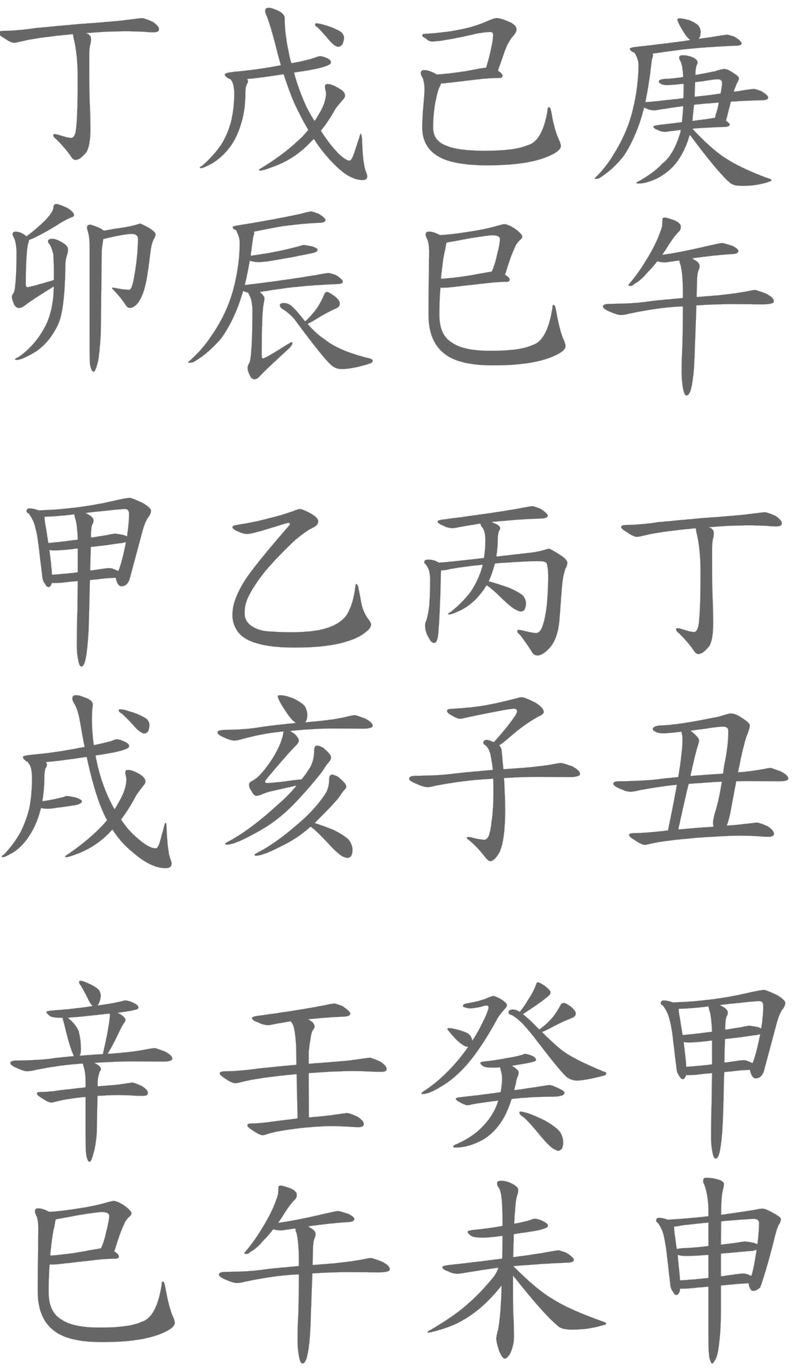
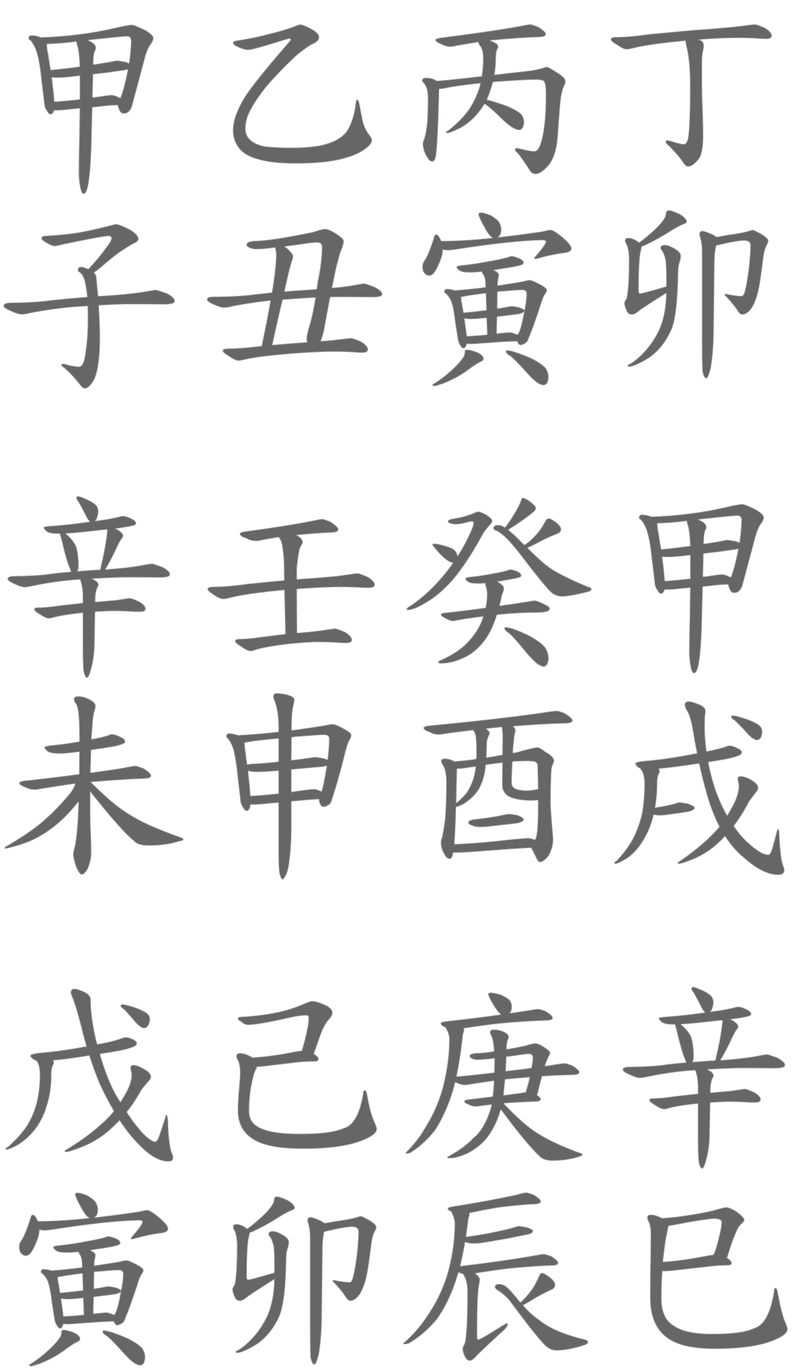


The Data Science Project by PingChi Tsai



Outline

- Executive Summary
- Introduction
- Methodology
- Data Wrangling
- Data Visualization
- Conclusion
- Appendix

Project:

Does Chinese traditional calendar System has any relationship with stock market?

