Final Report

Group 7: Julie Chea, Tianyu Gao

Table of Contents

Context of consultancy & Problem Statement

The aim of this consultancy project is to provide an understanding three of the most popular market index funds: the S&P500, Dow Jones, and FT100. Harnessing the power of Tableau and R, this project will include forecasting of future stock prices, as well as an observation of the relationship between the prices and returns of these indices with commonly associated economic variables such as unemployment rate, CPI and interest rates. We will also explore the effect of seasonality on these returns. Overall, this project will compare and contract these indices as well as explore their relationship with other factors. It might also help establish for beginner investors which index is best to unilaterally invest in given specific consideration and constraints. This consultancy is not aimed mostly at professional investors although it may certainly be helpful in giving them an overview of the market and their interconnections with economic factors, but at people trying to better understand these indices as they enter the finance field.
The background of this project is steeped in both financial and economic analysis as we attempt to bridge the two disciplines and reveal both the relationship between the macroeconomic data and the indices. In terms of the context, this project will be borrowing financial knowledge which allows us to calculate returns with data on closing price, as well as an understanding of standard deviations and how different stocks might better appeal to people depending on their risk profiles and return expectations because rather than any one of these stocks being the best, they are all best suited to certain needs. It is important to navigate the stock market within a broader understanding of the economic environment.

The focus of the Consultancy

Our consultancy is dedicated to help investors to better understand the trend of the index market and how will external macro effects have an impact on the indices. The dataset includes three major indices, namely, the Dow Jones Index, The Financial Times Stock Exchange 100 Index, and the Standard & Poor's 500 Index, which track the performance of large companies listed in the US. And we chose Unemployment Rate, CPI, and GDP to represent the macro factors that might have an external impact on the whole market which would finally change the prices of indices. Our client base is not limited to investors however, our visualizations can be easily viewed and understood by people who are relatively new to finance and willing to learn more, as well as wanting to compare which index looks best in terms of returns and how impacted it is by seasonality amongst the three as many people who are purely looking to combat inflations will put all of their money into just one index fund. It can also be useful for economists who make policies to see how variations in macroeconomic factors affect stock prices.

Data Quality Assurance

In terms of overall initial data quality, each dataset that was used is individually valid and credible, although there is work, we needed to do to make sure that the datasets were consistent with each other. Overall, the initial datasets can be said to have integrity as well as timelessness. Although

there are some missing values in each dataset which were not easy to notice, the datasets were relatively complete with little null values.

We can therefore consider that the initial data quality is relatively high. To make the data more useful for given purposes, we also decided to use R to categorize the data (for example categorizing returns into positive or negative).

After drawing the scatter of different variables, we found some influential points which are far away from the center of the data and have huge residuals, and we deleted these observations from our dataset. Also, because the data is from authoritative databases, the data is of high correctness and cleanliness. Overall, our data comes from three main datasets: any data about indices comes from Yahoo Finance and our economic data from FRED and the OECD website.

The final data is of high quality. It is complete, comes from credible sources, is timeless and consistent. It has multiple columns that are of value and can all be used to make interesting observations through the visualization process. From our initial statistical evaluations and descriptive analysis, it seems that there will be a lot of interesting relationships to observe which could provide value to different client bases. We believe the macro factor factors we chose are relative, for the CPI, Unemployment Rate, and Interest Rate can influence the prices of the market (Demir et al., 2017).

Data compilation and joining

Compilation was done through excel and Tableau. The initial market indices data was compiled together with unemployment and CPI data from FRED using excel. Data was downloaded for the last ten years, and copy pasted on one same sheet on Excel. The data for short and long-term interest rates were joined with the dataset through Tableau. In order to observe seasonality, we have also created an excel sheet where we created a column to assign months and joined it through Tableau PREP. Finally, we get our final dataset, as shown as "Output 3" in exhibit 1 below.

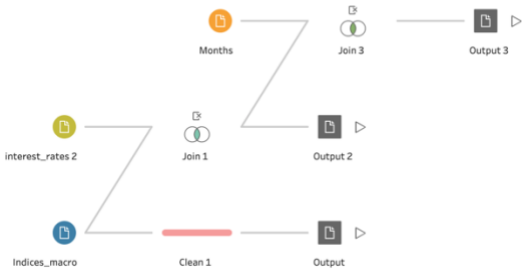

Exhibit 1. Tableau Capture of Joining Process

Data overview

In order to conduct our research, we chose three indices from Yahoo finance, we collected their opening price, closing price, the difference between the highest price and lowest price, and the trading volume in our dataset. Also, we collected some macro indicators to figure out how much the external conditions would impact the prices of the indices. The dataset is biweekly, and it starts on 12/01/2012, ends on 10/01/2022, it is a panel data containing 30 variables.

In the summary table below, we can see that the macro factors are relatively smaller compared to the prices of the indices. To standardize this gap, we used min—max standardization when doing some analysis in the report. Also, we can see the range of the prices is quite similar. We have categorized returns into positive and negative monthly returns. There are a total of 238 observations in the dataset, and most of the attributes contain no empty values, except for the DJI_Excess, GSPC_Excess, FTSE_Excess, CPI, and Unemployment where there are several empty values. Because the empty values concentrate on the beginning and the end, we directly deleted these observations when conducting feature engineering, and omitted them when drawing descriptive graphs.

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
Date	238	NA	NA	2012/12/01	2022/10/01
GSPC_Volume	238	7.866e+10	1.485e+10	5.813e+10	1.618e+11
DJI_Volume	238	5.212e+09	2.768e+09	1.483e+09	1.552e+10
FTSE_Volume	238	1.647e+10	3.376e+09	1.007e+10	3.741e+10
GSPC_Excess	236	0.00846	0.0423	-0.134	0.119
DJI_Excess	236	0.00905	0.217	-0.649	1.038
FTSE_Excess	237	0.000779	0.0375	-0.185	0.116
CPI	232	2.254	2.035	-0.200	9.060
Unemployment	236	5.333	1.865	3.500	14.70
ltinterest_perannum	238	0.0212	0.00661	0.00620	0.0398
stinterest_perannum	238	0.00890	0.00917	0.000900	0.0385
GSPC_Open	238	2,698	872.1	1,416	4,778
GSPC_High	238	2,790	911.7	1,448	4,819
GSPC_HighLow_price	238	183.8	150.5	45.93	944.9
GSPC_HighLow_percent	238	0.0642	0.0414	0.0199	0.318
GSPC_Close	238	2,715	867.2	1,426	4,766
DJI_High	238	23,835	6,758	13,366	36,953
DJI_Low	238	22,275	6,142	12,884	34,690
DJI_HighLow_proportion	238	1,560	1,220	304.4	8,889
DJI_HighLow_percent	238	0.0647	0.0425	0.0182	0.347
DJI_Close	238	23,184	6,439	13,104	36,338
FTS_Open	238	6,852	516.4	5,577	7,749
FTSE_High	238	7,052	483.5	5,997	7,904
FTSE_Low	238	6,641	551.8	4,899	7,541
FTSE_HighLow_price	238	410.7	220.3	128.1	1,958
FTSE_HighLow_proprtion	238	0.0607	0.0343	0.0189	0.297
FTSE_Close	238	6,862	508.6	5,577	7,749

Exhibit 2. Summary Table

Because the indices are publicly traded in open market, and the trend effect in each index's price, we suspect that there are variables correlated with each other, to figure this out, we draw Pearson Correlation heatmap. And there are some variables correlated with each other, represented by deep blue in the following graph.

