CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The FOREX(foreign exchange) market faces individual challenges for predicting because of its unique characteristics such as dynamic,non-linear,and fiendishly unstable.Therefore,using traditional approaches often weakly confront capturing these problems,applying advanced computational techniques.This project implements tackling these challenges and promoting the accuracy of FOREX predictions with the use of a machine learning model,LSTM(Long Short-Term Memory) .

In this project,adopting a structured methodology to transform raw financial data into valuable and actionable insights. First of all,collecting and seeking comprehensive and high-quality datasets which include historical exchange rates and additional indicators of macroeconomic and microeconomic.Then ensuring analyzing clean,consistent and veracity data,it is a crucial procedure to beforehand process data for preparation.Conducting engineering techniques to effectively extract meaningful and valuable patterns and relationships for facilitating the predictive power of the model.

The following subsections demonstrate every step of the methodology ,which is from data source description,data Collection and analysis to model evaluation.With the use of this structured approach,the project transforms raw ,non-pattern and complex data from the FOREX market to clear,structured and valuable data for stakeholders,and offers a set of forecasting insights.

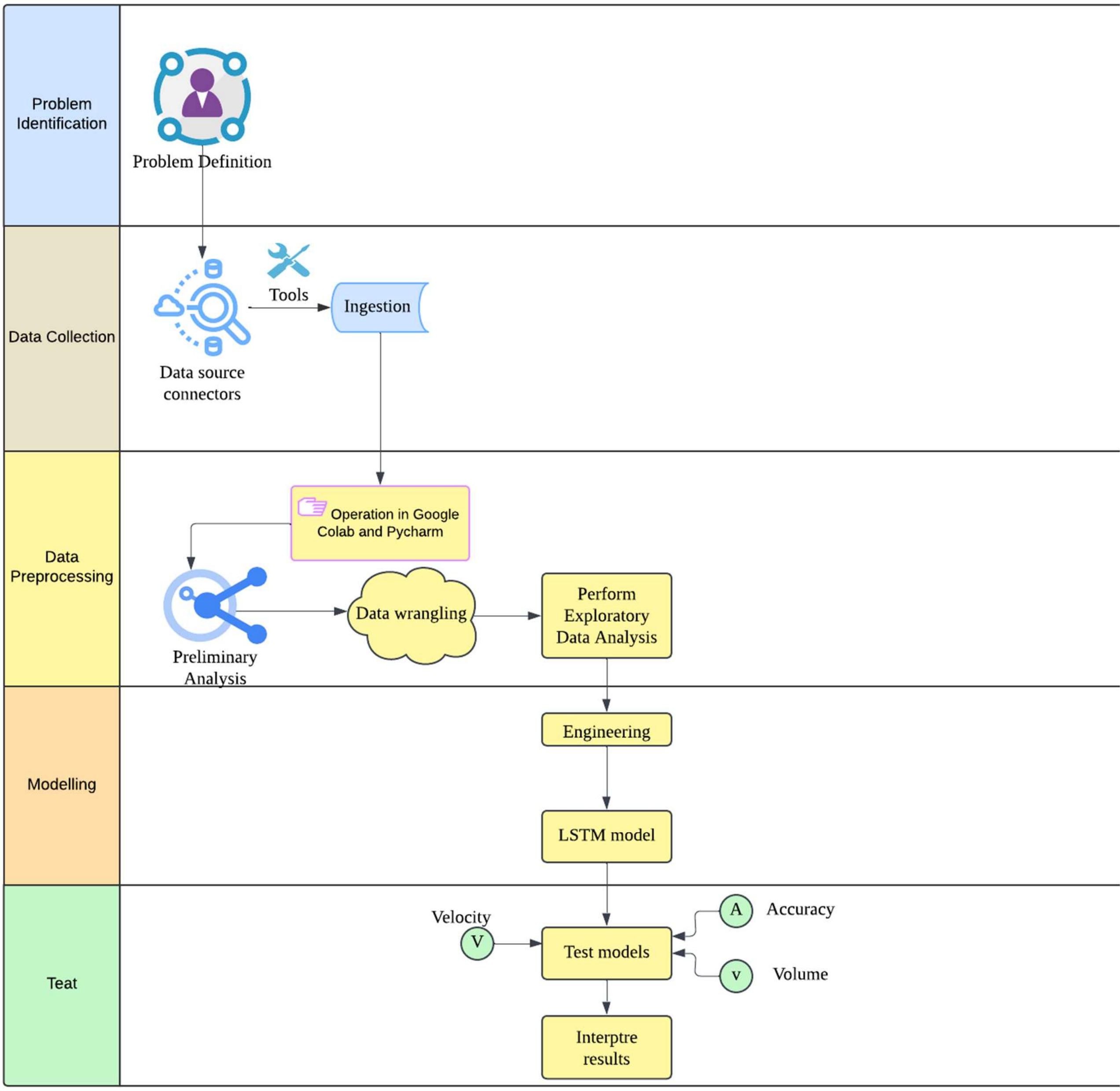


Table 3.1 Project Research Framework of FOREX prediction

3.2 Data Collection

In this project,it is used as a dataset from the website of Exchange Rates UK for pairs USD/CNY.In the meantime,also it is confirmed that dataset from the website of Alpha Vantage for pairs USD/CNY. The range of both Dataset are from 28th December 2014 to 27th December 2024.The size of the first one is 168 KB,and the second one is 112 KB.Due to improving the accuracy of the forecasting about the pairs USD/CNY,there are three additional factors which are 10-year Treasury yield,interest and inflation from US and China to need to concern.So,the dataset of 10-year Treasury yield derived from the Alpha Vantage ,and Chinese one is from People's Bank of China (PBOC).As for CPI data of US and China,they are both from IMF..Moreover,the dataset of US interest rate is from the Alpha Vantage while China one is from PBOC.It shows below Table 3.2

Datasets	Att ributes
China_10YBond.xlsx	Date: the time of the yield observation Rate: the bond interest rate
Exchange_PairsUSD&CNY_Fromexchangerates.xlsx	Date:the time of the exchange rate observation US Dollar to Chinese Yuan:the value of 1 USD in Chinese Yuan (CNY)
CPI_China.xlsx	Date:the time of the CPI Consumer Price Index:the raw CPI value for the given time era
CPI_US.xlsx	Date:the time of the CPI Consumer Price Index:the raw CPI value for the given time era
China_1YLPR.xlsx	Date:the time of the LPR observation 1 Y: interest rate for short-term loans

