Security Price Prediction Problem

Data

For the analysis, we have considered a broad market-wide index from India i.e. Nifty 50, along with securities (equity shares) of ten companies that are a part of Nifty 50.

The 10 companies considered for analysis are listed below:
1. Reliance Industries
2. Tata Consultancy Services
3. HDFC Bank
4. ICICI Bank
5. Hindustan Unilever
6. ITC
7. Infosys
8. State Bank of India
9. Bharti Airtel
10. Housing Development Finance Corporation (HDFC)

Source of Data:
- NIFTY 50: https://finance.yahoo.com/quote/%5ENSEI/history/
- Ten companies listed above: https://finance.yahoo.com/quote/yhoo/history/

We have considered data for time period from 02-Jan-2018 to 31-May-2023 for the analysis.

The data includes fields for:
1. Date - Date of trading session
2. Open - Opening Price for the trading session
3. High - Highest Price during the trading session
4. Low - Lowest Price during the trading session
5. Close* - Closing Price for the trading session (adjusted for splits)
6. Adj Close** - Adjusted Closing Price for the trading session (adjusted for splits and dividend and/or capital gain distributions)
7. Volume – Volume of securities traded during the trading session

* Adjusted for splits.
** Adjusted for splits and dividend and/or capital gain distributions.

The 'Adjusted Closing Price' in the data has been taken for analysis as it is adjusted for splits and dividends and/or capital gain distributions.

The data is downloaded programmatically in R using getSymbols function provided by quantmod library with the index and script data fetched from yahoo finance.

Data Pre-Processing

Based on the data downloaded, it was observed that the data for Nifty index was missing for 01-Jan-2019, 27-Oct-2019, and 14-Nov-2020. This was filled with the last day's value from the available data. The data for the other scrips was complete (no missing/NaN values).

Only the adjusted price for the stock was considered for further evaluation.

The initial analysis exhibited the following variations in the script as well as index movement across the time period specified, as shown in Figure 1.

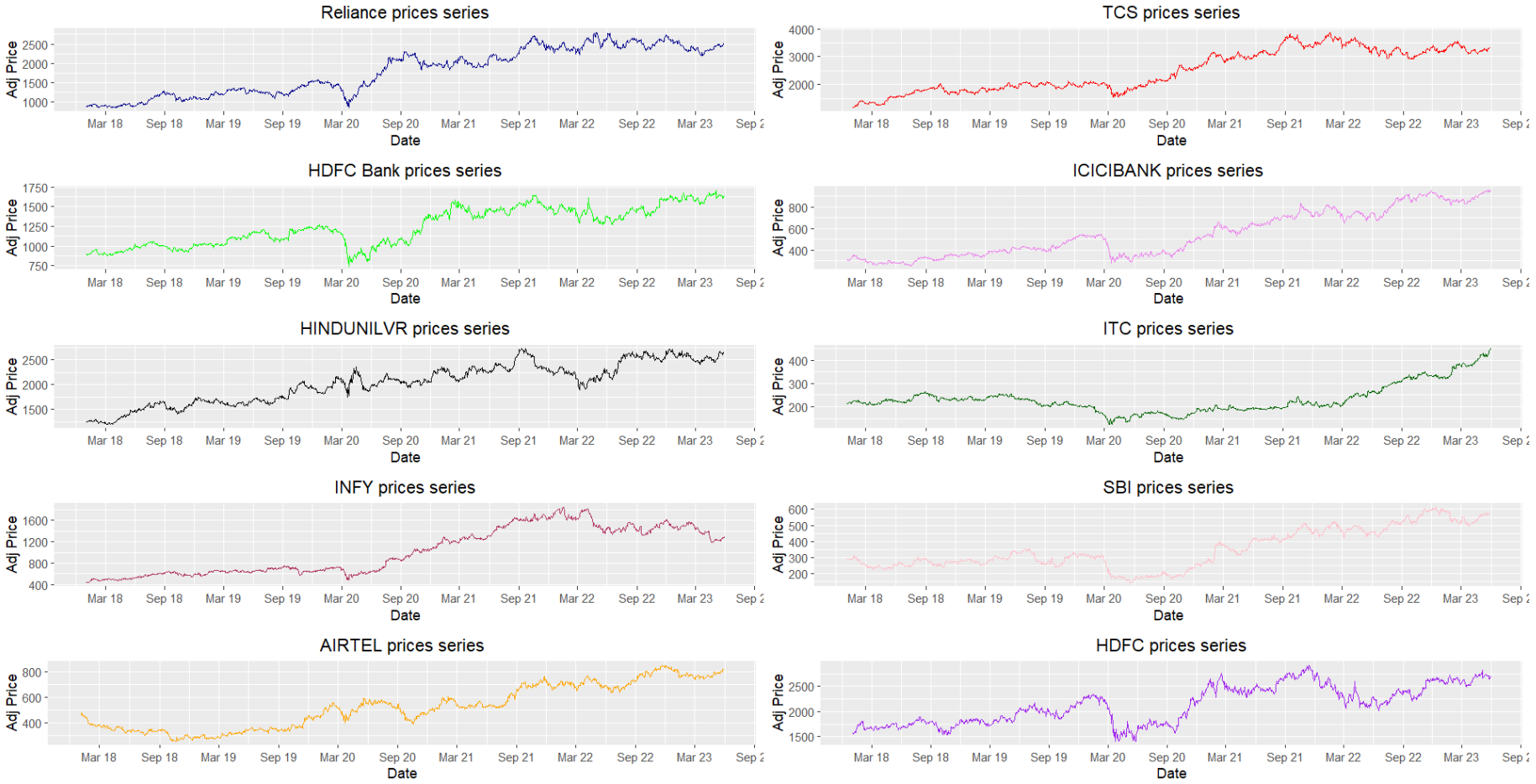


Figure 1: Price variations of various scrips across the time period

Figure 2 shows the Nifty50 price variation for the same time period. As can be seen, the price variation is more prominent post Mar-20 and while all the scrip above show a similar movement, there are variations in the movement, more prominently so post Dec-2021.