Executive Summary

Summary of Methodology

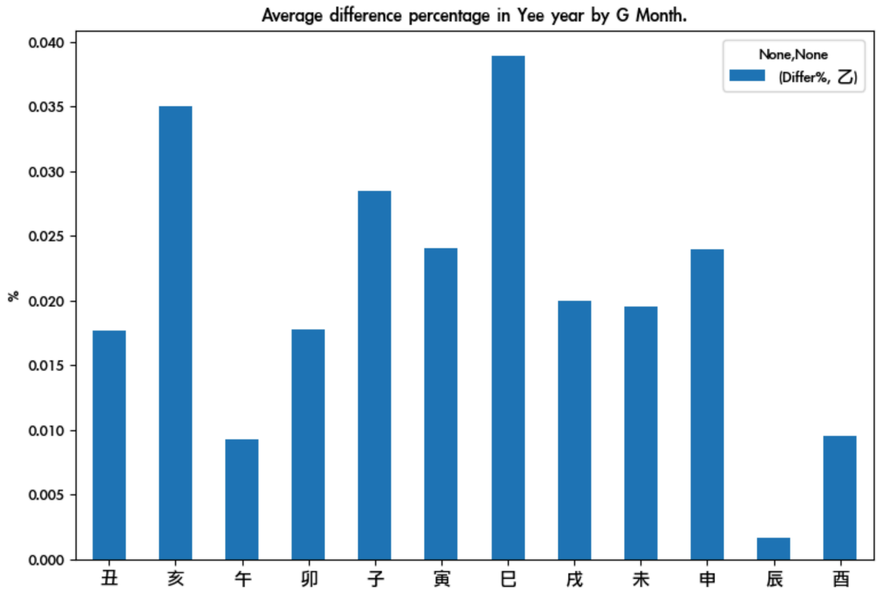
I collected data of monthly closing prices of the Dow Jones Industrial Average from June 1896 to March 2025 from S&P Global. The data were converted into a traditional Chinese calendar system, which was believed associated to nature environment and human behavior.

I used this data separated them into different groups, and use bar plots, box plots, scatter plots to analyze them to find the relationship between these data.

Summary of Result

This project indeed found in some calendar combinations, the risks were lower than other combinations. Specifically, in the 天干(Tian Gan) system the year of 乙(yee) had a better result than others. And the 天干 calendar system is the one recommended as a tool to build an investment plan comparing with the 地支(Di Zi) calendar system.

For the calendar system combined with 天干 and 地支, check out the appendix part. I used K-means and HDBSCAN clustering to find the combining system.



Introduction

The Gregorian calendar currently used worldwide is a calendar based on the revolution of the Earth around the Sun. The traditional Chinese calendar is based on the cycles of the moon and the solar cycle is incorporated into the calendar by adding the changes in solar seasons. For centuries the Chinese believed that this set of celestial rules had an impact on humanity.

This project is based on this calendar and studies the Dow Jones Industrial Average over more than 100 years to determine whether there is a certain correlation between this calendar statistics and the investors' confidence in the financial market.

This calendar is not composed of numbers but two counting elements: 天干(Tian Gan) and 地支(Di Zi). There are 10 天干 and 12 地支. Basically, the calendar counts are composed of 1 天干 and 1 地支, so there are 60 combinations in total.

Nowadays, people pay more attention to 地支, because the 12 months and 12 zodiac years can be matched with 地支, while 天干 are more like counting the 60-year combination cycle. This project refers to various information. In order to make it easier to correspond with current trading dates and trading behaviors, the calendar system in this project is used to count the year starting from 1864 as the year of 甲子, and January 1st of the Western calendar is the year of change, and November is the first month which is 子. In this method more attention is paid to the impact of the 天干地支 combination in a specific year.

The DJI just has 130 years history that is not enough for a 60-year combination cycle to have a solid conclusion. I separated two counting system to count the year. Here are the two counting system below:

天干(Tian Gan):

甲Gia, 乙Yee, 丙Bin, 丁Din, 戊Woo, 己Gee, 庚Gng, 辛Xin, 壬Ren,

癸Gue

地支(Di Zi):

子Zee, 丑Cou, 寅Yin, 卯Mao, 辰Cho, 巳See, 午Wuu, 未Way, 申San, 酉You, 戌Shu, 亥Hai

For the calendar system combined with 天干 and 地支, check out the appendix part. I used K-means and HDBSCAN clustering to find the combining system.

Methodology

The monthly closing prices of the Dow Jones Industrial Average from June 1896 to March 2025 were obtained from S&P Global, and the data were converted into 天干 Year, 地支 Year, and 地支 Month.

Calculate the price difference growth percentage for each month by dividing the closing price of a certain month by the closing price of the previous month, and add a new column to record the data. These growth percentage data are divided into different groups according to different needs:

1. Price difference percentage for each month
2. Average price difference percentage per month in each 天干 Year
3. Average price difference percentage per month in each 地支 Year
4. Average price difference percentage per month for a specific 天干 year
5. Average price difference percentage per month for a specific 地支 year

Each group was analyzed using a bar graph or a box plot.

Data Wrangling

Get the monthly closing data xls file from S&P Global from 1896 to 2025, use Pandas to read and add columns for each month's index increase or decrease and the percentage column for each month's index increase or decrease.