Table 3.2 Four factors data(10-year Treasury,interest and inflation

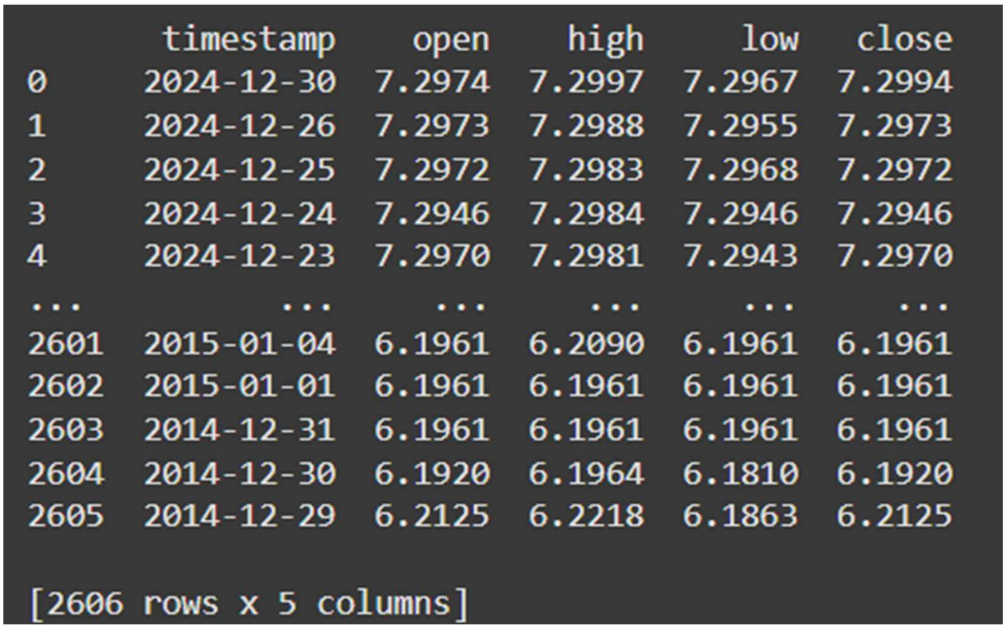
3.3 Data Pre-processing

Due to massive data ,it is quite crucial to beforehand process these various datasets,ensuring later machine learning model analysis works smoothly.It transform raw,non-pattern and chaotic data to clear,structured and comprehensible data for machine learning modeling.Below diagram illustrates the entire steps of Data Pre-processing.

3.3.1 Preliminary Analysis

Preliminary analysis plays an important role in data pipeline .Through the eight datasets which are CPI_US, CPI_China, BOP_China, BOP_US,interest_US ,interest_China,10YTreasury_China,10YTreasury_from China and US,for processing in mode stably,it is necessary to merge these eight datasets into one.Checking the information of these datasets by Figure 3.1 ,it is essential to unify the format of the date,which transforms Quart,daily to monthly.