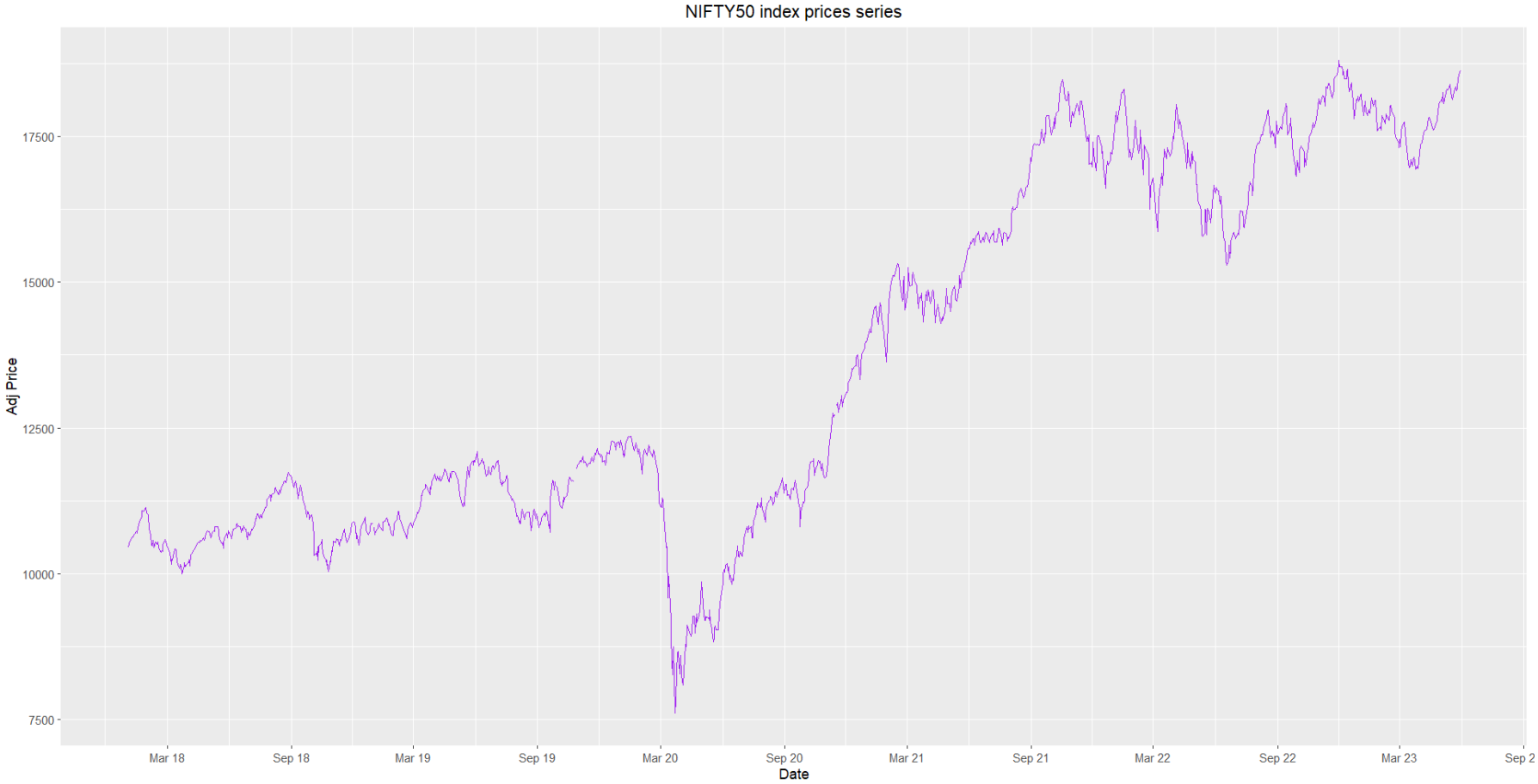


Figure 2: NIFTY 50 variation across the time period

Return processing based on the adjusted price

The scrip uses Return.calculate function provided by xts package to come up with the returns for individual scrips and Nifty50.
Table 1 and 2 indicate the illustrative absolute returns for the scrip and the index respectively.

Table1: Scrip returns for the start and end of the period

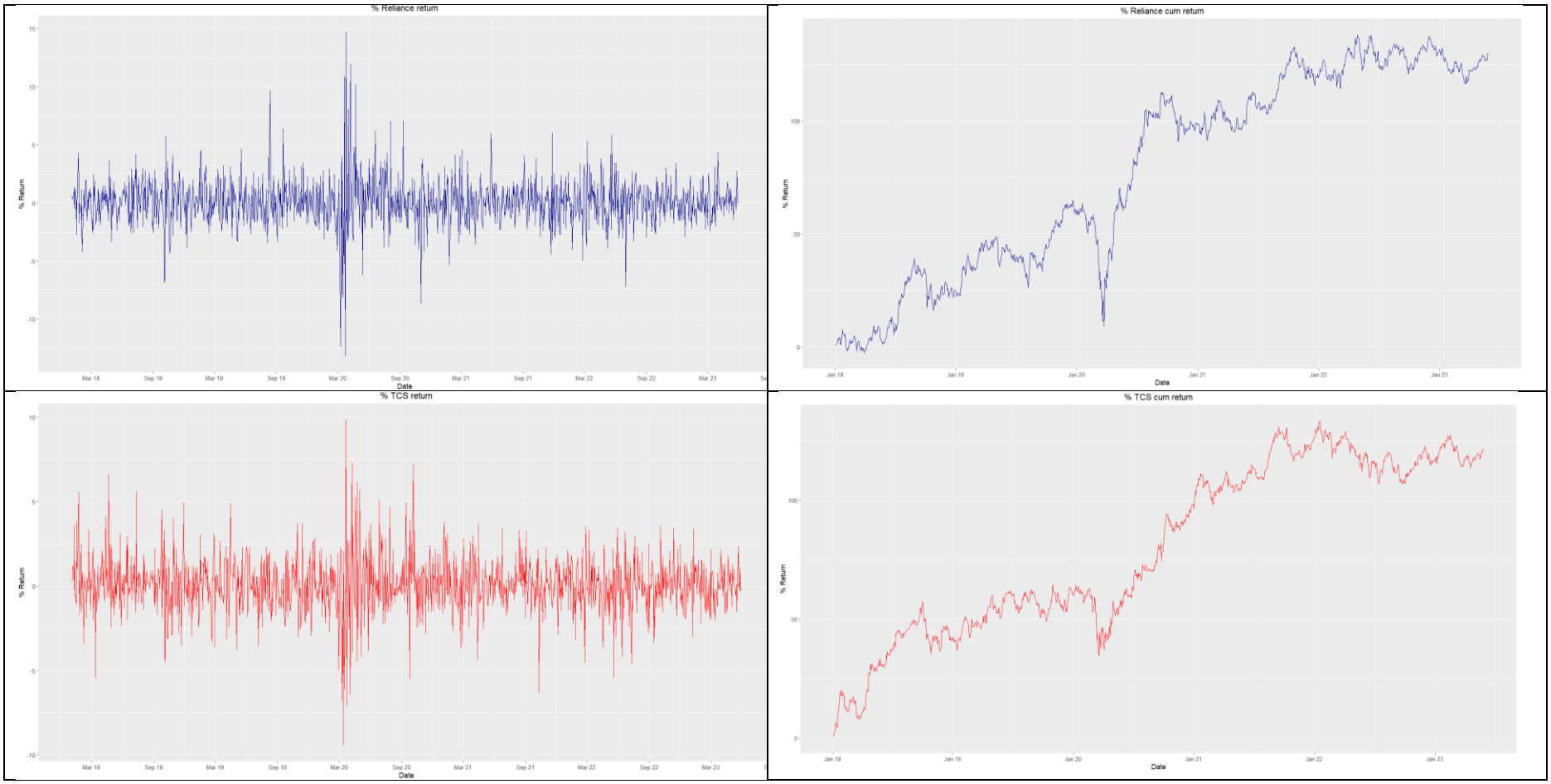
	REL	TCS	HDFCBANK	ICICIBANK	HINDUNILVR	ITC	INFY	SBI	BHaratiAIRTEL	HDFC
2018-01-03	0.4005920	0.2831419	-1.0521449	1.71134196	0.8777752	0.1534106	0.8157699	-0.1319047	0.3194264	-0.1438388
2018-01-04	0.6012297	0.6992573	0.3913288	-0.09522801	0.1407203	0.1531756	0.5776948	1.8656042	0.9262859	0.1410945
2018-01-05	0.3205465	1.2080570	0.1989370	-0.57198261	0.3179815	0.5161484	0.3348375	-0.6969224	3.2504759	1.2917026
2018-01-08	0.5740545	0.9370864	-0.1609883	0.39946813	0.8735359	1.0270032	2.3764856	-0.1795441	-4.3796221	0.9129341
2018-01-09	1.3354201	-0.1989328	0.1666194	-0.49336442	-0.5188497	1.8637119	0.5115586	-0.4905076	-1.2685208	-0.3389032
2023-05-24	-0.58259349	0.1910814	-1.30710377	-1.33731903	-0.50793659	1.0136318	-0.1076933	0.1891457	0.1812629	-1.2602360
2023-05-25	-0.01229582	-0.2981851	-0.38371538	-0.08505276	-0.67687533	1.7647043	0.4389127	-0.2488422	2.6578472	-0.9113913
2023-05-26	2.72751696	1.0748414	0.38519342	1.15444346	2.12147840	0.5553666	1.0043294	0.8171988	-0.5773672	0.1208807
2023-05-29	0.56254130	-0.2568357	1.21920724	-0.26822213	-0.07729179	1.2398597	-0.2314988	1.5358343	0.4034461	1.5072040
2023-05-30	-0.02579332	-0.1249912	0.07642923	0.18984209	0.23582236	0.7953372	0.7189529	-0.3697405	-0.3409422	0.1040698