Use astropy.time to convert the original time format, and define the 天干 and 地支 conversion program to create columns corresponding to the date. In the end, I only leave the required columns and data.

```
df=pd.read_excel(path)
df.head()
```

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	As of:	20250331	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Index Name	Effective Date	Close Value	TR Close Value	Net Change	Daily Volume	Open Value	Intraday High	Intraday Low	Theoretical Open	Theoretical High	Theoretical Low

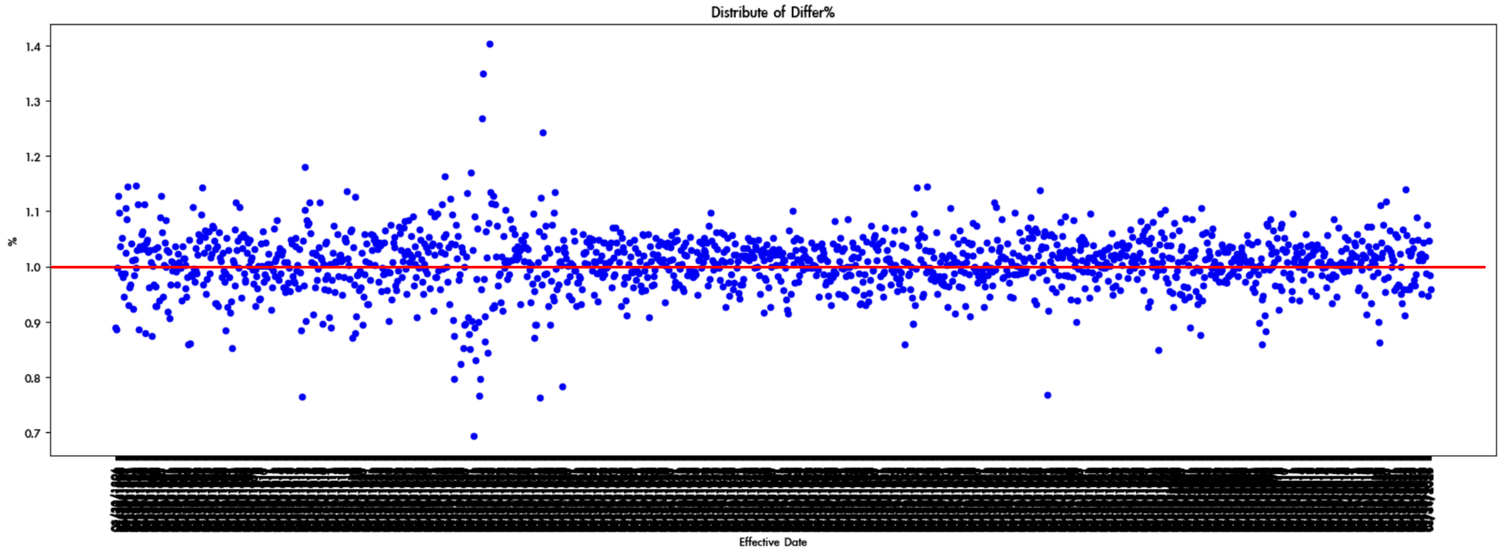
	Effective Date	Close Value	Difference	Differ%	gan Year	G Year	G Month
1522	08/31/2023	34721.91	-837.62	0.976	癸	卯	酉
1523	09/29/2023	33507.5	-1214.41	0.965	癸	卯	戌
1524	10/31/2023	33052.87	-454.63	0.986	癸	卯	亥
1525	11/30/2023	35950.89	2898.02	1.088	癸	卯	子
1526	12/29/2023	37689.54	1738.65	1.048	癸	卯	丑
1527	01/31/2024	38150.3	460.76	1.012	甲	辰	寅
1528	02/29/2024	38996.39	846.09	1.022	甲	辰	卯
1529	03/28/2024	39807.37	810.98	1.021	甲	辰	辰
1530	04/30/2024	37815.92	-1991.45	0.950	甲	辰	巳
1531	05/31/2024	38686.32	870.40	1.023	甲	辰	午
1532	06/28/2024	39118.86	432.54	1.011	甲	辰	未
1533	07/31/2024	40842.79	1723.93	1.044	甲	辰	申
1534	08/30/2024	41563.08	720.29	1.018	甲	辰	酉
1535	09/30/2024	42330.15	767.07	1.018	甲	辰	戌
1536	10/31/2024	41763.46	-566.69	0.987	甲	辰	亥
1537	11/29/2024	44910.65	3147.19	1.075	甲	辰	子
1538	12/31/2024	42544.22	-2366.43	0.947	甲	辰	丑
1539	01/31/2025	44544.66	2000.44	1.047	乙	巳	寅
1540	02/28/2025	43840.91	-703.75	0.984	乙	巳	卯
1541	03/31/2025	42001.76	-1839.15	0.958	乙	巳	辰

Data Visualization

At first, I used the original Gregorian calendar time and the Differ% column, which is the percentage of the closing price of the month and the last month, to visualize the scatter plot.

From the graph, we can see that except for the great fluctuations during the Great Depression in the 1930s, the data for other periods are distributed almost in the range of 1.2%-0.8%.

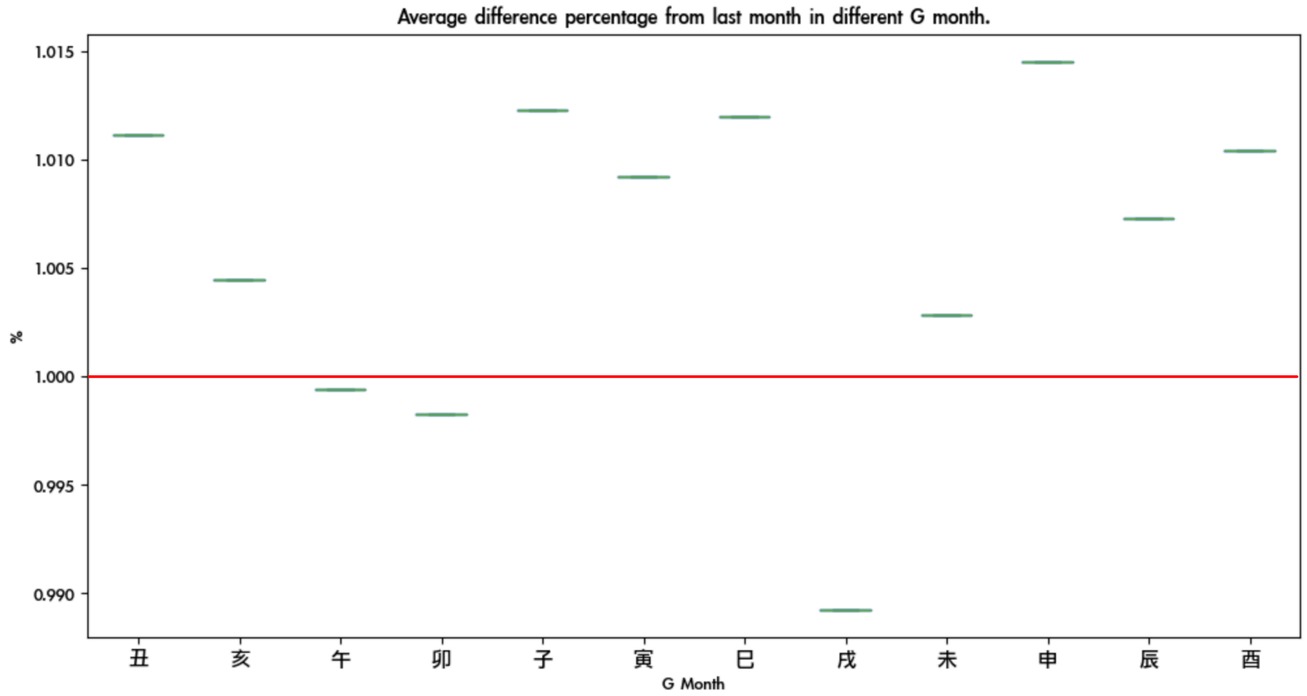
With a cycle of 10 or 12 years, the extreme values in the 1930s have little impact on nearly 130 years of data. The spirit of this project should not avoid such extreme values also, so all calculations later on include all closing price data.