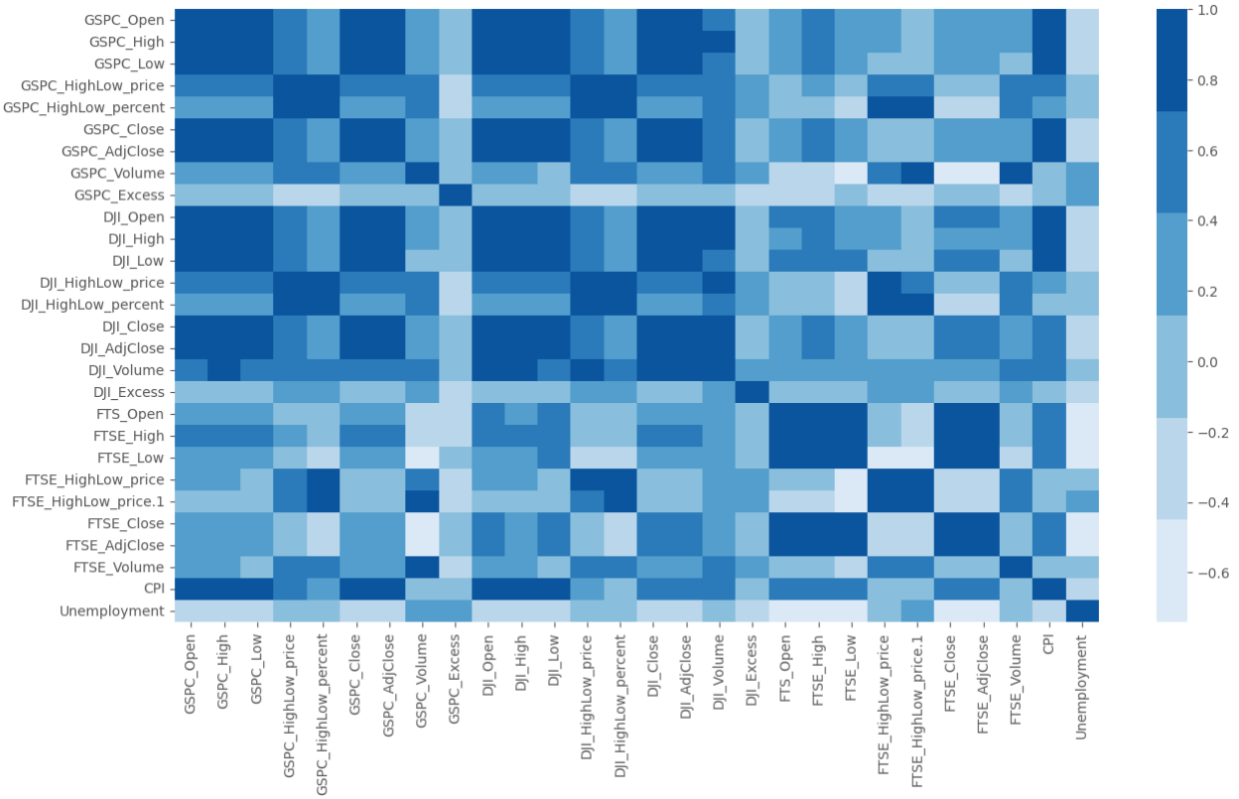

Exhibit 3. Pearson Correlation Heatmap

Also, we are interested in the distribution of indices in each month, which impacts our recommendation of the indices, therefore we draw violin plots and figure out that the prices of indices are much more volatile in the springs and the winters, so if the retail investors are risk-averse, we recommend they hold their share in the summer and fall and sell the shares before winter. However, if the investors are risk loving, they can try to buy shares in the winter when the prices are relatively low and sell when the prices are relatively high. By conducting such aggressive strategies, they might have hire revenues and they need to take more risks of loss.

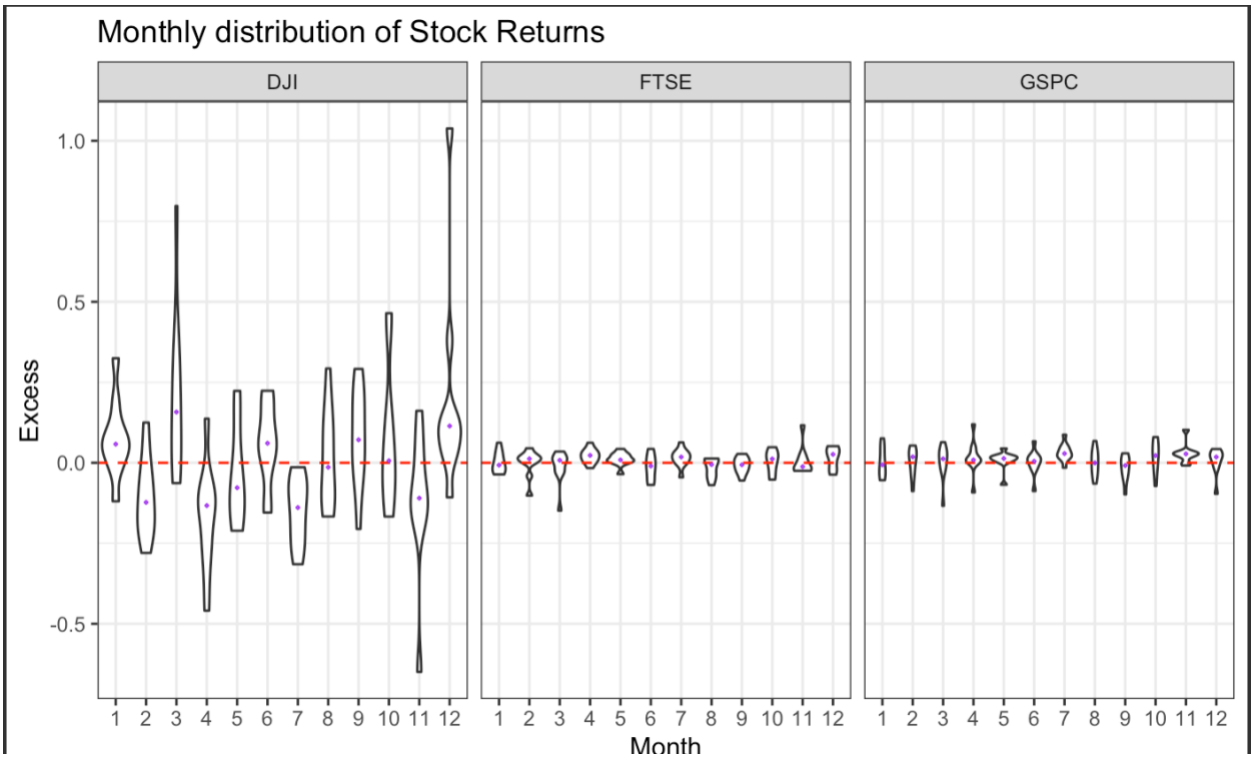

Exhibit 4. Violin plots of Excess_yield in Each Month

Visualizations and interpretations

From all the visualizations, we have noticed that during the prevalence of COVID-19, the stock market was behaving in a manner which was very abnormal, due to people's panic and some people would like to do speculation. This is observed throughout the following visualizations and insights.