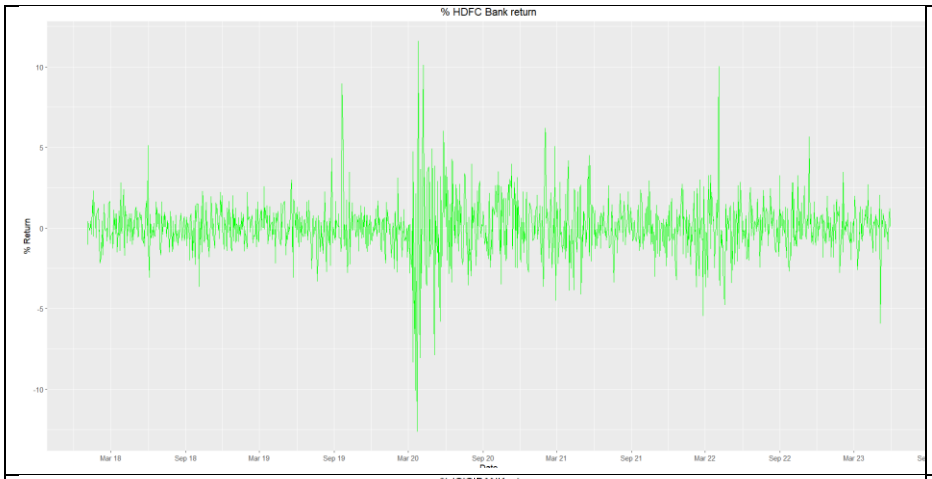
Table 2: Nifty50 index return for the start and end of the period

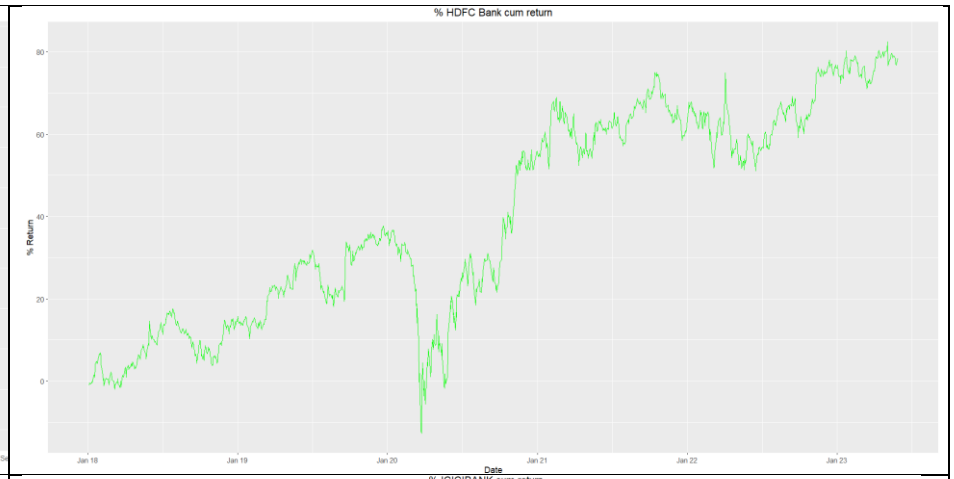
	Nifty_adj
2018-01-03	0.009576526
2018-01-04	0.589853763
2018-01-05	0.514524843
2018-01-08	0.613229683
2018-01-09	0.126137949
2023-05-24	-0.3411795
2023-05-25	0.1955112
2023-05-26	0.9726421
2023-05-29	0.5367799
2023-05-30	0.1892568

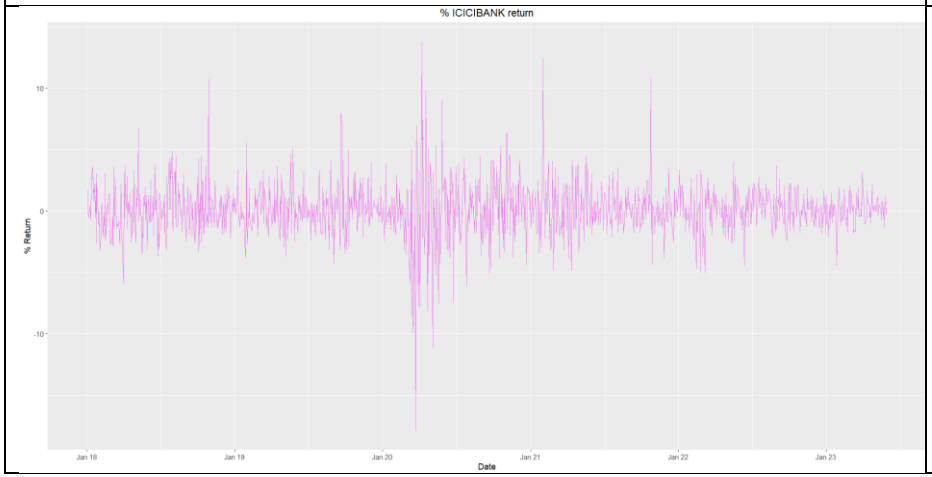
The visualization for the entire range is exhibited in Figure 3 below.

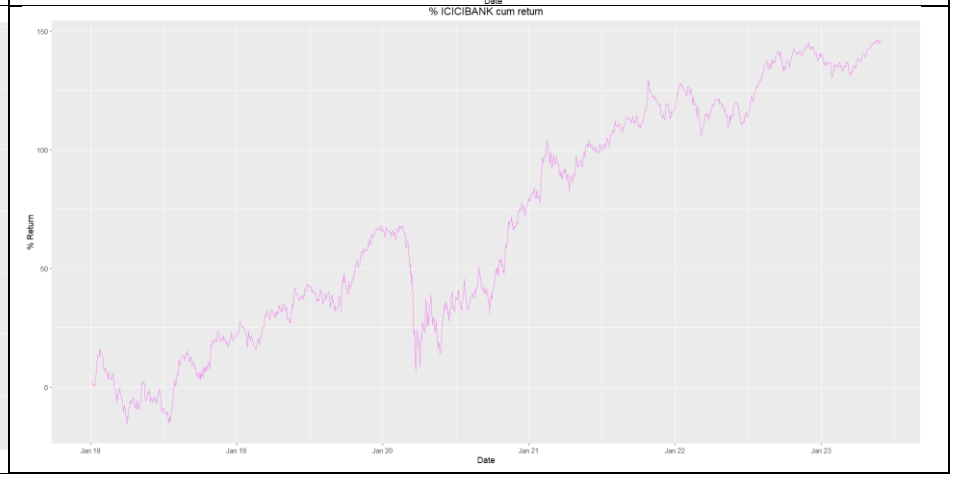
Figure 3: Adjusted price variation across the time period, both scrips and index