Data Visualization

Next, check whether there are any differences in the performance of each month. The value of the Differ% column is composed of the average Differ% of each month. It can be seen from the table that the return rate in most months is positive, and the average return in the three months of 戌 (September), 午 (May) and 卯 (February) is negative.

Among them, the return rate in the month of 戌 (September) is particularly poor, and this result is more obvious when looking at the box plot.

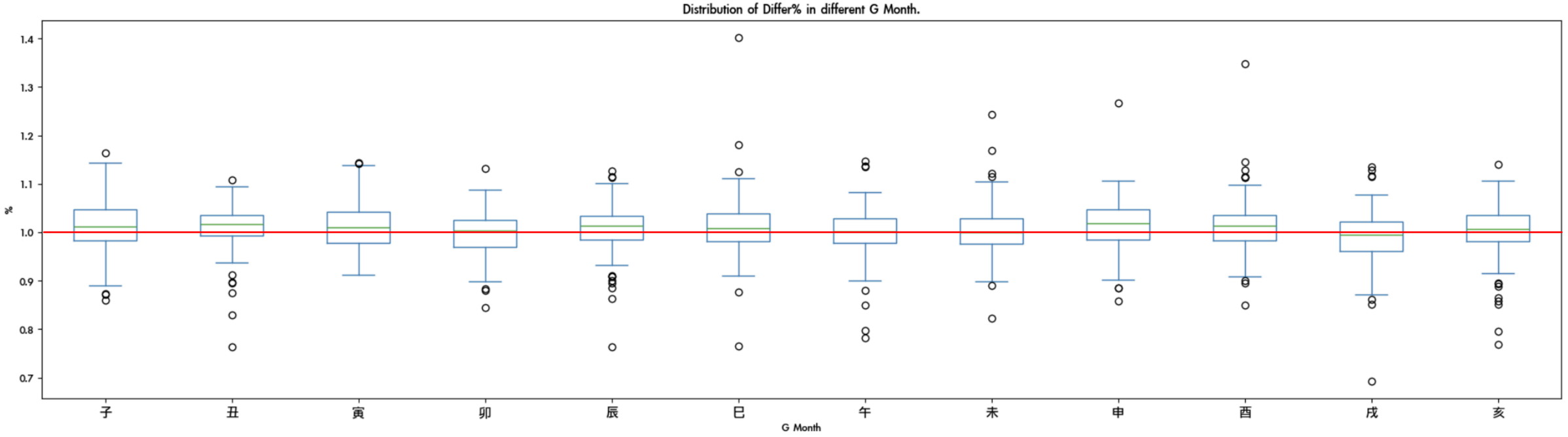


G Month	Differ%
丑	1.011171
亥	1.004469
午	0.999422
卯	0.998248
子	1.012273
寅	1.009209
巳	1.012008
戌	0.989242
未	1.002814
申	1.014527
辰	1.007295
酉	1.010445

Data Visualization

By using box plot to plot the distribution of 1542 data in each month, we can find that there are outliers in every month. The month of 戌 has the outlier with the lowest return rate, but it is not really much lower than other months, and the first quartile and median of the month of 戌 are indeed lower than other months.

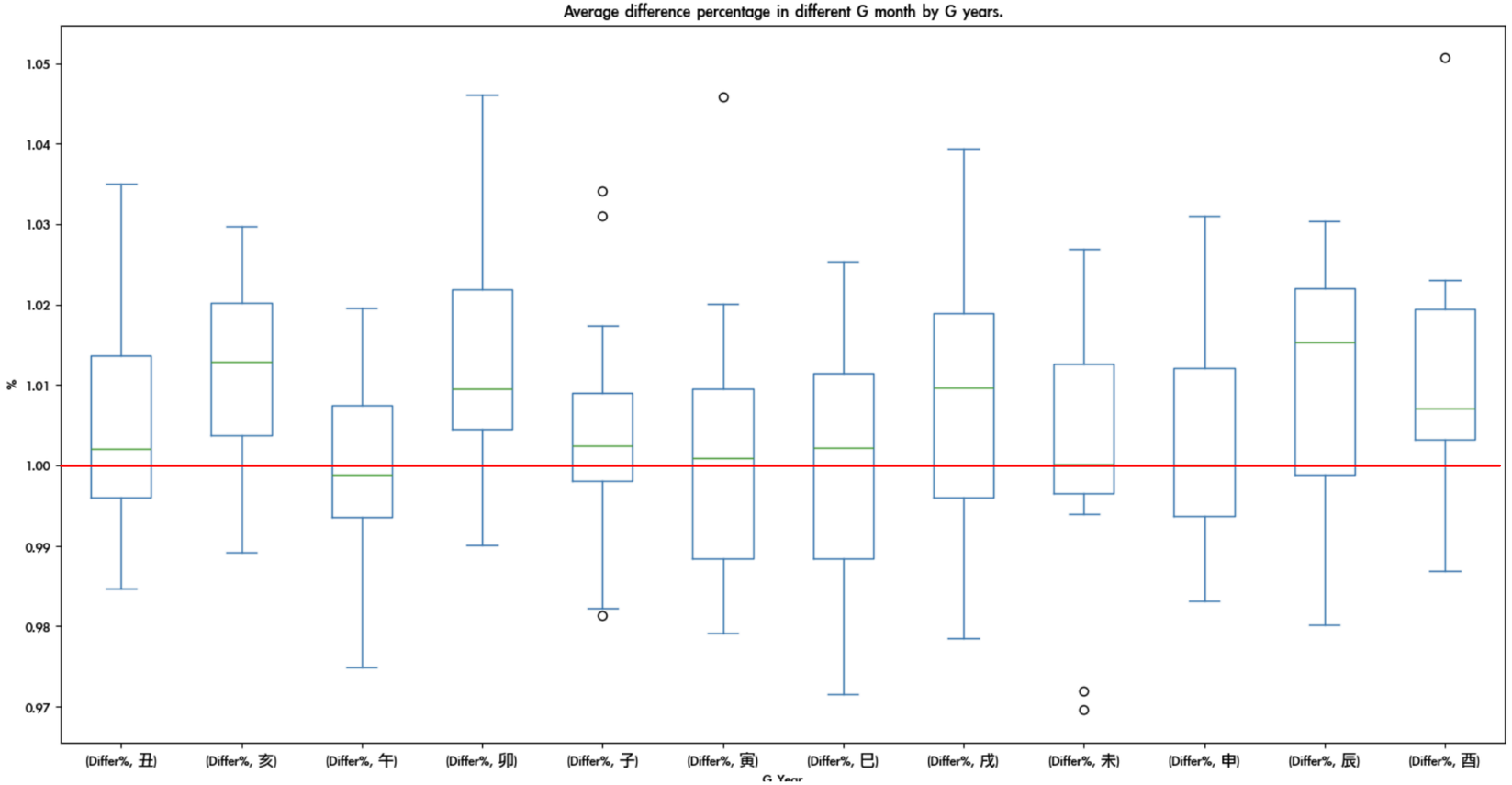
Compared with the 戌 month, the 亥 month and the 辰 month have more outliers with negative returns, but the average return rate is still positive, so it is correct to mention earlier that the extreme data of the Great Depression in the 1930s does not affect the overall assessment.