Bar Chart: Extremely Large trading volume during 2020 and overall volatility

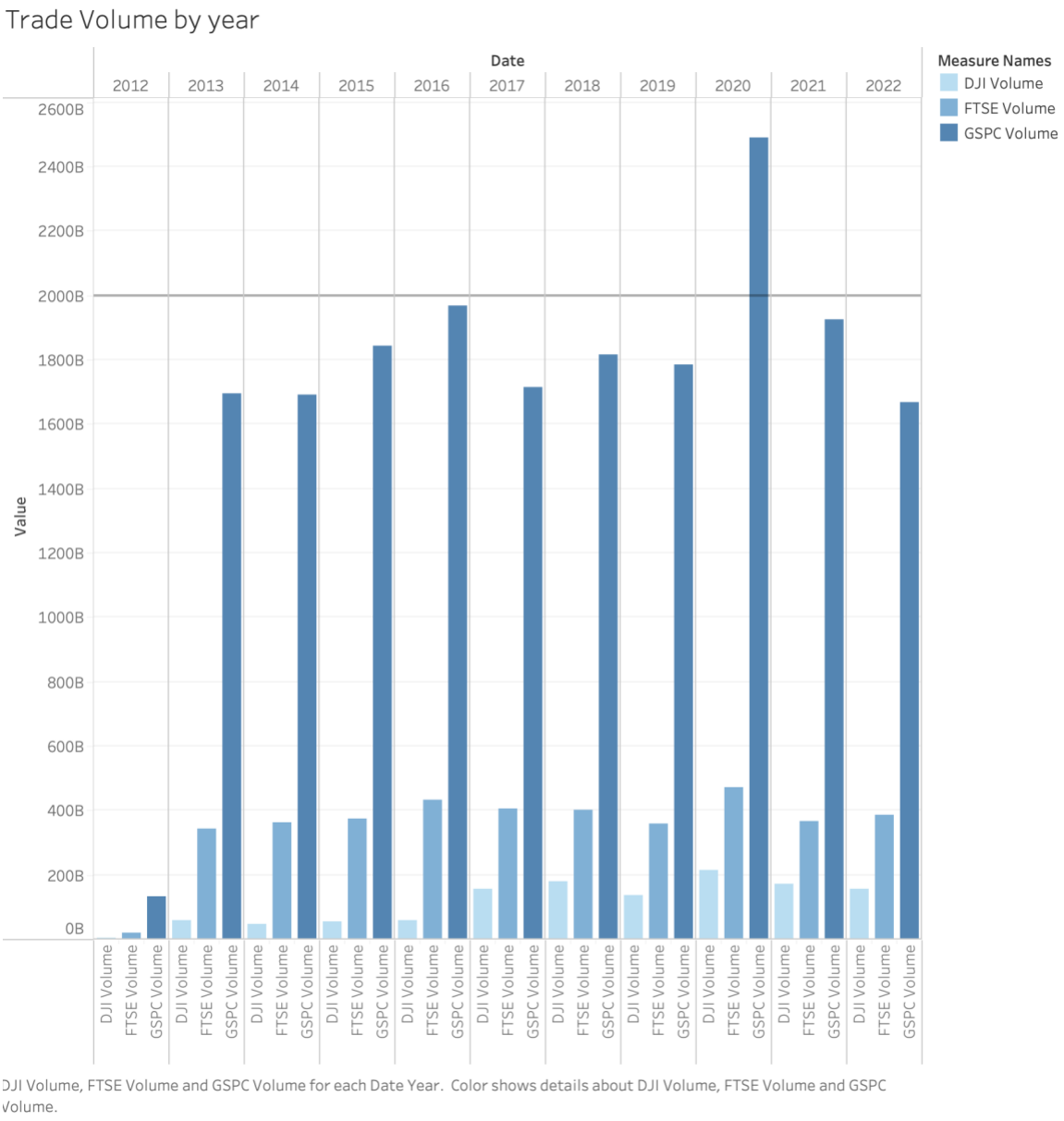

Exhibit 5. Trade Volume by Year

In 2020, the trade volumes of three indices are extremely high. Take GSPC for example, the trade volume exceeds 2500B, and in the other years the number is smaller than 2000B. This could be the result of panic selling due to the outbreak of COVID-19 and the following speculations in the market.

The overall bar chart shows that the index experiencing the highest trade volume is always the S&P500 but it appears that it and the Dow Jones is much more volatile on a yearly basis than the London-based FTSE.

Bar Chart: Profit or Loss in Month

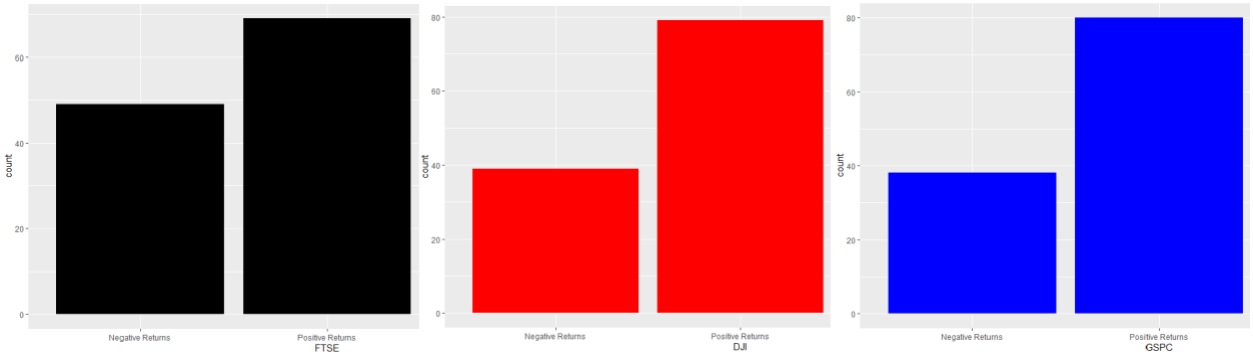

Exhibit 6. Monthly Returns Categorized into Positive and Negative for all three Indices

We categorized returns into positive and negative for each index. According to our bar chart visualizations for each index, it seems that FTSE records the most months with negative returns, at 50 months in comparison to only 40 months of negative returns for DJI and the S&P. According to these bar charts, it seems that the S&P500records the most months of positive returns relative to months with negative returns.

Bar Chart: Average excess yield per year per index

We computed the average excess yield throughout the year of each index, and we found that the excess yield of DJI is most volatile among all three, but it generates a lot of excess yield in 2016 and 2020, so for risk-loving investors, they should focus more on DJI so that they can make more money taking more risks.