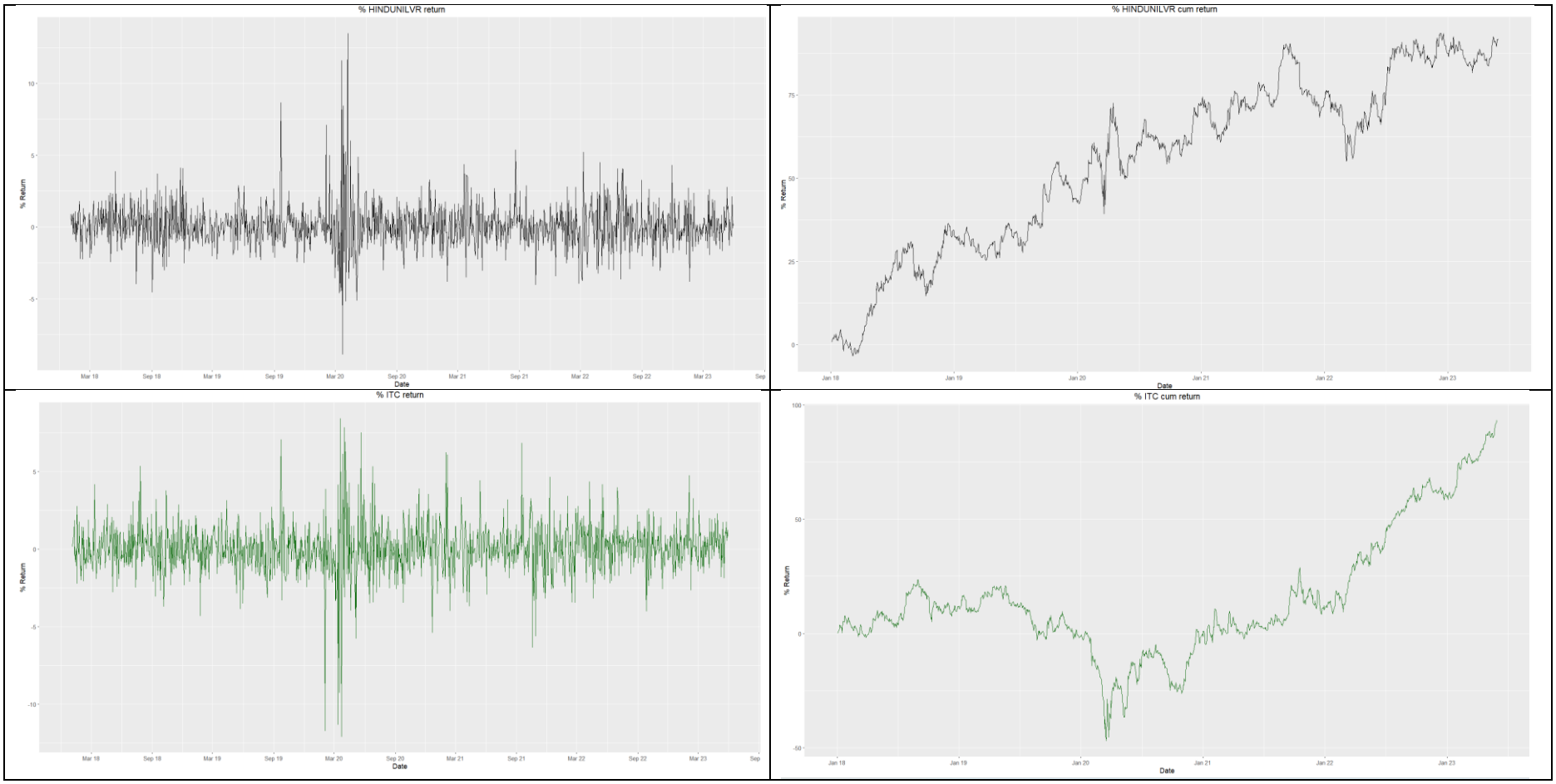


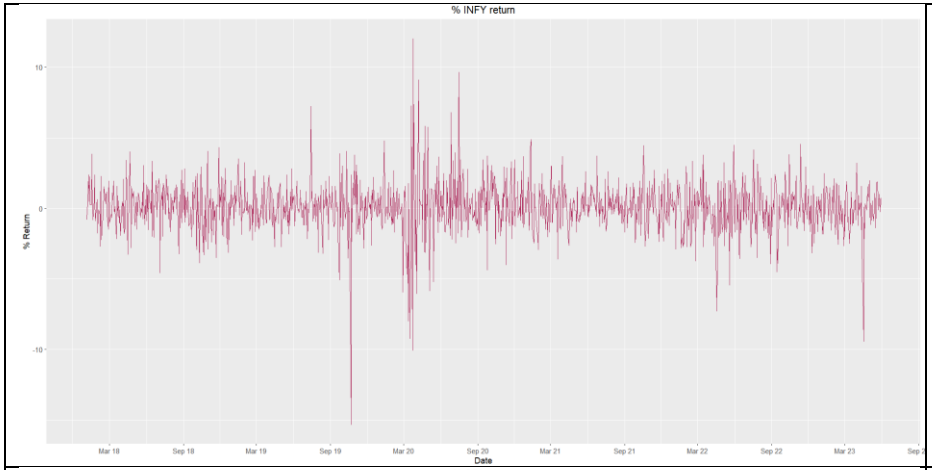

% HDFC Bank return
% Return
10
5
0
-5
-10
Mar 18
Sep 18
Mar 19
Sep 19
Mar 20
Sep 20
Mar 21
Sep 21
Mar 22
Sep 22
Mar 23

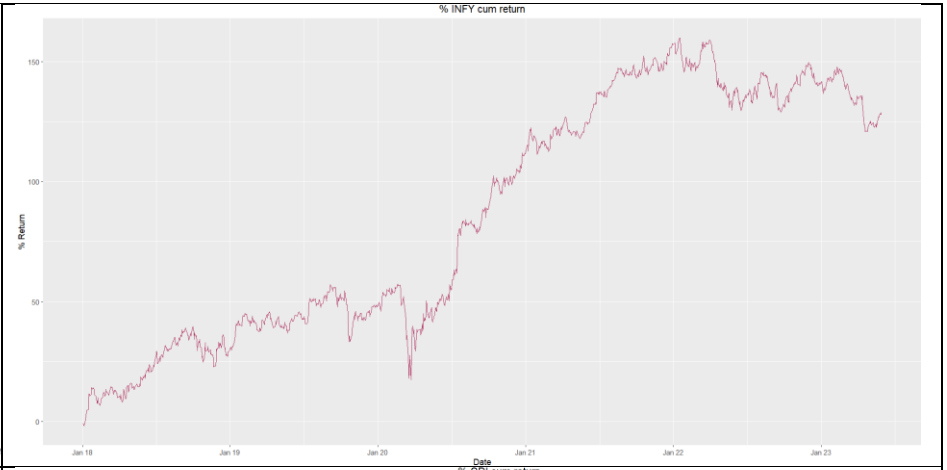

% HDFC Bank cum return
% Return
80
60
40
20
0
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22
Jan 23
Date

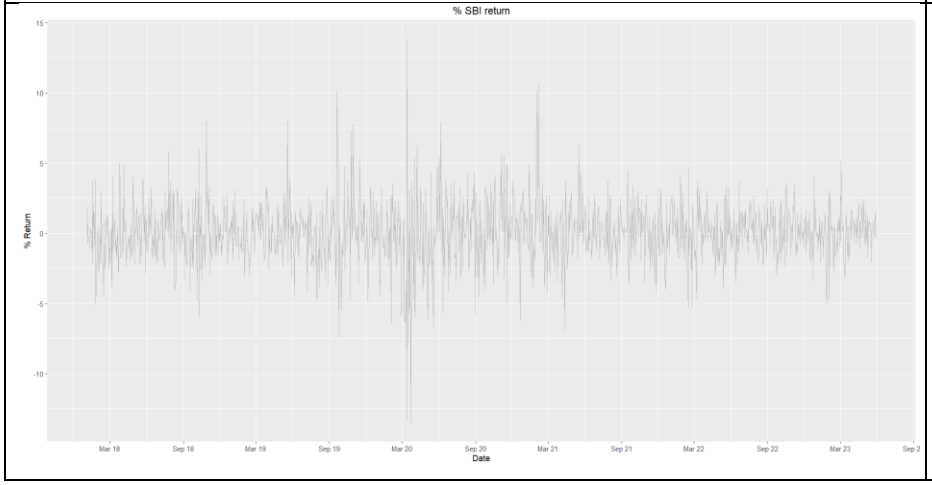

% ICICIBANK return
% Return
10
0
-10
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22
Jan 23
Date


