, investor sentiment, and risk preferences.

We often use the 10-year minus 2-year government bond YTM spread as a key indicator, measuring market expectations regarding economic cycles, monetary policies, and risk premium. Here, we compare the bond situations of both China and the United States. It can be noted that the spread in the US exhibits greater volatility compared to China. However, during the onset of the COVID-19, both spreads showed an upward trend, indicating a more positive market outlook for the future.

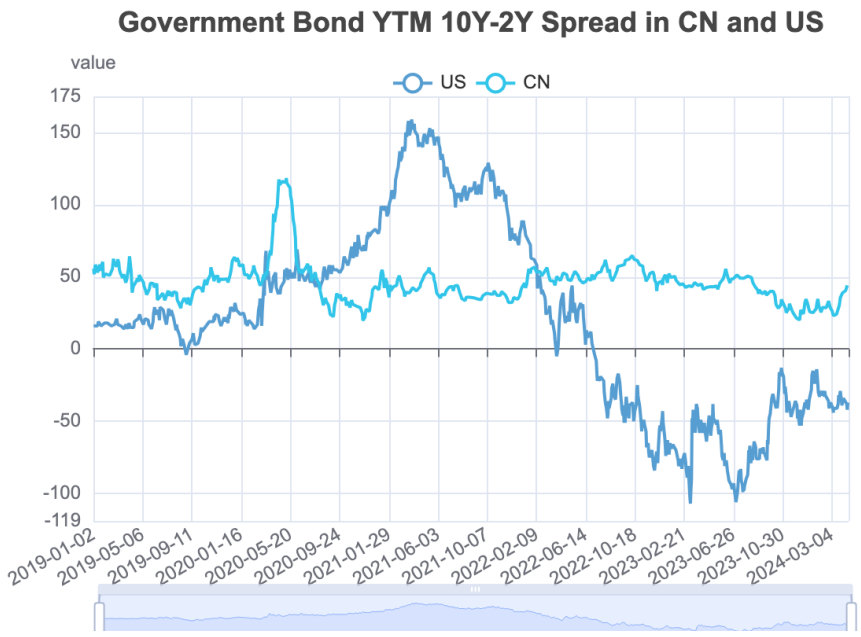

Figure 9: CN US YTM spread

6.2.2 Derivative market analysis

iVX is short for SSE 50 ETF Volatility Index, which is derived based on the SSE 50 ETF option contracts, reflecting the investors' expectations for the volatility of the 50 ETF in the next 30 days. When iVX rises, the investors expect a decreasing market and the market becomes more volatile.

Figure 10: VIX

From the figure we can see that during the beginning of Covid-19, iVX index increased and reached its local maximum in March, 2020, then decreased with the Covid-19 being effectively controlled. Although the financial market and macroeconomy still need a long time to recover, the investors regain confidence from April 2020.

6.2.3 Stock market analysis

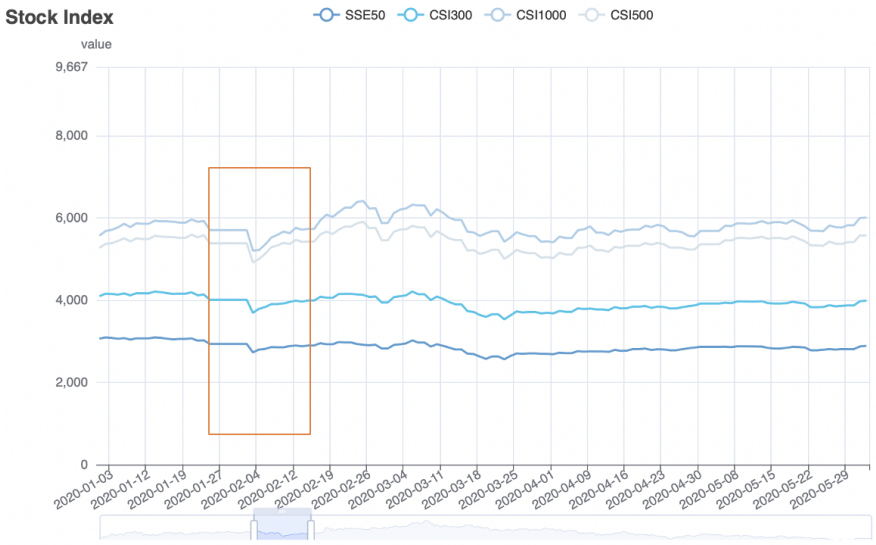

Figure 11: Stock Index

As a leading indicator to reflect the economy, the stock index acted fast and decreased in January, 2020. After a short period, the stock market began to recover but still performed worse than before the Covid-19 for a very long time. There are other indicators which reflect the condition of the stock market.