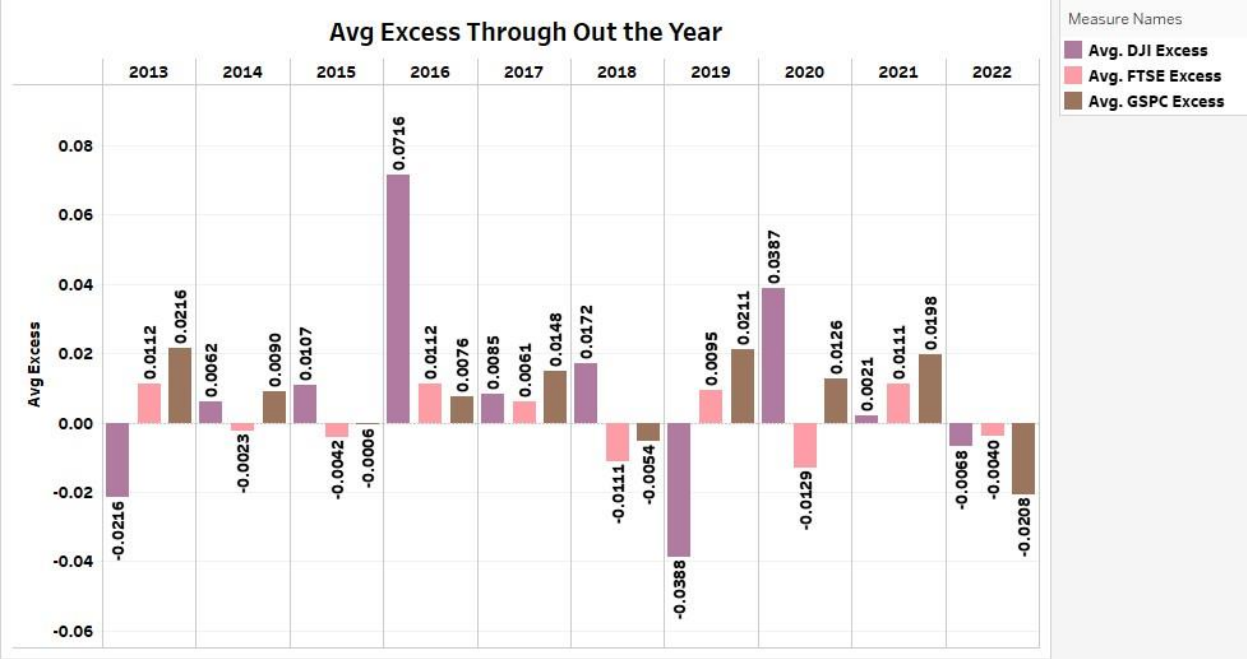

Exhibit 7. Average excess yield per annum of each index

Time Series plot: The shock of three indices in March

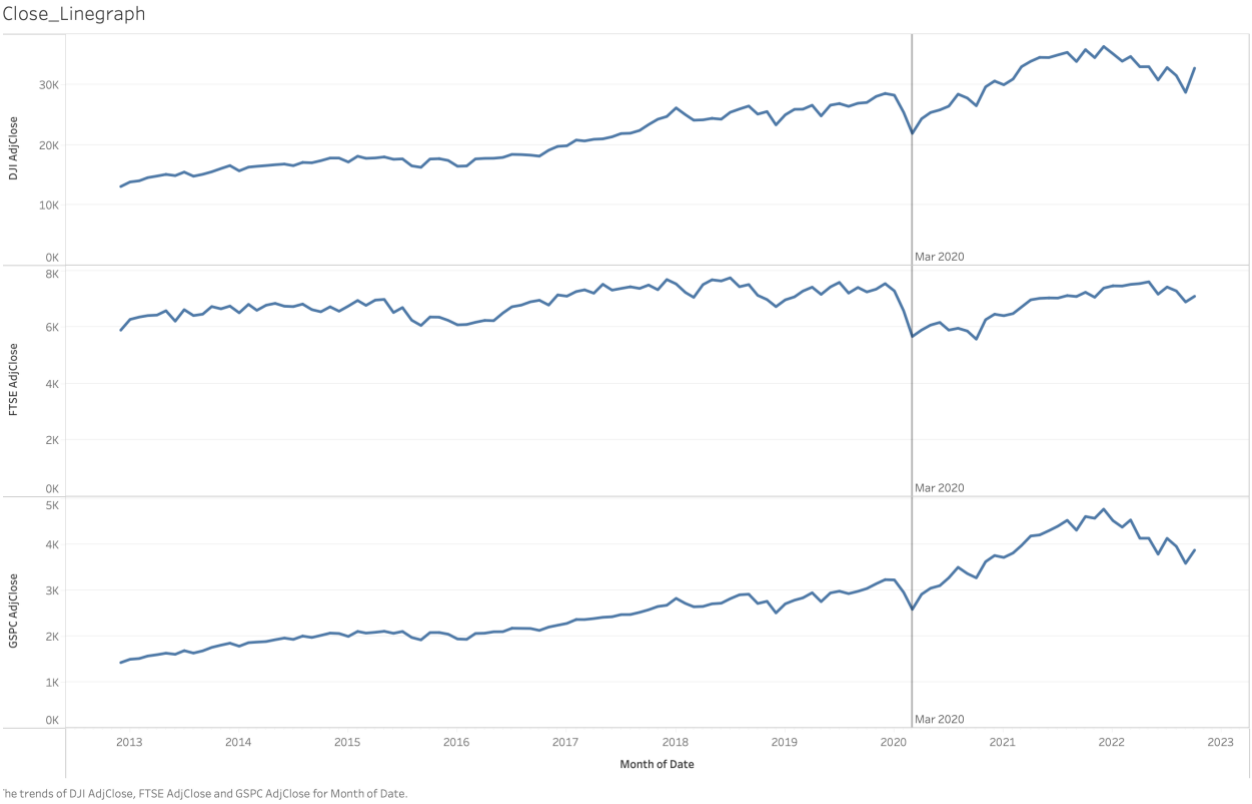

Exhibit 8. Close Prices of indices

In March 2020, when COVID-19 began its global prevalence, three indices dropped down drastically. Especially for FTSE, for which the price dropped to its lowest point since 2013. However, thanks to the people who made a lot of effort to mitigate the impact, including government, volunteers and people changed their living mode due to the quarantine, the shock of the market was weaker after one month the prices went up again. This is a representation of financial stability in modern society.
The overall plot shows that these the close price of these indices are quite similar in their trends, seasonality, and ways of going up and down. On first view, they seem highly correlated.

Tree map: COVID Recession and overall unemployment throughout the years

During 2020, the unemployment rate was 8.092%, which is its peak from our available data. This is the result of the bankruptcy of many firms caused by COVID-19 and the following quarantine. Also, in contrast to the mild CPI in years before in line with the 2% target, the average CPI is only 1.237. To mitigate these problems, the government took easy fiscal policies in 2021 and 2022, leading to a high inflation rate, which triggered a tight monetary policy on the federal reserve that had to increase the federal funds rate by a lot in 2022.