% ICICIBANK cum return
% Return
150
100
50
0
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22
Jan 23
Date

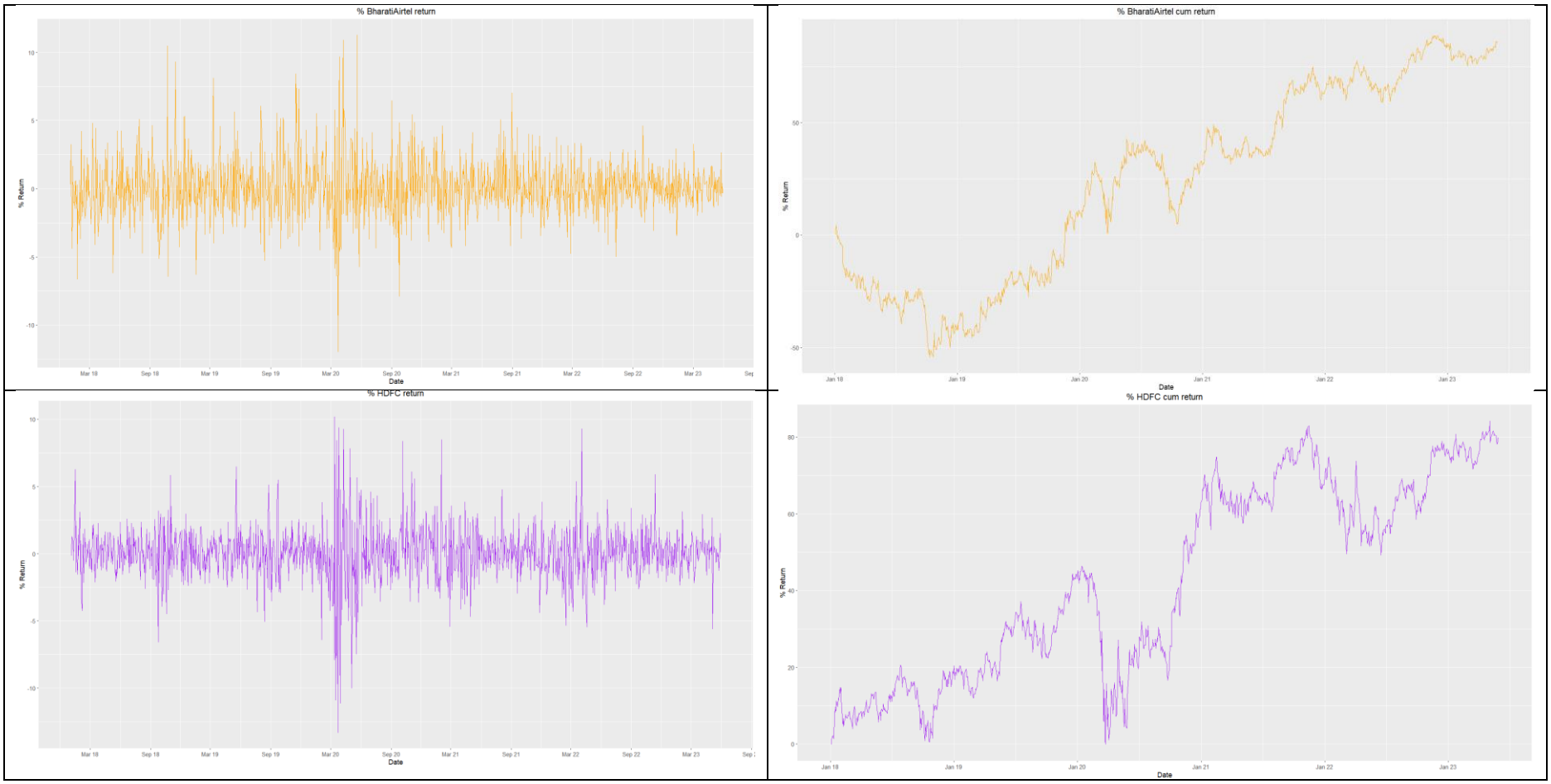

% HINDUNILVR return
% Return
Date
% HINDUNILVR cum return
% Return
Date
% ITC return
% Return
Date
% ITC cum return
% Return
Date

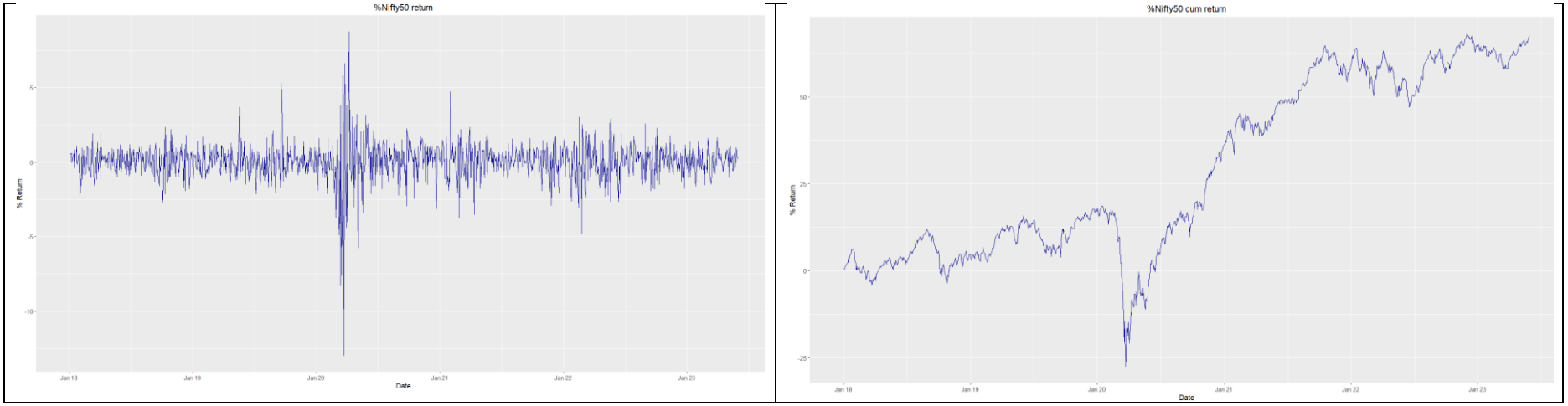

% INFY return
% Return
10
0
-10
Mar 18
Sep 18
Mar 19
Sep 19
Mar 20
Sep 20
Mar 21
Sep 21
Mar 22
Sep 22
Mar 23
Date


% INFY cum return
% Return
150
100
50
0
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22
Jan 23
Date


% SBI return
% Return
15
10
5
0
-5
-10
Mar 18
Sep 18
Mar 19
Sep 19
Mar 20
Sep 20
Mar 21
Sep 21
Mar 22
Sep 22
Mar 23
Date


% SBI cum return
% Return
100
50
0
-50
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22
Jan 23
Date


% BharatiAirtel return
% Return
Date
% BharatiAirtel cum return
% Return
Date
% HDFC return
% Return
Date
% HDFC cum return
% Return
Date



Due to advent of pandemic in Mar-20, most of the scrips, including Index, show a sudden dip. Interestingly, Hindustan Unilever and Bharti Airtel did not show this behaviour.