Data Visualization

I added the 地支 year to the analysis, used groupby() function to average the monthly returns within each 地支 year, and then obtained this box plot. It can be found that the average monthly returns in the years of 卯, 酉 and 亥 are mainly positive, which means that it's safer to invest in these years whenever the time is.The median of year 午 is negative, and it means it's a year to trim long positions.

In the next, I will separate each month of the year to see if there are more details.



Data Visualization

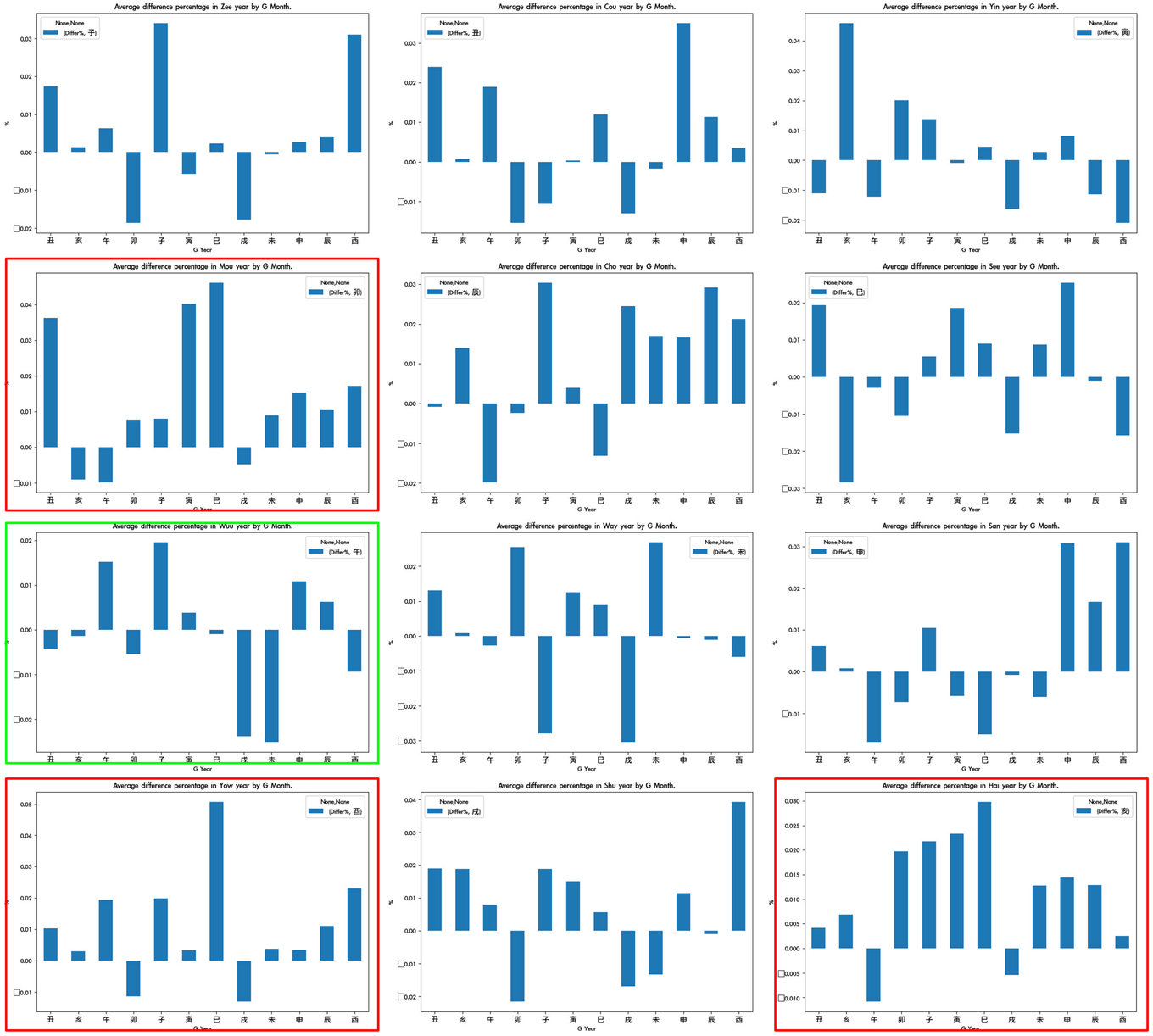
DJI各地支年各月平均成長百分比與一之差

To make it easier to see positive and negative returns, I subtract 1 from the return rate.

The years in the red box are the 卯, 酉 and 亥 years. The year 卯 and the year 酉 basically rely on certain months to raise the rate of return of that year. The year 亥 is the type where small amounts add up to a lot.

The green box is the year 午. The monthly return rates in this year are indeed lower than that in other years.