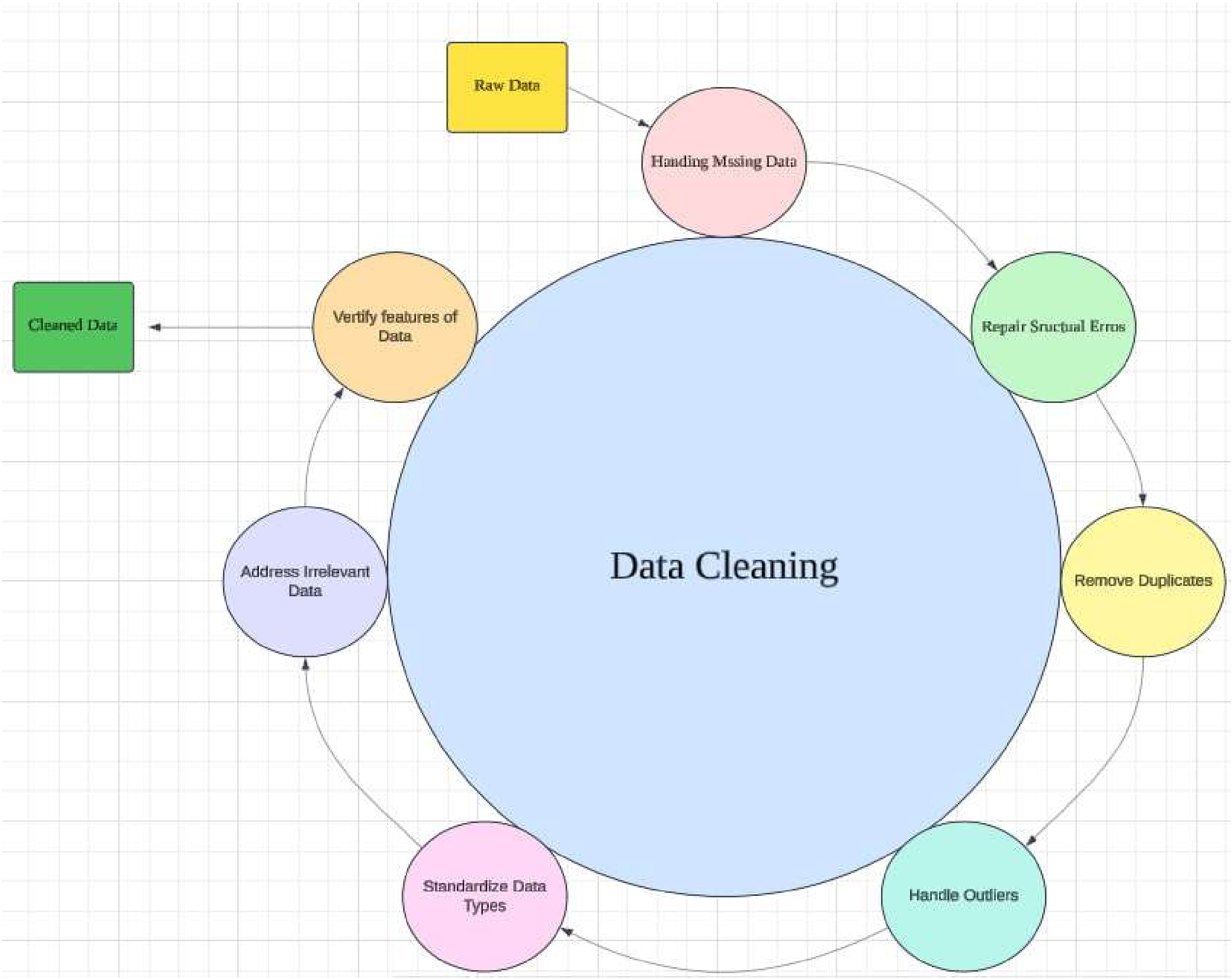
Figure 3.1 realtime exchange rate between China and US



```
       timestamp    open    high     low   close
0     2024-12-30  7.2974  7.2997  7.2967  7.2994
1     2024-12-26  7.2973  7.2988  7.2955  7.2973
2     2024-12-25  7.2972  7.2983  7.2968  7.2972
3     2024-12-24  7.2946  7.2984  7.2946  7.2946
4     2024-12-23  7.2970  7.2981  7.2943  7.2970
...          ...     ...     ...     ...     ...
2601  2015-01-04  6.1961  6.2090  6.1961  6.1961
2602  2015-01-01  6.1961  6.1961  6.1961  6.1961
2603  2014-12-31  6.1961  6.1961  6.1961  6.1961
2604  2014-12-30  6.1920  6.1964  6.1810  6.1920
2605  2014-12-29  6.2125  6.2218  6.1863  6.2125

[2606 rows x 5 columns]
```