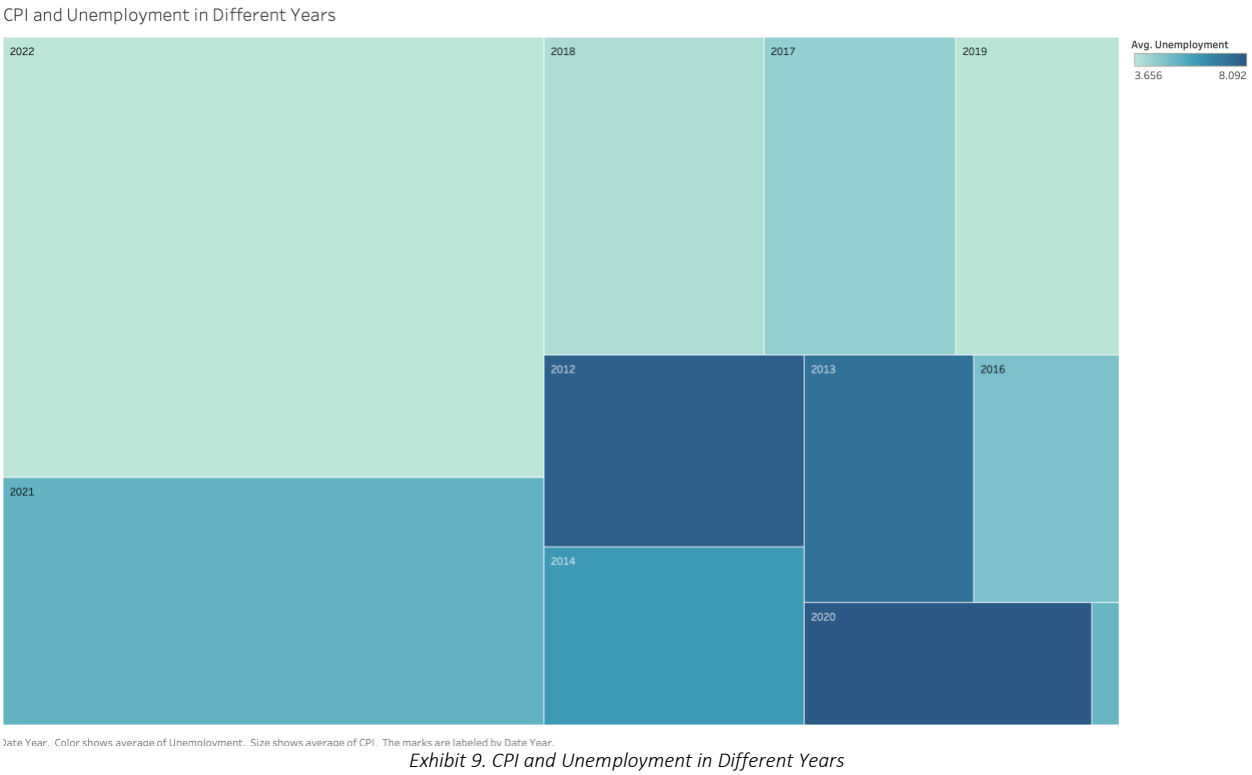

Exhibit 9. CPI and Unemployment in Different Years

Scatter plot: Different trends in the trades

We can observe a positive relationship between Dow Jones returns and volume. The relationship does not seem very strong. This scatterplot shows that the relationship would be stronger with just data preceding 2016. It seems that the general trading volume has increased despite the returns not being higher since 2018.

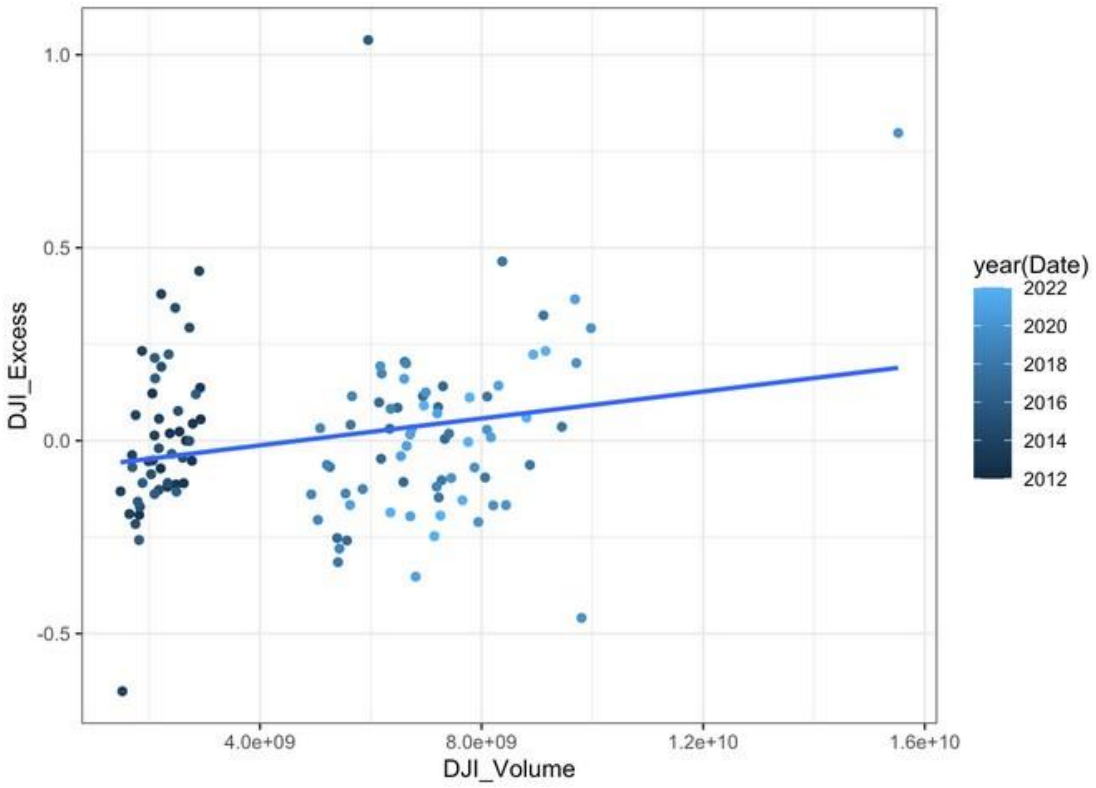

Exhibit 10. Excess yield vs trade volume of Dow Jones Index

Factor analysis: variability in the model

Because some of the variables are highly correlated, we try to use some dimension reduction methods to describe the variation in the model. Here we use factor analysis.

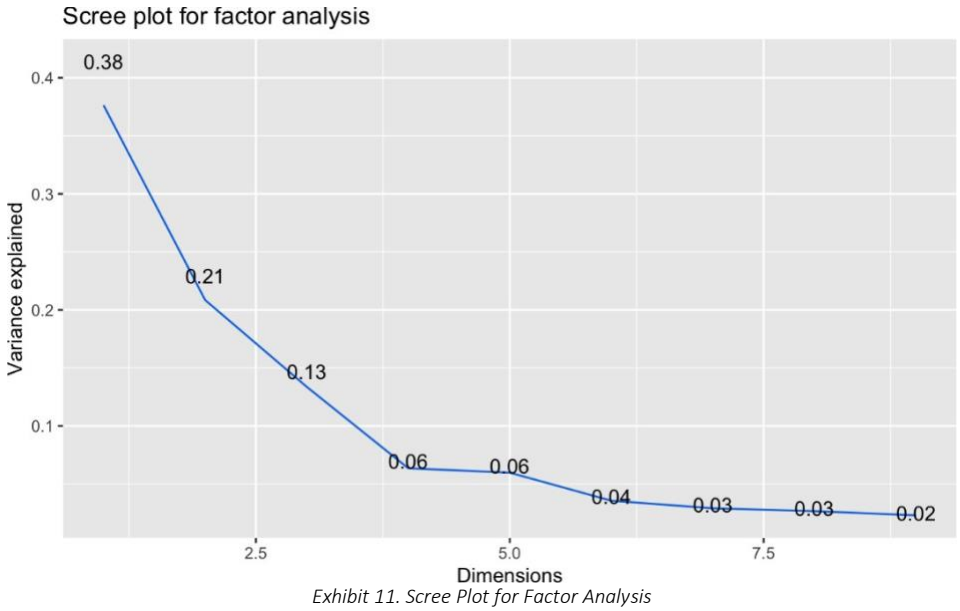

Exhibit 11. Scree Plot for Factor Analysis

Based on the scree plot and elbow technique, we chose the first four factors naming them excess yield factor, performance factor, macro factor, and volume factor. To find out how they interact with each other, we introduce the ellipse plots and facet plots.