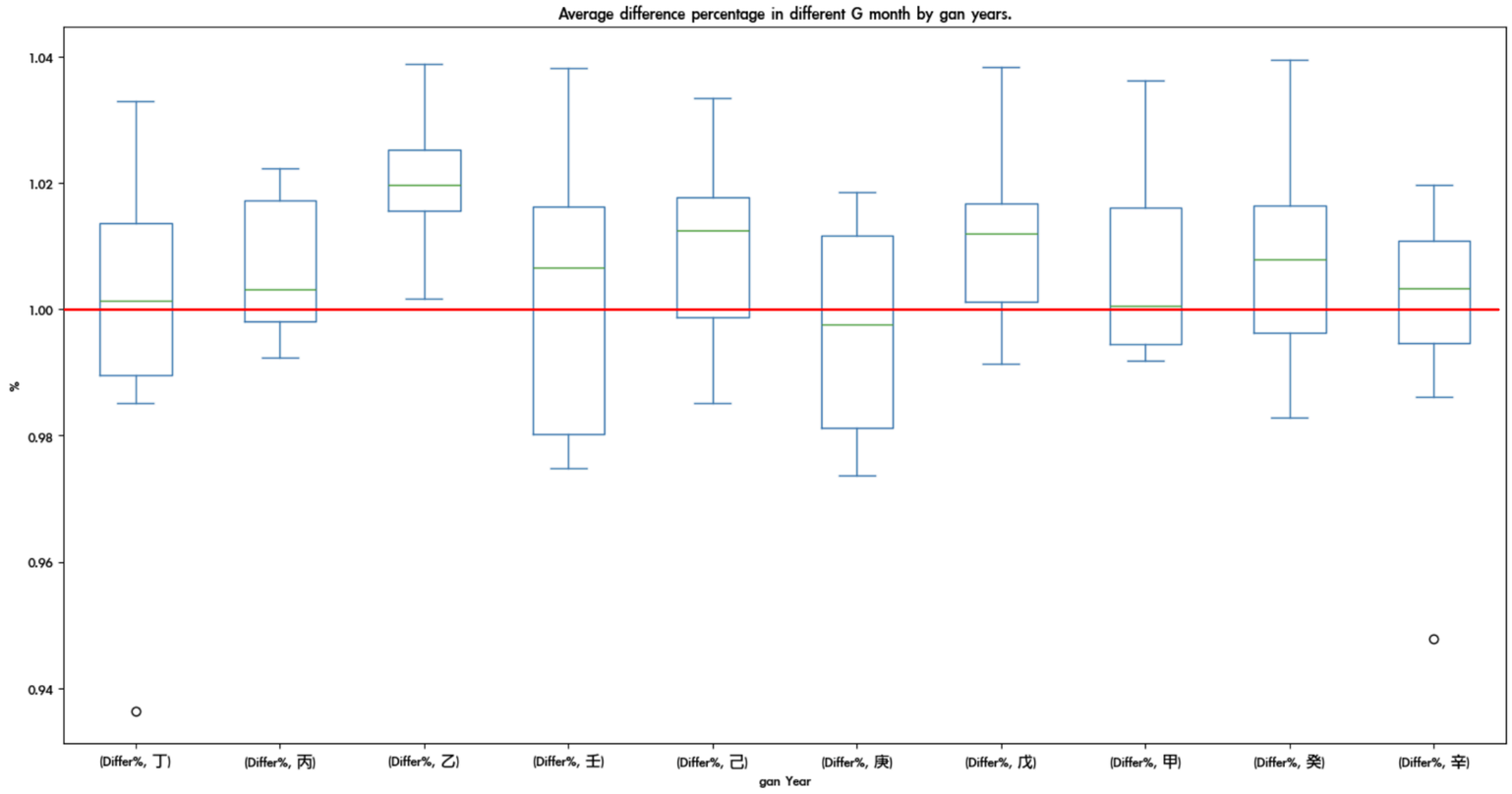
In addition, from the bar chart of each year, we can see that the rate of return in the month of 戌 (September) is indeed not good.



Data Visualization

Switching to the 天干 calendar system to look at the data, I also used the average monthly returns for each 天干 year to draw this box plot.

It can be found that the rates of return in Year 乙 is positive and Year 戊 is better than other except Year 乙. The median of the year 庚 is negative.



Data Visualization

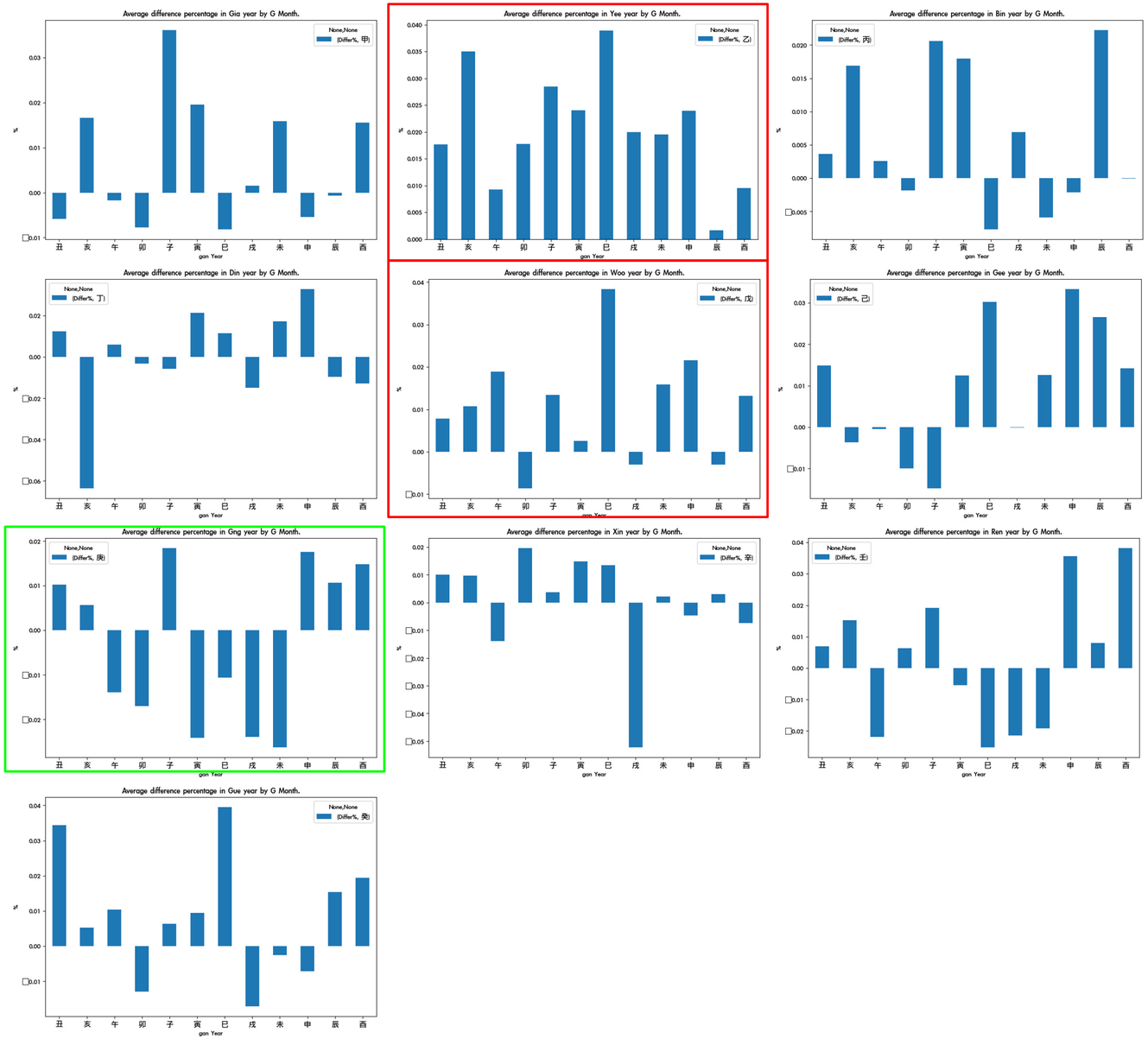
DJI各天干年各月平均成長百分比與一之差

To make it easier to see positive and negative returns, I subtract 1 from the return rate.