3.3.2 Data Cleaning

Data Cleaning is a significant step of the Data science analysis that engages a set of preparation of raw data for later machine learning modeling.It significantly implements the goals which is to attain a veracity ,complete,and consistent data.In addition, it would facilitate the quality of insights and predictions.From figure 3.2,it is clearly to showcase the circle flow of the Data Cleaning.

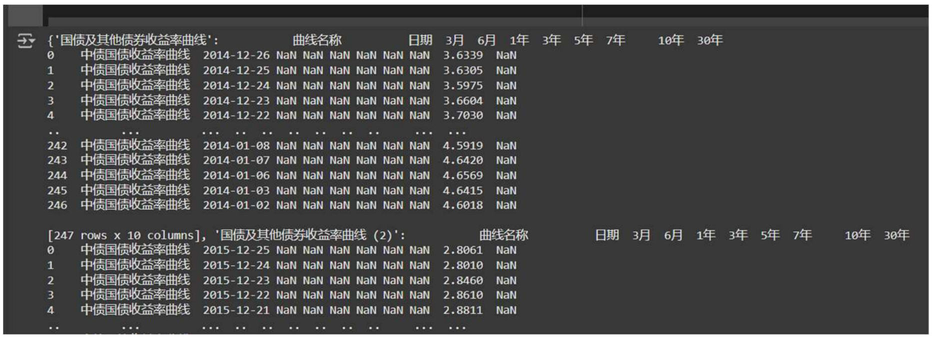


Figures 3.2 Data cleaning cycle

First of all,according to the above circle of Data Cleaning steps,dealing with the 10-year Treasury yield of the US and 10-year Treasury yield of China.Hading missing data of these two datasets is the first step.Checking Figure 3.3 ,it is straightforward to find the messorder of this dataset,China_10YBond.xlsx, missing value ,extra columns and language format.So,using pandas tool to tackle these problem for ensuring appropriate data order .After cleaning,Figure 3.4