Equal weighted portfolio returns

To create the equal-weighted portfolio for the 10 scrips, a weight of 10% was assigned for each scrip. Then it calculates the returns of the portfolio across the same time period as above.

Table 3 below illustrates the absolute returns for the above equal-weighted portfolio.

Table 3: Absolute returns for the above equal-weighted portfolio

	portfolio.returns
2018-01-03	0.16020298
2018-01-04	0.42440936
2018-01-05	0.55103688
2018-01-08	0.21791782
2018-01-09	0.05864678
2023-05-24	-0.3718048
2023-05-25	0.1811828
2023-05-26	1.0499771
2023-05-29	0.4426383
2023-05-30	0.1459692

Figure 4 shows the absolute and cumulative returns for the above portfolio across the time period.

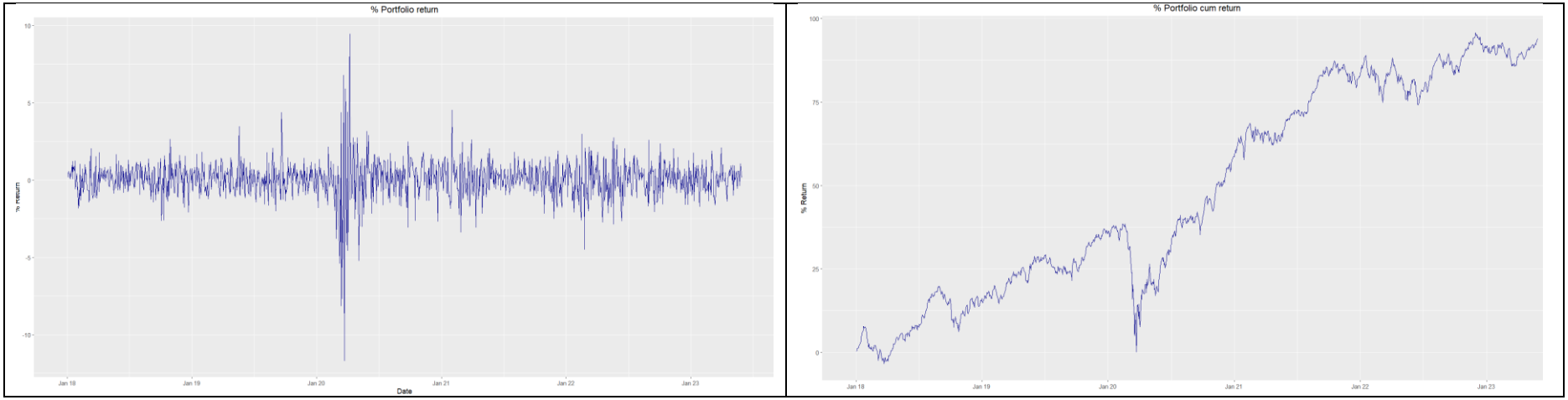


Figure 4: Absolute and cumulative returns for equal-weighted portfolio

As can be seen in Figure 4, this is very close to the trend seen in the Figure 3 for the Nifty 50 index.

Rebalancing the portfolio is important because over time, based on the returns of the investments, each asset class's weighting will change, altering the risk profile of the portfolio. To ensure that the portfolio is composed in a manner that is resilient to the market dynamics, rebalancing is an important practice. Therefore, we also attempted the rebalance the portfolio using monthly rebalancing technique. The rebalanced portfolio exhibits better annualized returns compared to the standard equal weighted portfolio as shown in Table 4.

Table 4: Performance comparison between standard equal weighted portfolio and the rebalanced portfolio

	Non-Rebalanced	Monthly Rebalanced
Annualized Return	0.1737	0.1959
Annualized Std Dev	0.1857	0.1889
Annualized Sharpe (Rf=0%)	0.9353	1.0371

Regression Problem

Training

The dataset obtained above is split into 80:20 ratio for training (in-sample) and testing (out-of-sample) (Nau, 2023) respectively with random sampling.
Predictor variable: Market return
Estimated variable: Equal-weighted portfolio return

The correlation graph indicates a strong correlation between the predictor and estimated variable, as shown in Figure 5. The correlation is observed to be 0.96.

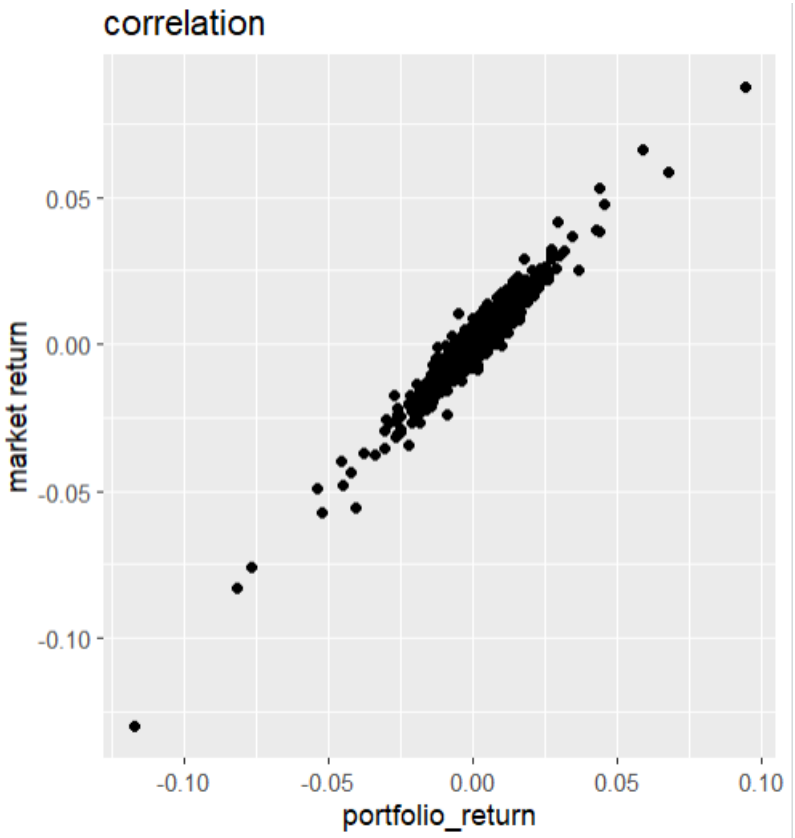


Figure 5: Correlation between the model variables

The Augmented Dickey-Fuller test indicates the statistic of -10.322 and -9.8326 for the portfolio and Nifty returns respectively. It concludes that the null hypothesis is rejected, indicating that the time series is stationary.

The linear regression model fit as per the above, shows the following summary.

```
Call:
lm(formula = portfolio.returns ~ Nifty_adj, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0146909 -0.0018427 -0.0000254  0.0019177  0.0136030

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0001738  0.0000941   1.847   0.0651 .
Nifty_adj   0.9407631  0.0077263 121.762   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003072 on 1065 degrees of freedom
Multiple R-squared:  0.933,     Adjusted R-squared:  0.9329
F-statistic: 1.483e+04 on 1 and 1065 DF,  p-value: < 2.2e-16
```

Figure 6: Linear regression model summary

As can be seen in Figure 6, Nifty adjusted price is explaining the variance in the portfolio by up to 93% and the correlation is significant. Since p value is less than 5% (close to 0) we reject the null (null hypotheses in lm is intercept of independent variable is zero or no significance). Thus, the correlation is significant and positive. Standard error is very small indicating the average of the difference between predicted and actual value is small and indicate higher predictability of the portfolio returns given the Nifty50 return.