Exhibit 12. Ellipse plot for Performance and other factors

In the model we found the performance of the indices is highly correlated with macro, and slightly correlated with the volumes, so we decide to perform a time-series regression on it.

Before we run the model, we need to make sure that the variables in the model are stationary, so we draw line graphs and run Augmented Dickie Fuller tests on them, and we found that the Performance and Macro are not stationary, but the Volume is stationary, so we take differentiation on the former two variables, and their differences are stationary.

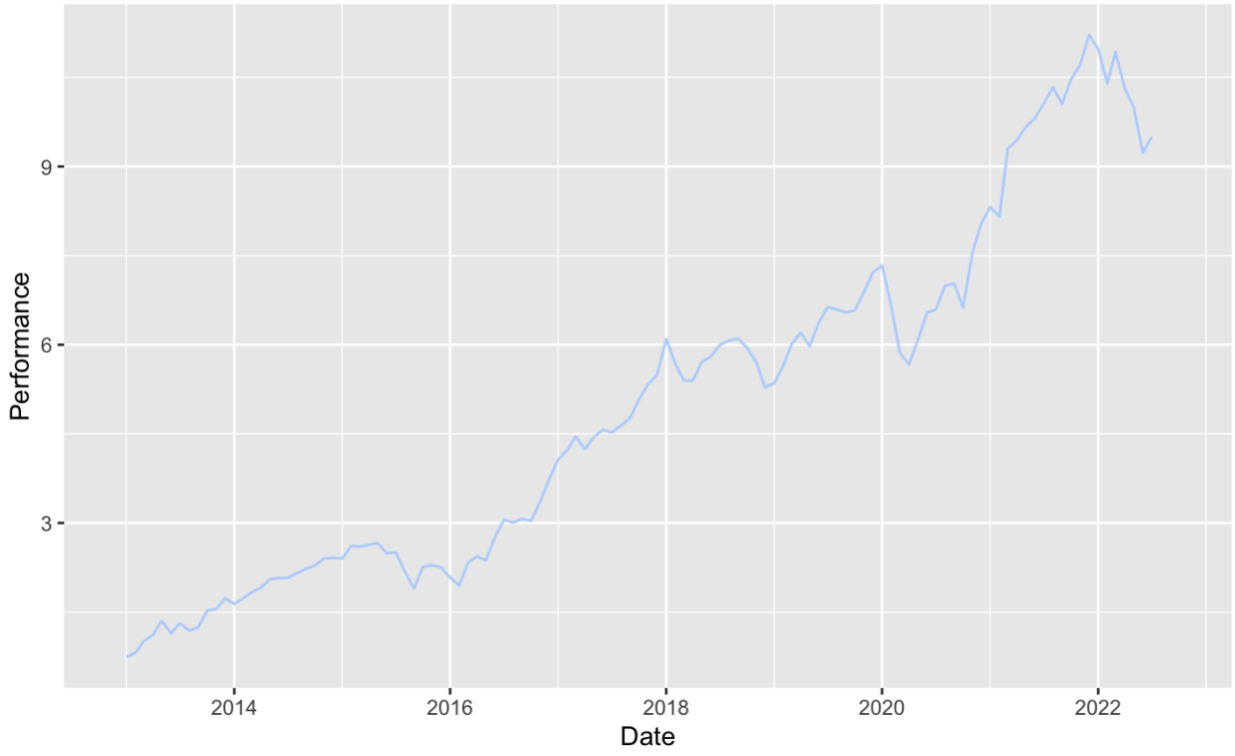

Exhibit 13a. Line Graph of the Performance

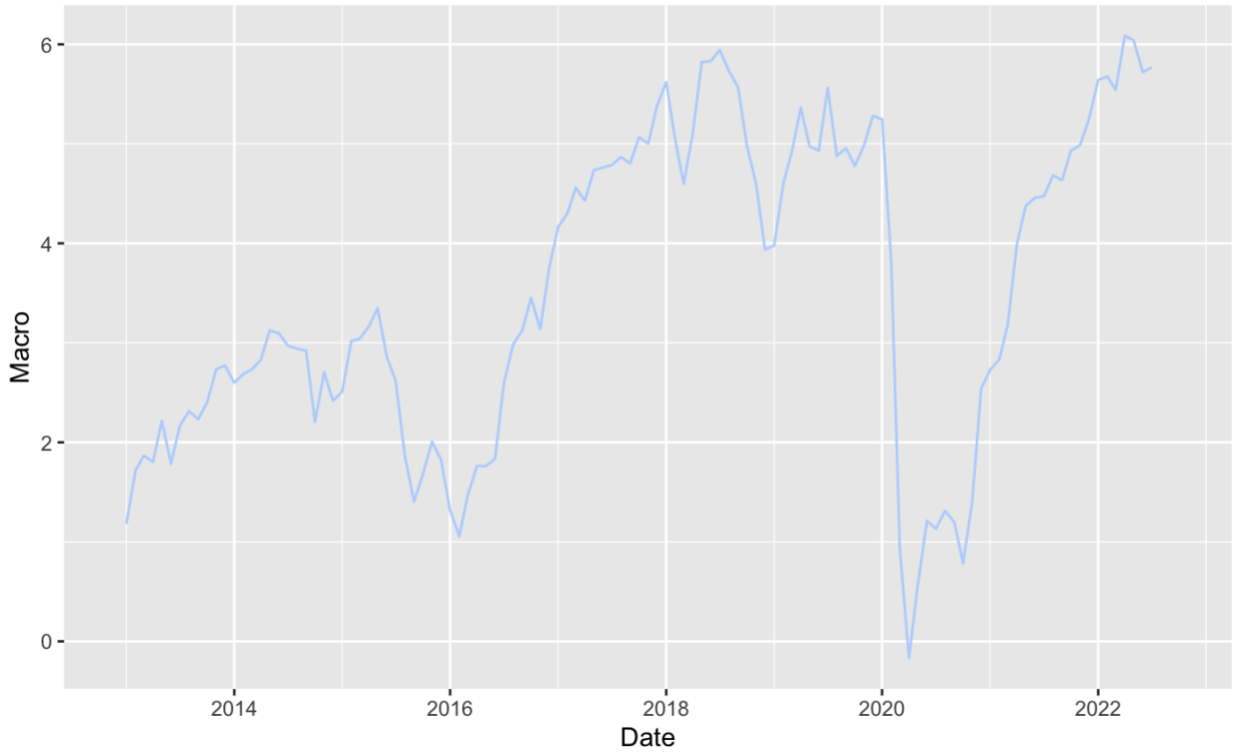

Exhibit 13b. Line Graph of Macro Factors

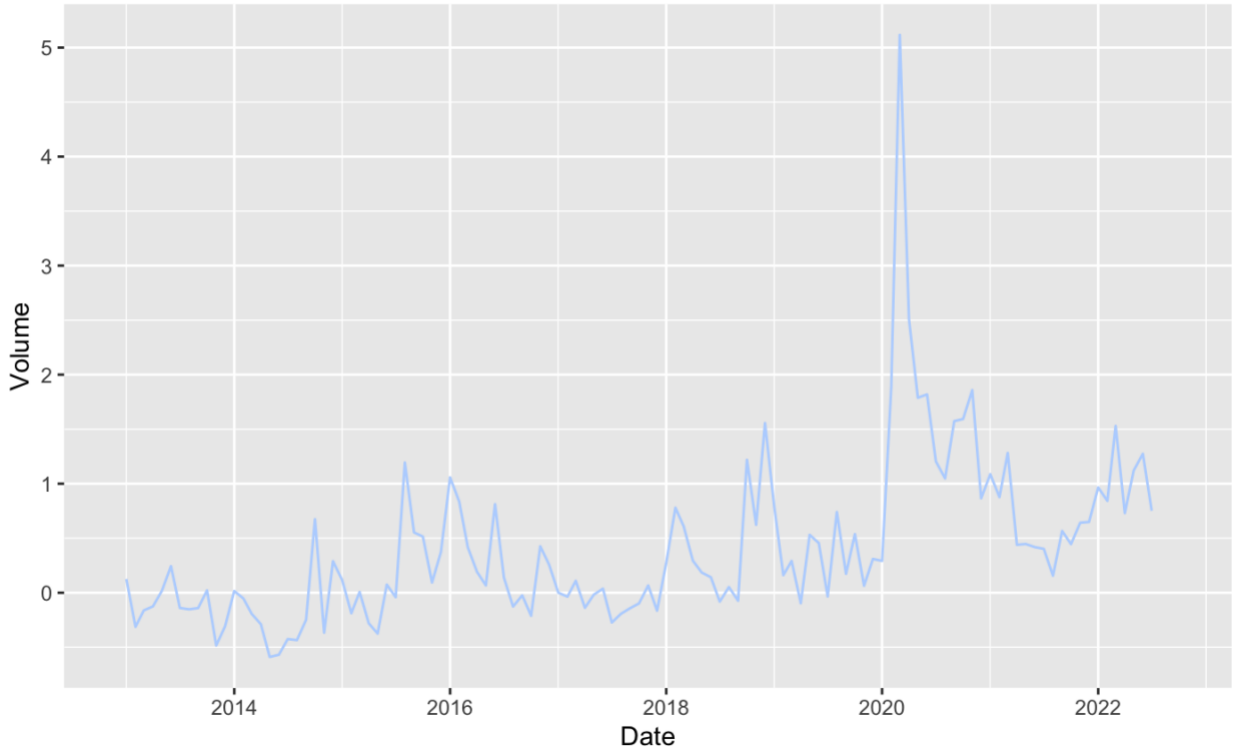

Exhibit 13c. Line Graph of Trading Volume

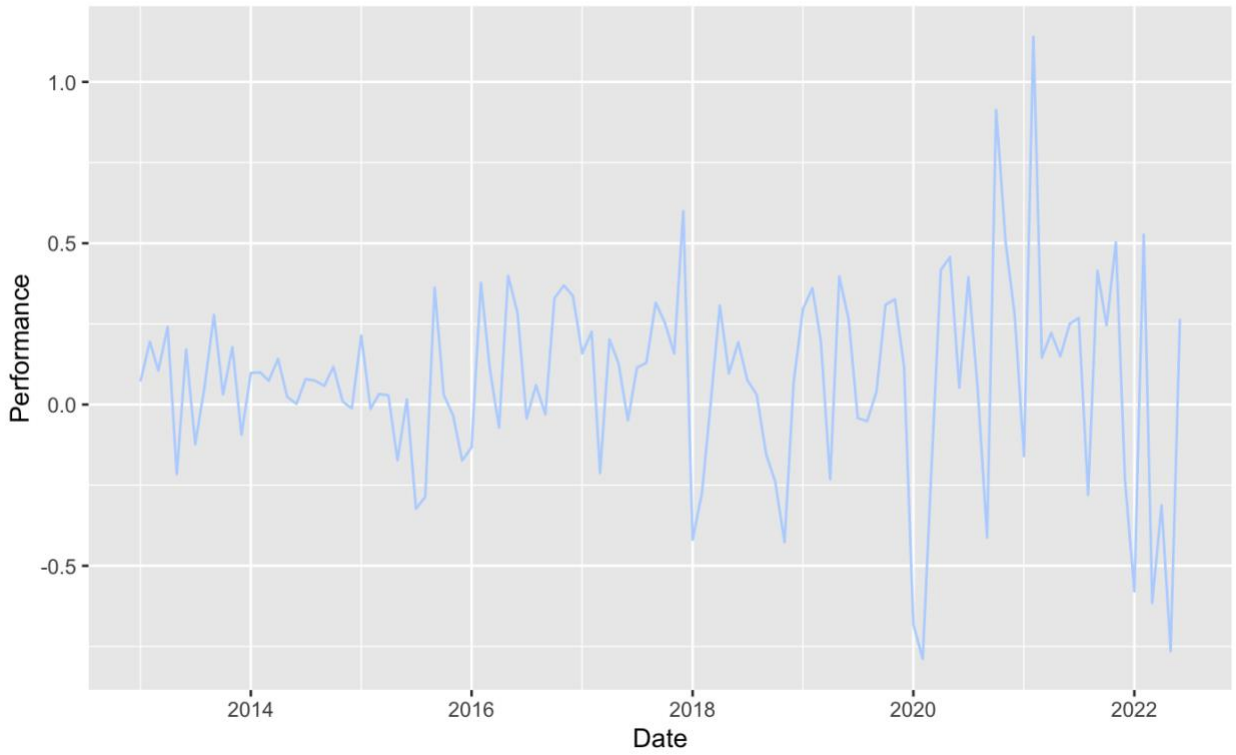

Exhibit 14a. Line Graph of Differenced Performance

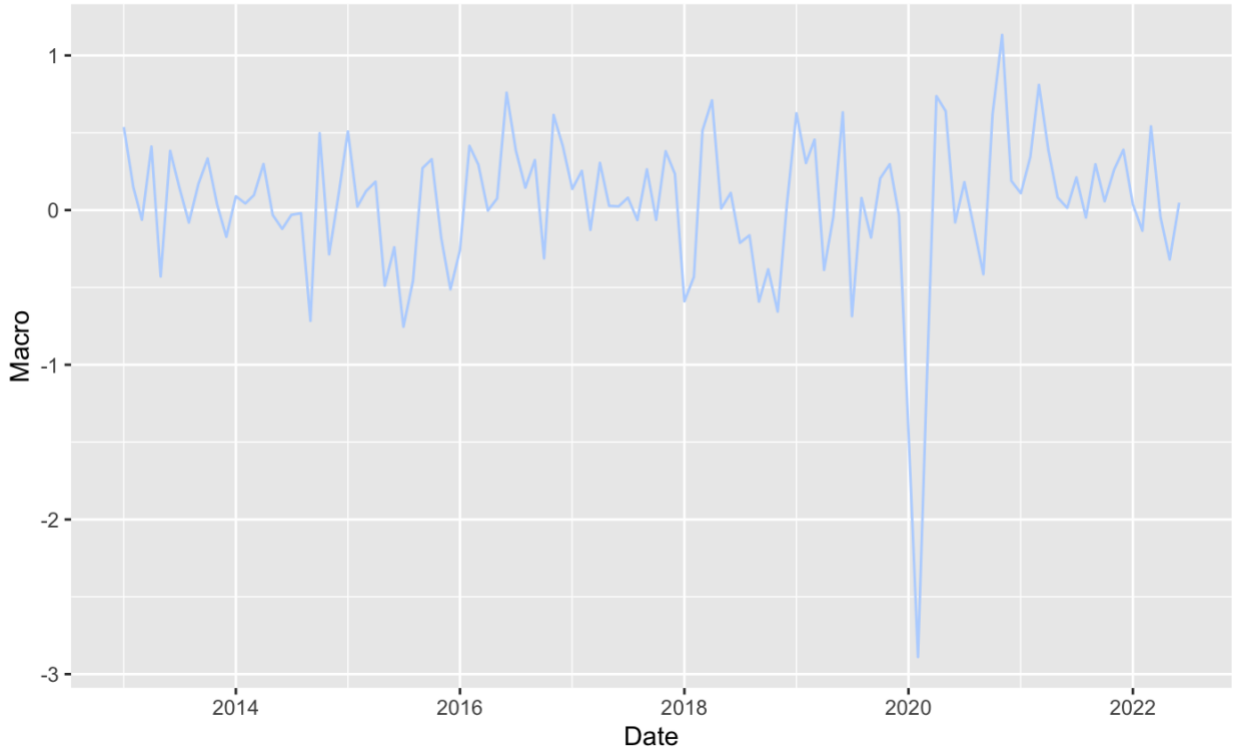

Exhibit 14b. Line Graph of Macro Data

It could be seen from the line graphs above, after taking differences, the macro and the performance become now stationary, and it is proved by the ADF test.

As suggested by the ADF test and the line graphs, the Performance and Macro are non-stationary, and after taking differences, the linear relationship is clear in the scatter plot below. Also the residuals of the regression are stationary, which means no extra dynamic factors are contained in the residuals.

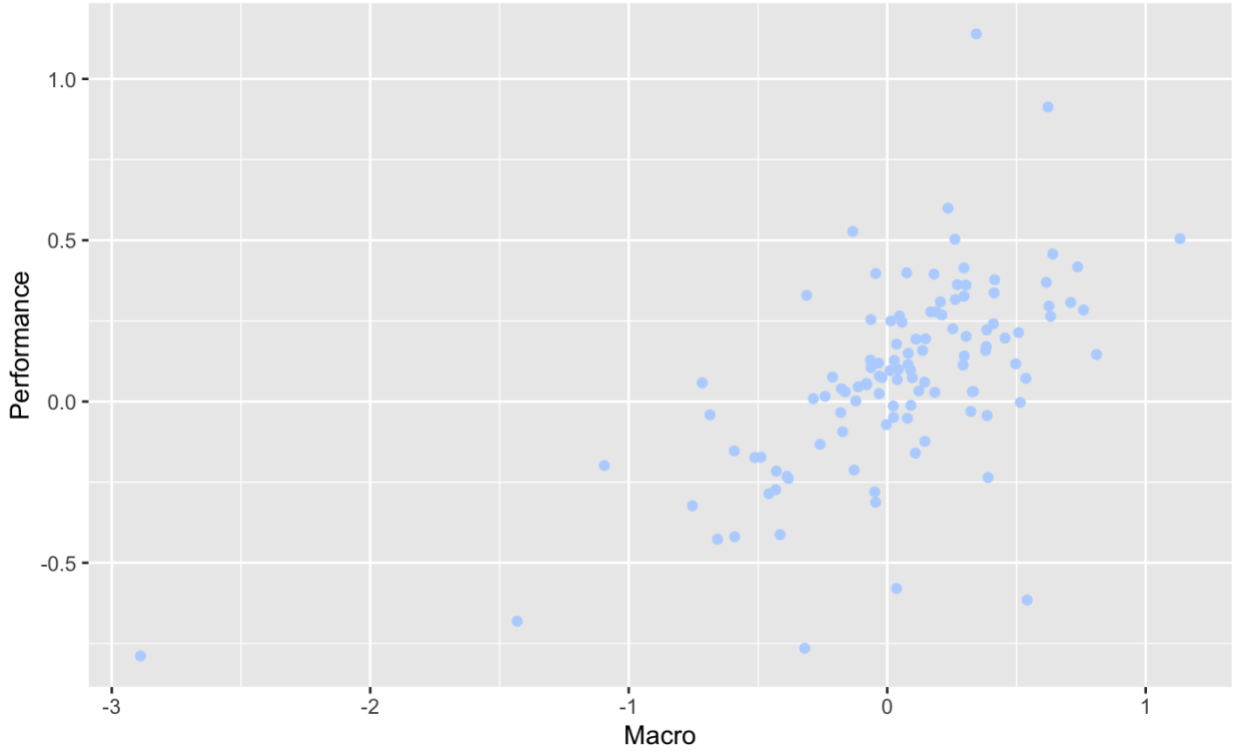

Exhibit 15. Performance vs. Macro scatter plot, both variables are differenced