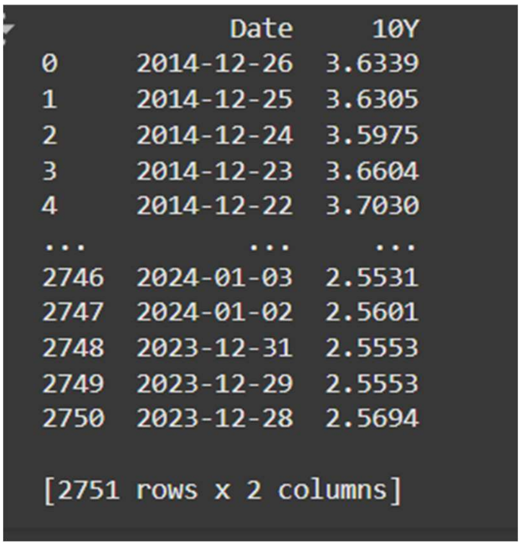
displays that the table is transformed to the valid data table.

Figure 3.3 Raw Chinese interest data



```
{'国债及其他债券收益率曲线':          曲线名称         日期  3月  6月  1年  3年  5年  7年     10年  30年
0    中债国债收益率曲线  2014-12-26 NaN NaN NaN NaN NaN NaN  3.6339  NaN
1    中债国债收益率曲线  2014-12-25 NaN NaN NaN NaN NaN NaN  3.6305  NaN
2    中债国债收益率曲线  2014-12-24 NaN NaN NaN NaN NaN NaN  3.5975  NaN
3    中债国债收益率曲线  2014-12-23 NaN NaN NaN NaN NaN NaN  3.6604  NaN
4    中债国债收益率曲线  2014-12-22 NaN NaN NaN NaN NaN NaN  3.7030  NaN
..         ...         ...  ..  ..  ..  ..  ..  ..     ...  ...
242  中债国债收益率曲线  2014-01-08 NaN NaN NaN NaN NaN NaN  4.5919  NaN
243  中债国债收益率曲线  2014-01-07 NaN NaN NaN NaN NaN NaN  4.6420  NaN
244  中债国债收益率曲线  2014-01-06 NaN NaN NaN NaN NaN NaN  4.6569  NaN
245  中债国债收益率曲线  2014-01-03 NaN NaN NaN NaN NaN NaN  4.6415  NaN
246  中债国债收益率曲线  2014-01-02 NaN NaN NaN NaN NaN NaN  4.6018  NaN

[247 rows x 10 columns], '国债及其他债券收益率曲线 (2)':          曲线名称         日期  3月  6月  1年  3年  5年  7年     10年  30年
0    中债国债收益率曲线  2015-12-25 NaN NaN NaN NaN NaN NaN  2.8061  NaN
1    中债国债收益率曲线  2015-12-24 NaN NaN NaN NaN NaN NaN  2.8010  NaN
2    中债国债收益率曲线  2015-12-23 NaN NaN NaN NaN NaN NaN  2.8460  NaN
3    中债国债收益率曲线  2015-12-22 NaN NaN NaN NaN NaN NaN  2.8610  NaN
4    中债国债收益率曲线  2015-12-21 NaN NaN NaN NaN NaN NaN  2.8811  NaN
..         ...         ...  ..  ..  ..  ..  ..  ..     ...  ...
```