The ones in the red box are the years of 乙 and 戊. It can be found that the return rates in each month of the year 乙 are positive, even in the month of 戊 (September).

In the year of 戊, the better return rate for the whole year was maintained by low negative returns in just a few months and a higher returns in the month of 巳 (April).

The one in the green box is the year 庚. The year 庚 has more negative return months, and even in the positive return months, the return rates are still not high.



Conclusion

The research of this project did not find strong evidence that a particular combination of this Chinese calendar system has absolutely correct relationships to the ups and downs in share market, but it did find higher or lower rates of return in certain year-month combinations. The following are some of the key findings:

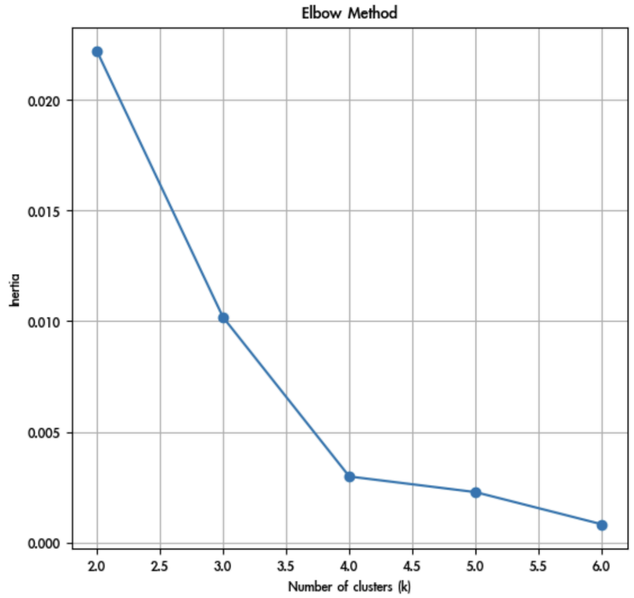
- The calendar system of 天干 and 地支 can indeed find combinations with relatively good returns.
- The return rates in September were relatively poor.
- In the 地支 year system, the average monthly return rates in the years of 卯, 酉 and 亥 were higher than in other years. They generally had fewer months with negative average returns. In addition, the worse return rates in September were more obvious under the 地支 year system.
- In the 天干 year system, the average monthly return rates in the Year of 乙 were all positive. The year of 戊 has relatively fewer and lower retreat in months and their average returns. In the year 庚, half of the months had negative average monthly return rates . The bad return rate in September is not obvious under the 天干 year system.
- The outliers in the 天干 year system were fewer than those in the 地支 year system, and the rate of return of each month in the 地支 year system is more scattered than that in the 天干 year system. The 天干 year system would be a better tool for building an investment plan.

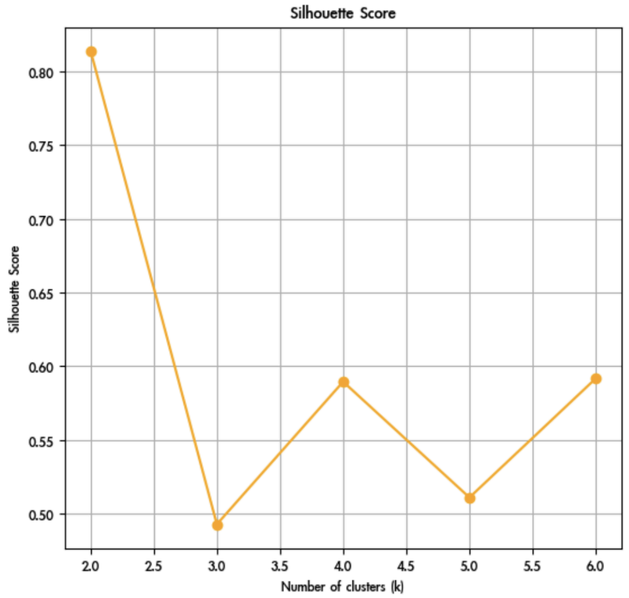
Appendix

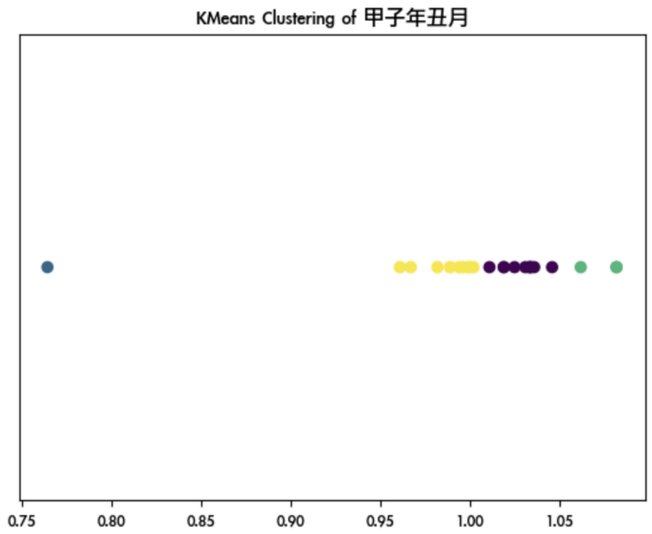
It takes 60 years for the 天干地支 calendar system to complete a cycle, but stock market data does not extend long enough to be analyzed. My idea was to carefully include the characteristics of the 天干 year system and 地支 year system, and applied machine learning packages to complete this task. Here were my steps:

1. Sort out all the price data for the year 甲 and the year 子.
2. Separate each month from the two data frames.
3. Use K-means or HDBSCAN to select more suitable clusters.
4. Box plot and bar chart of the months in the year 甲子.

This was the cluster using K-means of the month 丑 in the below.