Then we run a regression using and find out that around 40% percent of the variation in the performance can be explained by the macro.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0528	0.0278	1.9001	0.0601 .
L(dperformance, 1)	-0.1529	0.0802	-1.9064	0.0592 .
dmacro	0.4608	0.0561	8.2203	0.0000 ***
volume	0.0456	0.0347	1.3110	0.1926

$R^2_{adj} = 0.4025$

F-statistic: 26.15 on 3 and 109 DF, p-value: 8.076e-13

Exhibit 16. Regression Table, the model is $\Delta Performance_t = \beta_0 + \beta_1 \Delta Performance_{t-1} + \alpha \Delta macro_{t-1} + \gamma volume_{t-1} + \epsilon_t$

Reflection about The Process

It would be useful for us to have more categorical data like monthly extreme weather events data and location-based data. Some additional services that we could provide on top of observing the fluctuations in trade volume and the relationship between index prices and macroeconomic variables could be looking at the relationship of stock prices with commodity prices.

We have begun this process through this correlation matrix but ideally, we could use much more data and observe relationships, only keeping in the matrix the factors with the highest correlations to each other to observe these relationships.

Interestingly, we could also provide useful forecasting services for different businesses in various industries trying to forecast their sales by exploring the relationship of sales with specific commodity prices and macroeconomic variables. Overall, there is a lot of potential for this consultancy project to be explored later.

It would also be possible to use this data in a more complex manner not yet explored due to lack of technical ability: on the dashboard, using this data we could toggle a specific risk preference and a return range and obtain the best index for these purposes. We