Figure 3.4 cleaned Chinese interest data

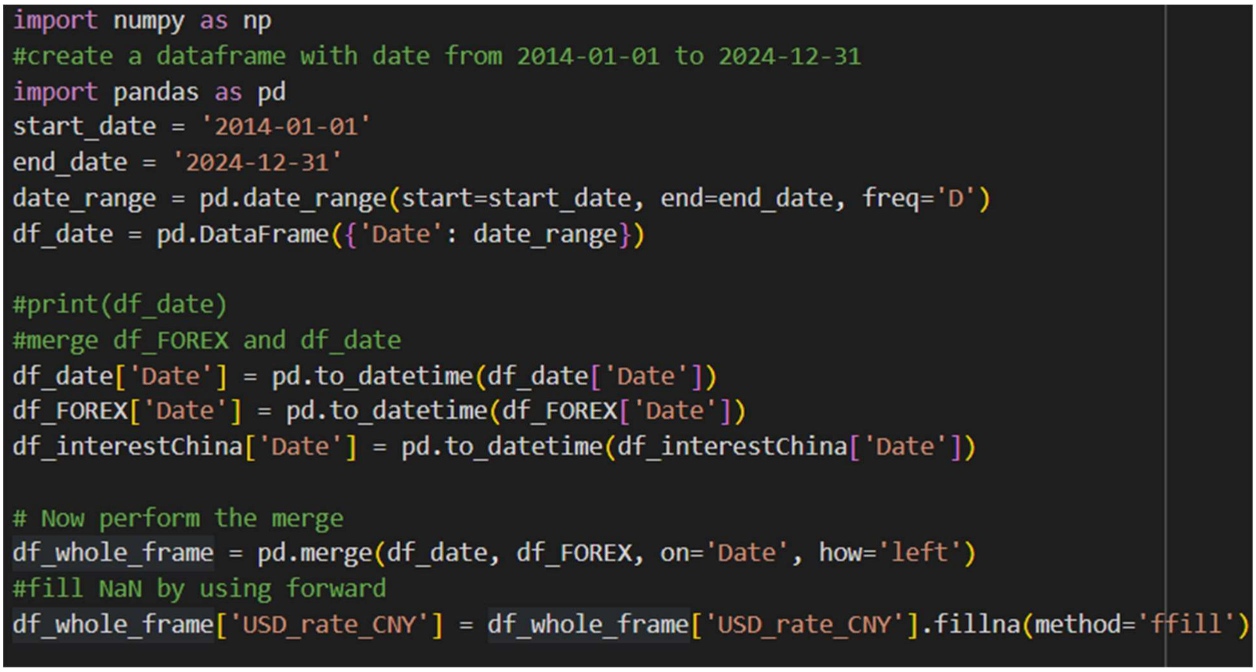


```
            Date     10Y
0     2014-12-26  3.6339
1     2014-12-25  3.6305
2     2014-12-24  3.5975
3     2014-12-23  3.6604
4     2014-12-22  3.7030
...          ...     ...
2746  2024-01-03  2.5531
2747  2024-01-02  2.5601
2748  2023-12-31  2.5553
2749  2023-12-29  2.5553
2750  2023-12-28  2.5694

[2751 rows x 2 columns]
```