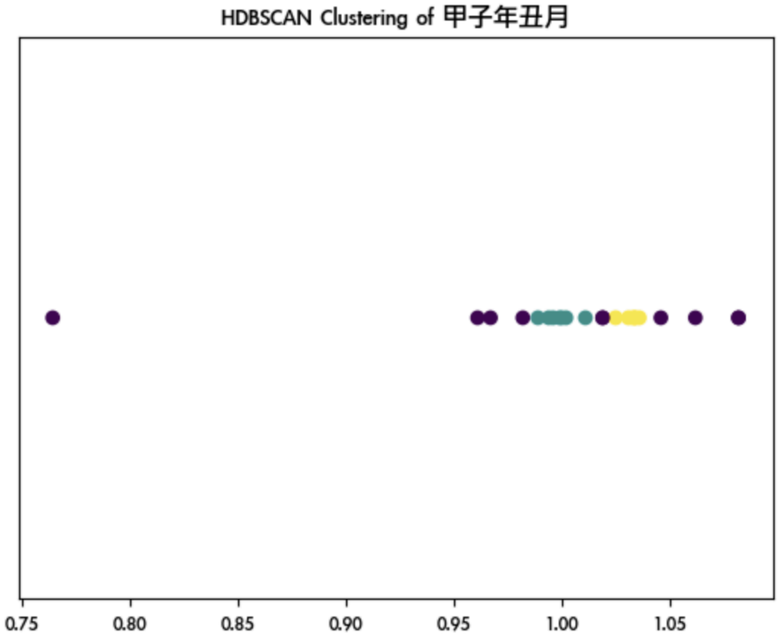


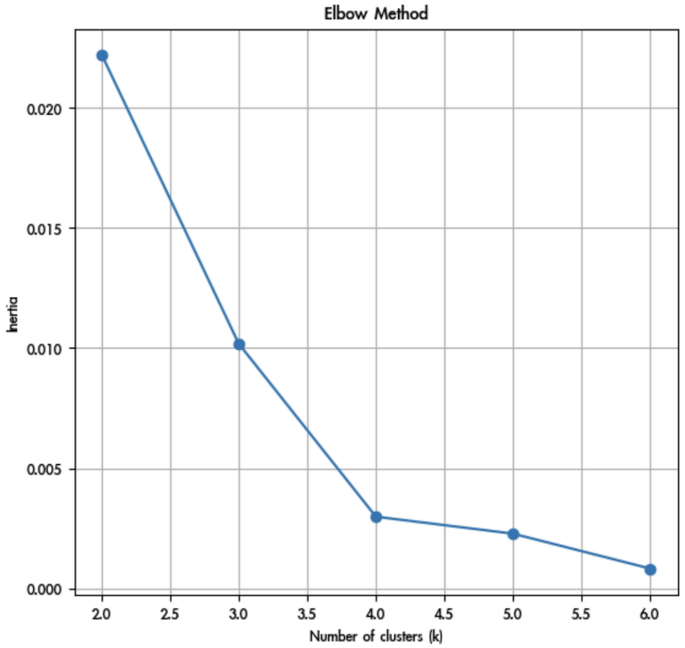


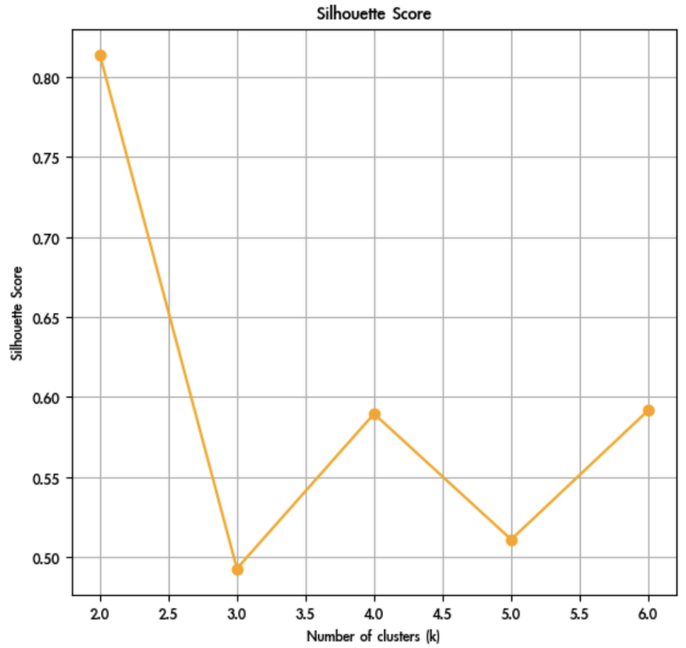


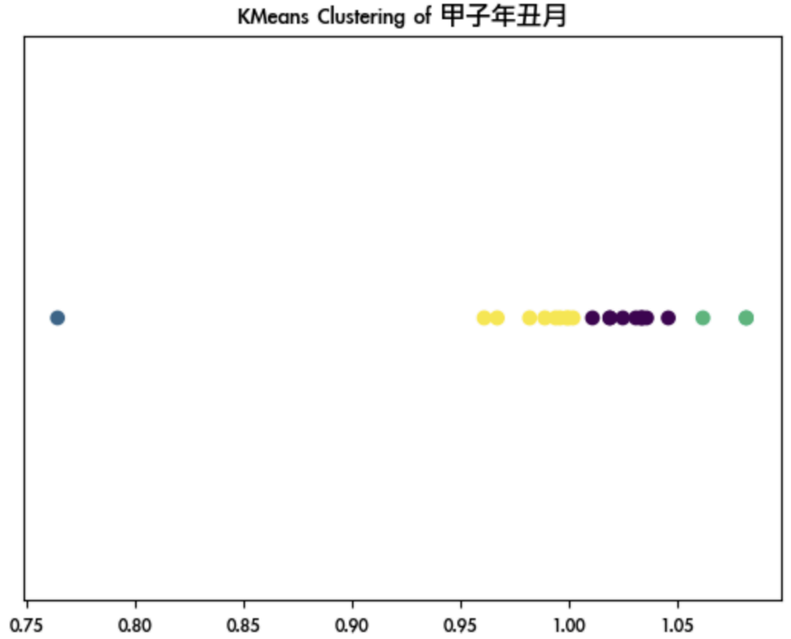
Appendix

It was sometimes difficult to achieve ideal clustering using K-means. Although HDBSCAN was not perfect either, it was better than K-means most of the time. So I used HDBSCAN in most months, and I tried to increase the min_cluster_size of HDBSCAN and filter out the non-clustered noise (the purple dots in the picture on right up side). For K-means, try to choose the number of groups that matches the Silhouette Score and Elbow Method.









Appendix

Finally, the cluster data was plotted as a box plot. Took the average of each month to get the bar plot. We can roughly observe the monthly return rate range for the 甲子 year and estimate the return rate for that year.

For the whole machine learning process check the url below

https://github.com/BinciTsai/DJI-project/blob/23c612f268c0dda0f21f72cdf3a507bfe64d208b/project1%20ML.ipynb