could also toggle to obtain the Sharpe ratio which means that the return and risk are optimal. A visualization that gave the Sharpe ratio across time of the different stocks would definitively be more interesting and informative than what we currently have.

Using the dataset created from multiple sources, we created a product that could be helpful for retail investors to help them better understand the indices market and create value for them. In the pre-proccing, we used Tableau Prep to merge the datasets from different sources. Also, we used R to create categorization for visualization. Tableau is easy to use, which only requires users to click the mouse and input the parameters to get good-looking plots, however, such pre-setting method limits users to customize their visualization demands. R, on the contrary, is not so user-friendly because we need to write a script to run on it, however, customizing scripts allows us to develop personalized graphs. Each software has strong tastes, and we should manage each software. Also, Tableau supports R scripts for feature engineering and other plotting, allowing it to have stronger abilities. Tableau allows a rapid overall visualization of the dataset from which we can gain a better understanding of the data, and R allows us to do a deeper analysis of the dataset. Whatever we use, we need to make sure that the visualizations are conducive to better decision-making and understanding to generate more value from the raw data.

Reference

[1] Demir, E., Alıcı, Z. A., & Chi Keung Lau, M. (2017). Macro explanatory factors of Turkish tourism companies' stock returns. Asia Pacific Journal of Tourism Research, 22(4), 370–380. doi:10.1080/10941665.2016.1271