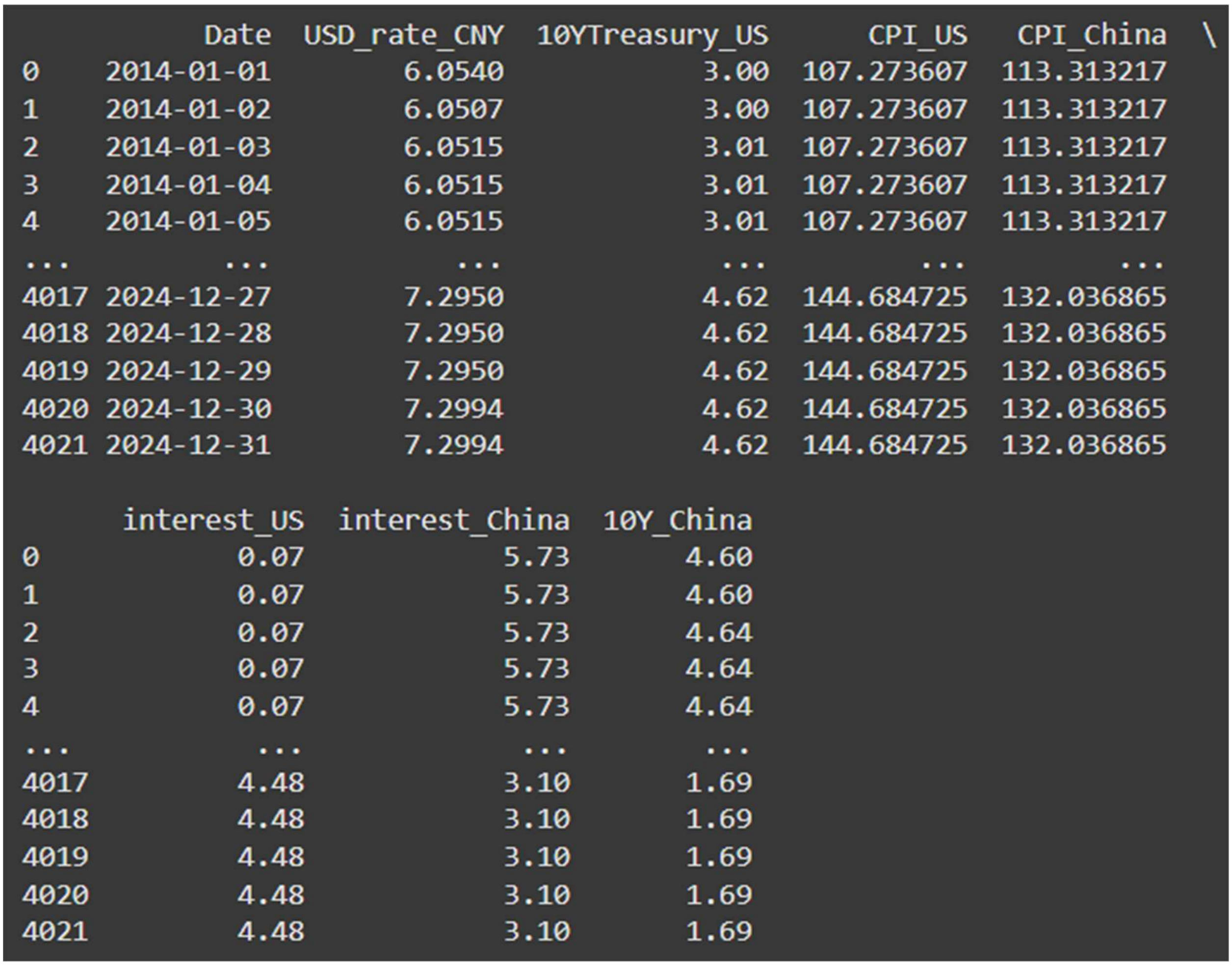
3.3.3 Data Merging

For smoothly and beforehand processing in the analysis and later modeling,it is quite necessary to integrate all key data frames into one comprehensive dataframe.So,using pd.merge() of pandas package is to combine df_FOREX,df_10Y_US, df_cpi_us,df_cpi_china,df_interest_US,df_interestChina and df_10Y_China,and then it will get a overall dataset:df_whole_frame,which is illustrated in Figure 3.5 below.

Figure 3.5 Merging among data frames



```
import numpy as np
#create a dataframe with date from 2014-01-01 to 2024-12-31
import pandas as pd
start_date = '2014-01-01'
end_date = '2024-12-31'
date_range = pd.date_range(start=start_date, end=end_date, freq='D')
df_date = pd.DataFrame({'Date': date_range})

#print(df_date)
#merge df_FOREX and df_date
df_date['Date'] = pd.to_datetime(df_date['Date'])
df_FOREX['Date'] = pd.to_datetime(df_FOREX['Date'])
df_interestChina['Date'] = pd.to_datetime(df_interestChina['Date'])

# Now perform the merge
df_whole_frame = pd.merge(df_date, df_FOREX, on='Date', how='left')
#fill NaN by using forward
df_whole_frame['USD_rate_CNY'] = df_whole_frame['USD_rate_CNY'].fillna(method='ffill')
```