The Alkaike’s Information Criteria (AIC) as well as Bayesian Information Criteria (BIC) values for the model are observed to be -9314.36 and -9299.44 respectively. Both these signify the penalty for including additional variables in the model and lower the better. The values for AIC and BIC for the above model are large and negative, which is good.

As seen in Figure 7, the residuals follow almost standard distribution.

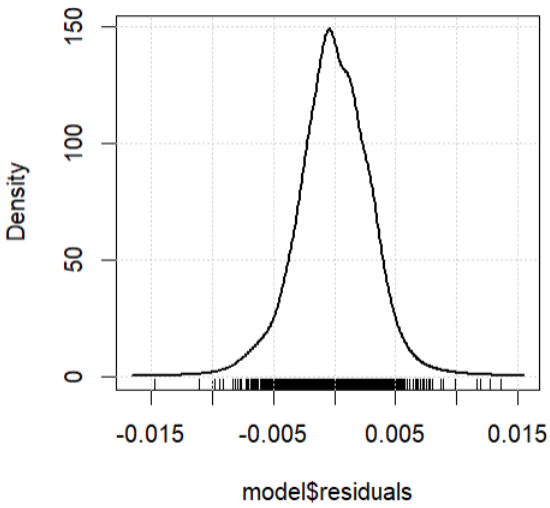


Figure 7: Residual error density distribution

As shown in Table 5, there are three outliers for the fitted model as far as training dataset is concerned.

Table 5: Outlier testing for the model prediction on the training dataset

	rstudent	unadjusted p-value	Bonferroni p
693	-4.836588	1.5154e-06	0.0016170
542	4.478820	8.3162e-06	0.0088734
543	4.221872	2.6298e-05	0.0280600

Figure 8 below shows further deep dive into residual. The points are not equally distributed throughout the range of fitted values indicating possible heteroscedasticity. In residuals vs leverage all points are within cook's distance indicating that there are no points that are high influencers.

Running NCV Test on the fitted model indicates the p-value of p = 0.0043022, which is low and so the null hypothesis that “there is no heteroscedasticity” is rejected. i.e. the model does exhibit heteroscedasticity.

Similarly, Breusch Pagan test as well as Standardized Breusch Pagan test show the p-value of 0.004302 and 0.03071. Both these are lower than 0.05 p-value and hence, null hypothesis that “the parameters are homoscedastic” is rejected. i.e. heteroscedasticity is present.

Durbin Watson test indicate p-value of 0.4699, which is significantly higher than 0.05. So the null hypothesis of “autocorrelation is not present” cannot be rejected. i.e. autocorrelation is absent. Similarly, Breusch-Godfrey shows statistic of 0.9405 which is significantly higher than 0.05. So the null hypothesis of “autocorrelation is not present” is failed to be rejected. i.e. autocorrelation is absent in the residuals.

Model validation

Since the model is found to be good fit and also adheres to the conditions needed for CLRM, the model is validated with the 20% test data that was separated in the previous step.

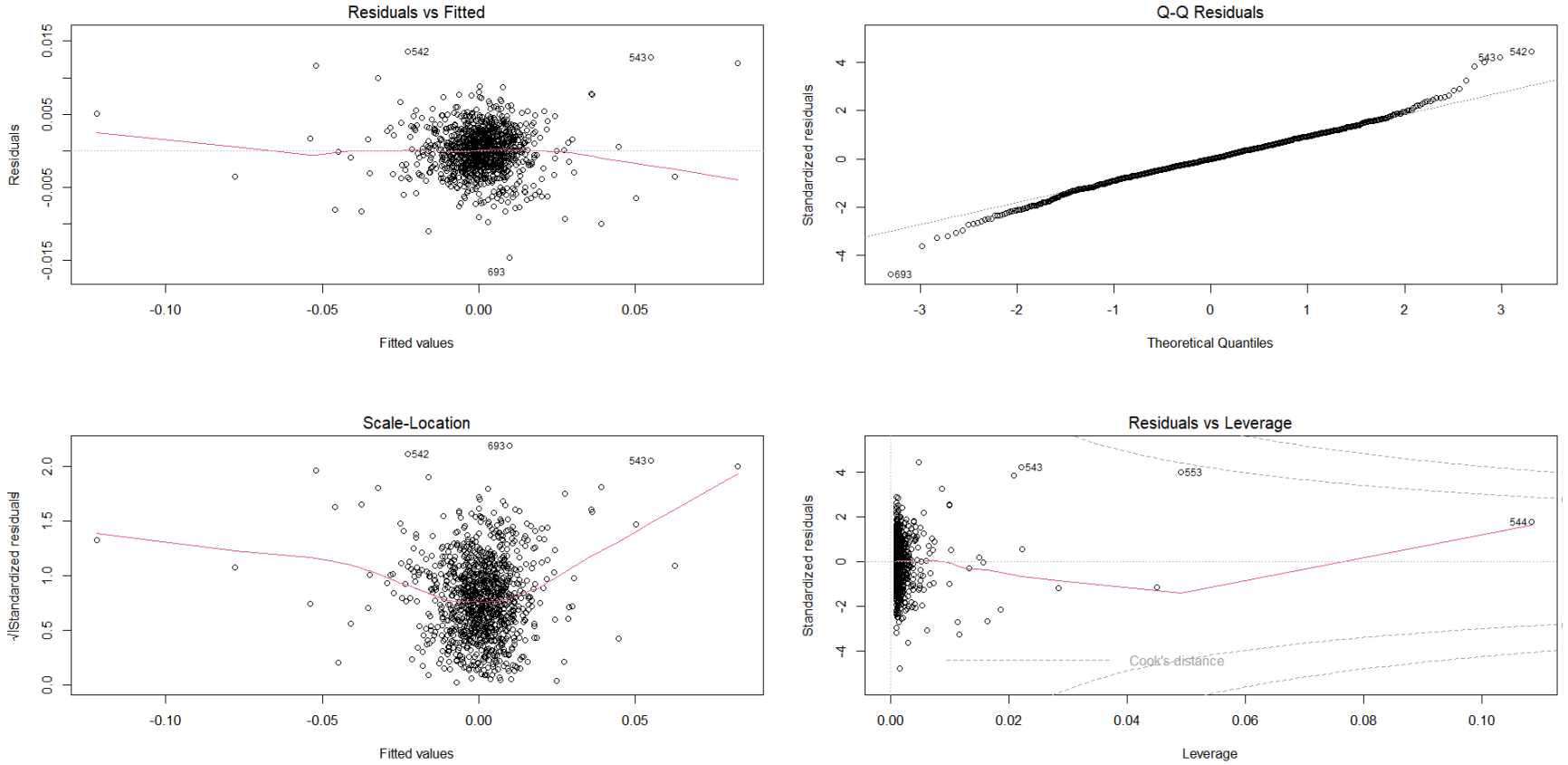


Figure 8: Residual analysis for the model

Figure 9 shows the actual vs predicted model performance and the predicted values closely resemble the actual values for the portfolio return. The correlation between the predicted and actual values was found to be 0.959, indicating highly correlated.