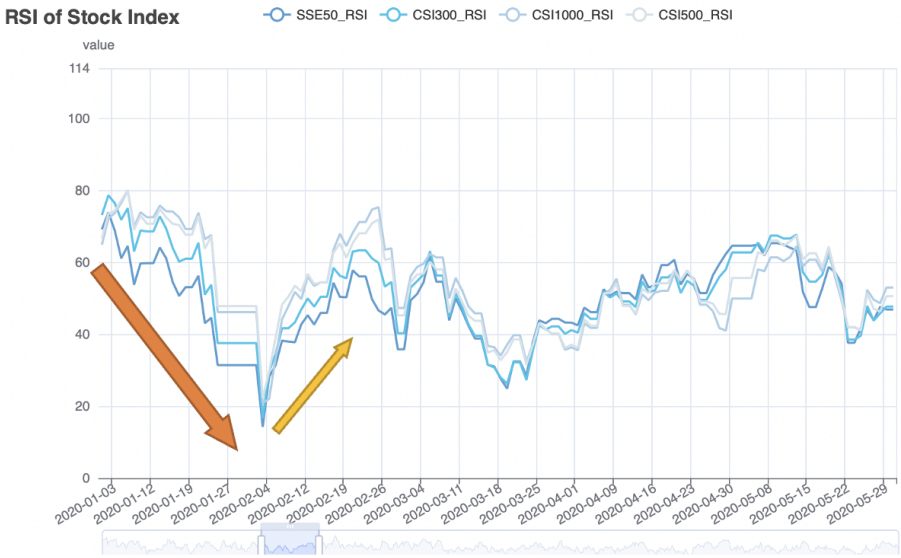

Figure 12: RSI of Stock Index

The above figure shows the trend of Relative Strength Index (RSI), which is derived based on the average of price gains and losses over a certain period, typically using data from 14 trading days.

The RSI showed a substantial and sustained decline at the beginning of 2020, which indicated a shift in market sentiment towards pessimism, with investors reducing their buying pressure on stocks, leading to a further decline until April, 2020.

The above index values and indicators help us to analyse the impact of Covid-19 to different markets, showing a common decreasing trend at the beginning of 2020.

6.3 Conclusion and future work

For the analysis of fixed income markets, particularly focusing on bond markets, derivative markets, and stock markets, we utilise data such as Chinese government bond yields, bond futures prices, and US Treasury bond yields for comprehensive analysis.

Significant fluctuations are observed in fixed income markets, influenced by factors including macroeconomic conditions, central bank policies, and investor sentiment. Key indicators such as the 10-year minus 2-year government bond yield spread reflect market expectations regarding economic cycles and risk.

The iVX index in derivative markets reflects investor expectations for market volatility. Stock market indicators such as the Relative Strength Index (RSI) show changes in market sentiment.

Further research is needed to deepen our understanding of key indicators in different markets. Exploring additional technical and sentiment indicators will enhance our ability to understand and predict market trends.

Looking ahead, we envision designing a more elegant and intuitive web UI for streamlined PDF output. Additionally, integrating ChatGPT-based viewpoint generation will add a layer of sophisticated analysis, enhancing the depth and relevance of insights provided in our reports.

7 Conclusion

This study has effectively demonstrated the integration and sophisticated application of database systems for the detailed analysis of macroeconomic and financial data. By leveraging comprehensive datasets from key financial instruments like stock indices and bonds, and employing advanced data processing technologies, we have significantly enhanced the accuracy and depth of our financial analysis. This has allowed us to gain a more granular understanding of the interrelationships between macroeconomic indicators and financial market dynamics, thus improving our ability to forecast economic conditions.

Looking towards the future, our project will focus on further enhancing the accessibility and real-time analytical capabilities of our database systems. We plan to refine the user interface and incorporate cutting-edge technologies such as natural language processing and report generating. These improvements will automate and enrich the analytical processes, potentially transforming how financial insights are derived and utilized, thereby supporting more informed and strategic decision-making in the financial sector.

8 Appendix

- Github: https://github.com/EndeavorSu/QF-5214/tree/main