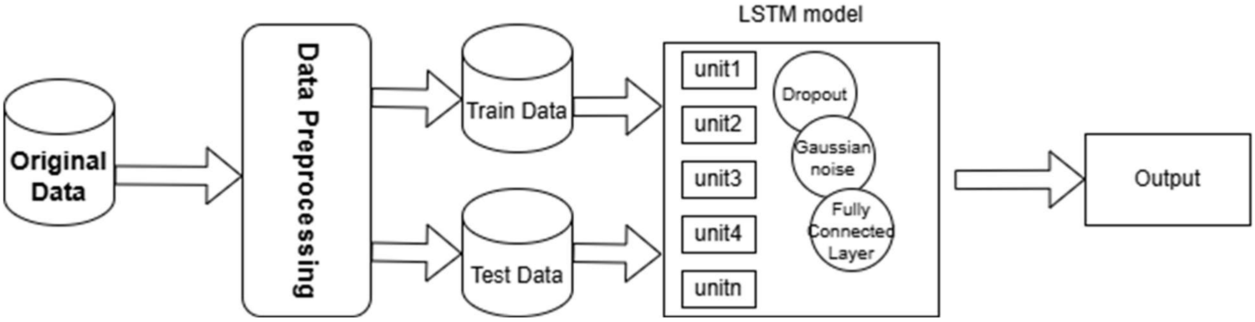
```
            Date  USD_rate_CNY  10YTreasury_US      CPI_US   CPI_China  \
0     2014-01-01        6.0540            3.00  107.273607  113.313217
1     2014-01-02        6.0507            3.00  107.273607  113.313217
2     2014-01-03        6.0515            3.01  107.273607  113.313217
3     2014-01-04        6.0515            3.01  107.273607  113.313217
4     2014-01-05        6.0515            3.01  107.273607  113.313217
...          ...           ...             ...         ...         ...
4017  2024-12-27        7.2950            4.62  144.684725  132.036865
4018  2024-12-28        7.2950            4.62  144.684725  132.036865
4019  2024-12-29        7.2950            4.62  144.684725  132.036865
4020  2024-12-30        7.2994            4.62  144.684725  132.036865
4021  2024-12-31        7.2994            4.62  144.684725  132.036865

      interest_US  interest_China  10Y_China
0            0.07            5.73       4.60
1            0.07            5.73       4.60
2            0.07            5.73       4.64
3            0.07            5.73       4.64
4            0.07            5.73       4.64
...           ...             ...        ...
4017         4.48            3.10       1.69
4018         4.48            3.10       1.69
4019         4.48            3.10       1.69
4020         4.48            3.10       1.69
4021         4.48            3.10       1.69
```

3.4 Data Modelling

In this project, using the LSTM model as a major analysis tool. Long Short-Term Memory (LSTM) networks is a special form of the RNN (Recurrent Neural Network) architecture,which is for facilitating the performance of a neural network with past data inputting and solving some traditional RNNS problems such as vanishing gradient.From the Figure 3.4.1,it is understandable to figure out the flow of LSTM working when integrating LSTM into Data science life circle.Firstly,preparing dataset with the characteristic of time series correspond the structure of LSTM.During the Data Preprocessing,it is significant to normalise the data under the standard of LSTM.When the Data enter into model circle,LSTM will work stably and appropriately with these processed data.

Through a set of crucial steps of the Data Preprocessing,splitting the processed data to two same datasets for the accurate forecation of LSTM.One named Train Data and another named Test Data .Within the processing of LSTM model,it will implement three important phrases:Dropout ,Gaussian noise and Fully Connected Layer.During th training ,Dropout plays a role in preventing overfitting by memorising long-term dependencies.After the dropout phrase,the data will go through the Gaussian Noise ,which ensures that the inputs makes the LSTM model more steadfast to noisy.After passing the Gaussian Noise phase,Fully Connected Layer will be the final one of the circle.And it will summarize the outputs from the LSTM to offer the expected values.

Figure 3.6 LSTM working flew


LSTM model
Original Data
Data Preprocessing
Train Data
Test Data
unit1
unit2
unit3
unit4
unitn
Dropout
Gaussian noise
Fully Connected Layer
Output