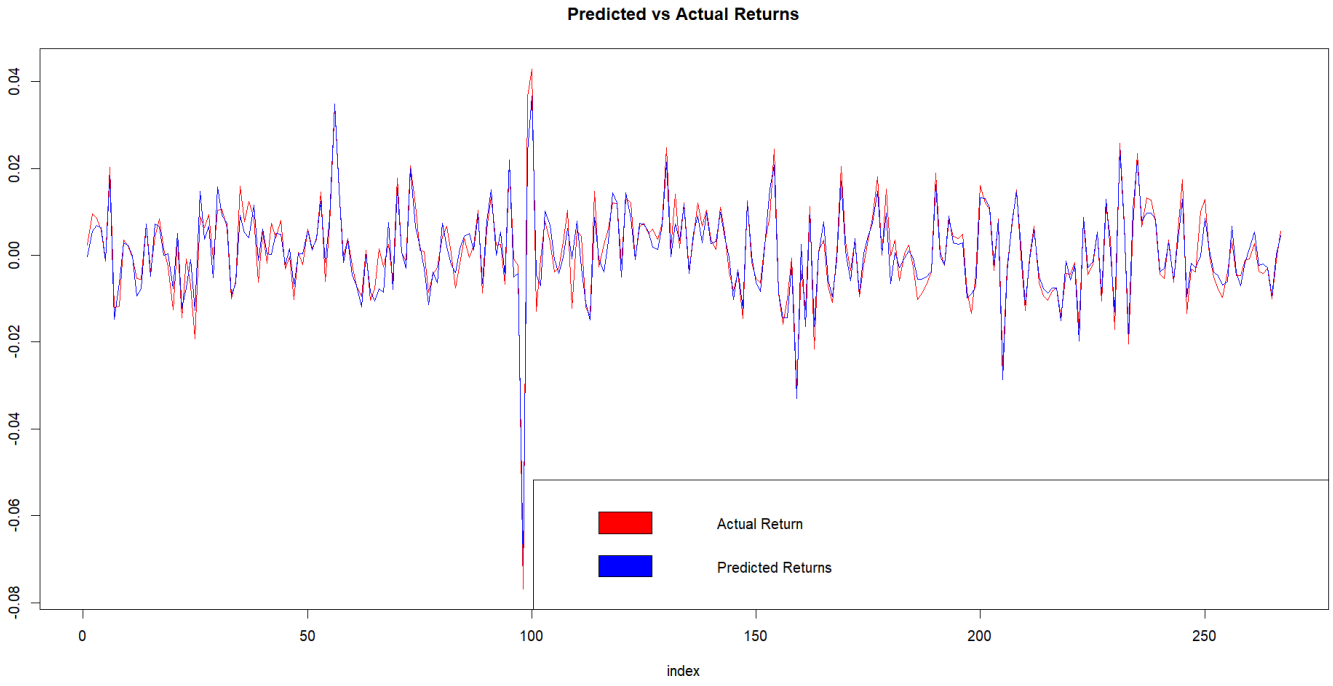


Figure 9: Actual vs predicted returns for the portfolio

Table 6 shows that as compared to the Naïve model, the model fitted with the Nifty50 as the predictor variable performs better on all the performance measures.

Table 6: Model performance on various error indicators

	Naive	Market
RMSE	0.01103904	0.003158856
SMAPE	1.75213311	0.571794751
RMSLE	0.01108788	0.003153441
Average	0.59142001	0.192702349

Classification Problem

Another linear model is created with the UpDown variable indicating whether the portfolio return is positive or negative. The resultant model indicates very low p-value and hence, it shows significantly correlated UpDown and Nifty50 index values.

This fitted model is then tested with the validation data, providing different treatment for the predicted values using different thresholds of 0.4, 0.5, 0.6, 0.7. i.e. if the fitted value is greater than the threshold, UpDown variable is considered to be true otherwise false.

Table 7: Model performance across Linear, Logit and Probit

Threshold	Accuracy	Sensitivity	Specificity	Class
<dbl>	<dbl>	<dbl>	<dbl>	<chr>
0.4	0.777	0.507	0.997	Linear
0.5	0.885	0.808	0.947	Linear
0.6	0.856	0.977	0.757	Linear
0.7	0.724	0.998	0.5	Linear
0.4	0.898	0.852	0.935	Logit
0.5	0.899	0.881	0.913	Logit
0.6	0.897	0.923	0.876	Logit
0.7	0.889	0.948	0.842	Logit
0.4	0.894	0.839	0.939	Probit
0.5	0.899	0.881	0.913	Probit
0.6	0.897	0.925	0.874	Probit
0.7	0.883	0.952	0.827	Probit

As seen in the Table 7, the model shows good performance across the thresholds. We observe that the model accuracy is consistently lower in Linear model compared to the Logit and Probit. For sensitivity and specificity, different models perform better for different values of threshold.

- For threshold of 0.4, Logit model has highest sensitivity, and Linear model has highest specificity
- For threshold of 0.5, Logit and Probit model have comparable sensitivity (higher than Linear model), and Linear model has highest specificity (with Logit and Probit model having comparable specificity
- For threshold of 0.6, Linear model has highest sensitivity, and Logit model has highest specificity (slightly higher than Probit model)
- For threshold of 0.7, Linear model has highest sensitivity, and Logit model has highest specificity

As shown in Figure 10, the performance of the models indicate that TPR and FPR for the models is almost similar across Linear, Logit and Probit models. The AUC for all Linear, Logit and Probit models is 0.9669, indicating that the model is able to make good prediction.

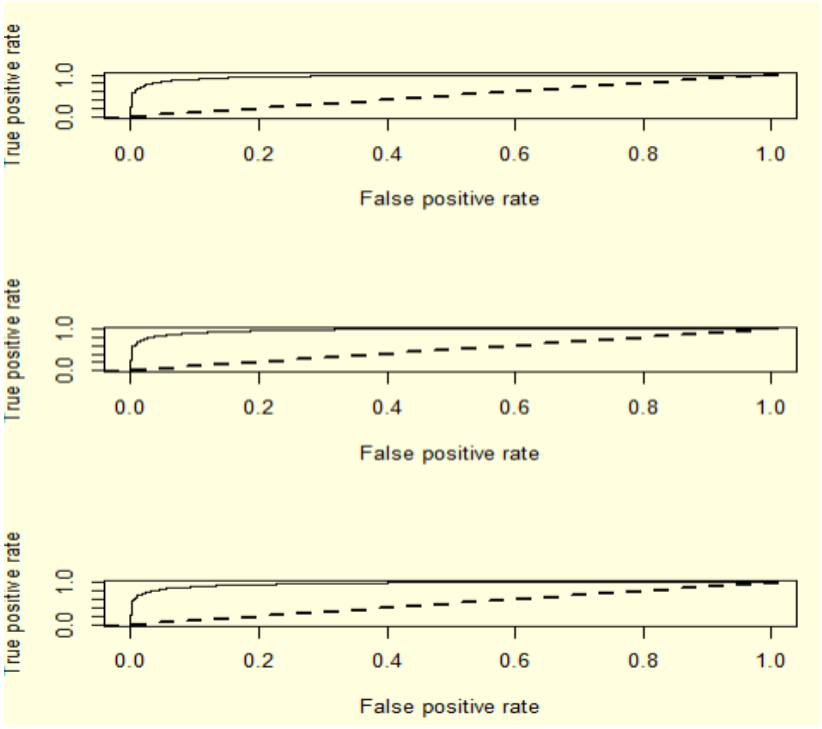


Figure 10: TPR vs FPR for various models (Linear, Logit and Probit)

Portfolio Rebalancing

We have also performed the analysis by considering portfolio rebalancing on monthly basis.

The key conclusions from this model are:

- There is not much difference in model performance with or without rebalancing, except for heteroscedasticity. The model with rebalancing performs slightly better.

- Multiple R-squared is slightly higher at 0.9337.

- For heteroskedasticity, using Breusch-Pagan Test, we find that p-value is greater than 0.05 so we failed to reject null hypothesis, and we conclude that there is no heteroscedasticity (unlike for equal-weighted portfolio without rebalancing, which showed heteroscedasticity).