

## To what extent can regression models predict future index prices?

### Introduction

Predicting future index prices is a crucial task for those working in the financial sector. Index prices are a major indicator of a country's current health and success. However, indexes on their own provide no indication of the future index price. The purpose of this report is to determine if a regression model can successfully forecast future index values based on historical index data.

Regression models are a type of machine learning algorithm that predict continuous outcomes based on a set of input data. They are especially suited to the application of index price prediction, where they have already been heavily utilised.

The main objectives of this report are to develop two models that can accurately predict future index prices for all indexes. Primarily to what extent can these models predict future index prices.

### Methodology and Dataset

<https://www.kaggle.com/datasets/mattiuzc/stock-exchange-data> provides three different data sets each with their own use. The data dataset is the original data collected on multiple indexes around the world. It is important to note that this dataset has not had any processing performed on it. Therefore, there could be null entries in the fields. The processed dataset, however, improves on the problems on the data dataset. Now, there is also an additional column as close price in USD which will provide a meaningful attribute for prediction. Additionally, the processed dataset has had null entries removed. Finally, the info dataset is a detail dataset simply storing further information to each of the indexes. All the columns in this dataset have no relevance to the regression models thus, omitted. To summarise, here is a set of tables stating the attributes for each dataset.

Included within the data dataset are the following attributes.

Attribute	Description
Index	Ticker Symbol for Indexes
Date	Date of Observation
Open	Opening Price
High	Highest price during trading day
Low	Lowest price during trading day
Close	Close price adjusted for splits
Adj Close	Adjusted close price adjusted for both dividends and splits
Volume	Number of shares traded during trading day

Included within the processed dataset are the following attributes.

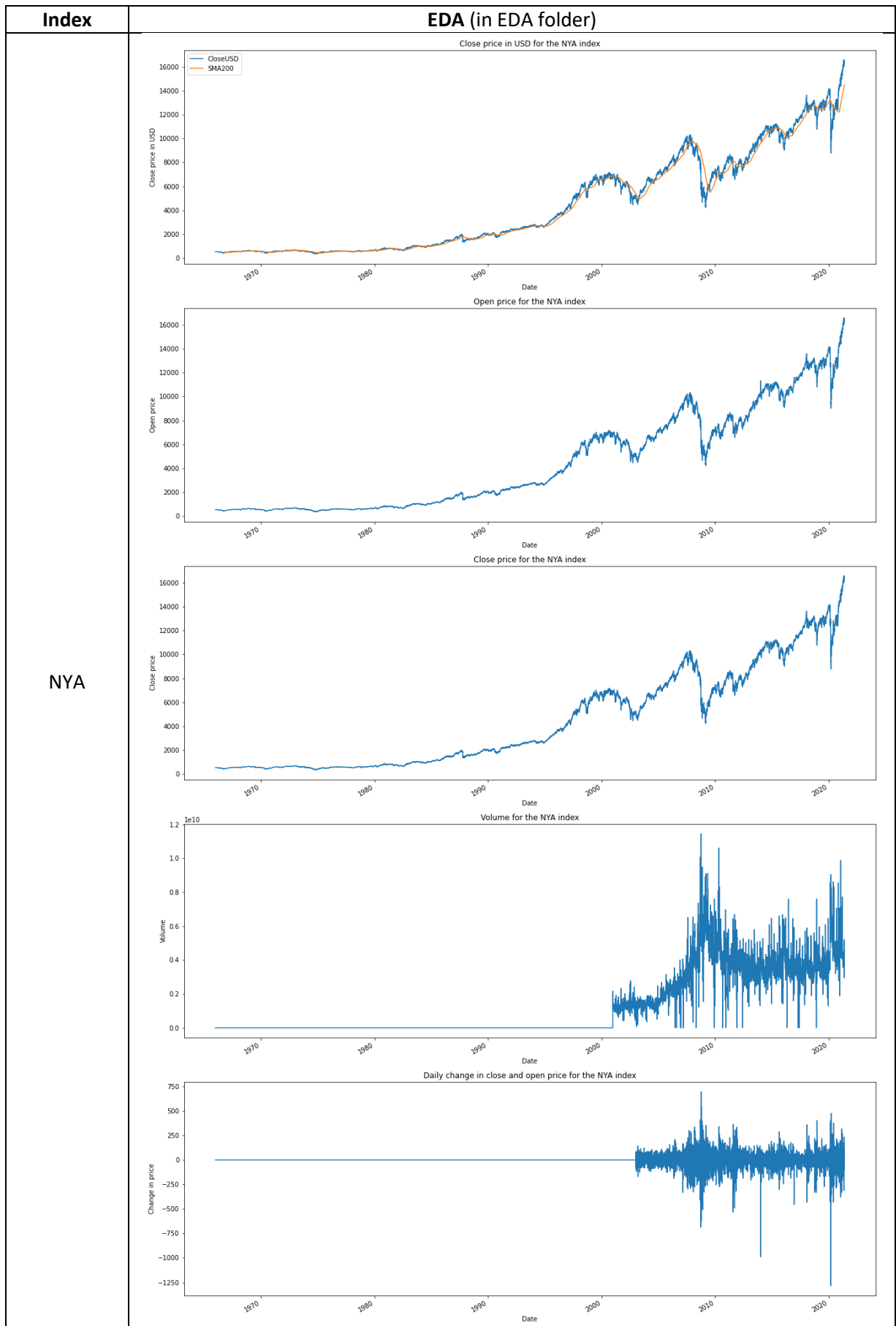
Attribute	Description
Index	Ticker Symbol for Indexes
Date	Date of Observation
Open	Opening Price
High	Highest price during trading day
Low	Lowest price during trading day
Close	Close price adjusted for splits
Adj Close	Adjusted close price adjusted for both dividends and splits
Volume	Number of shares traded during trading day
CloseUSD	Close price in terms of USD

Included within the info dataset the are following attributes.

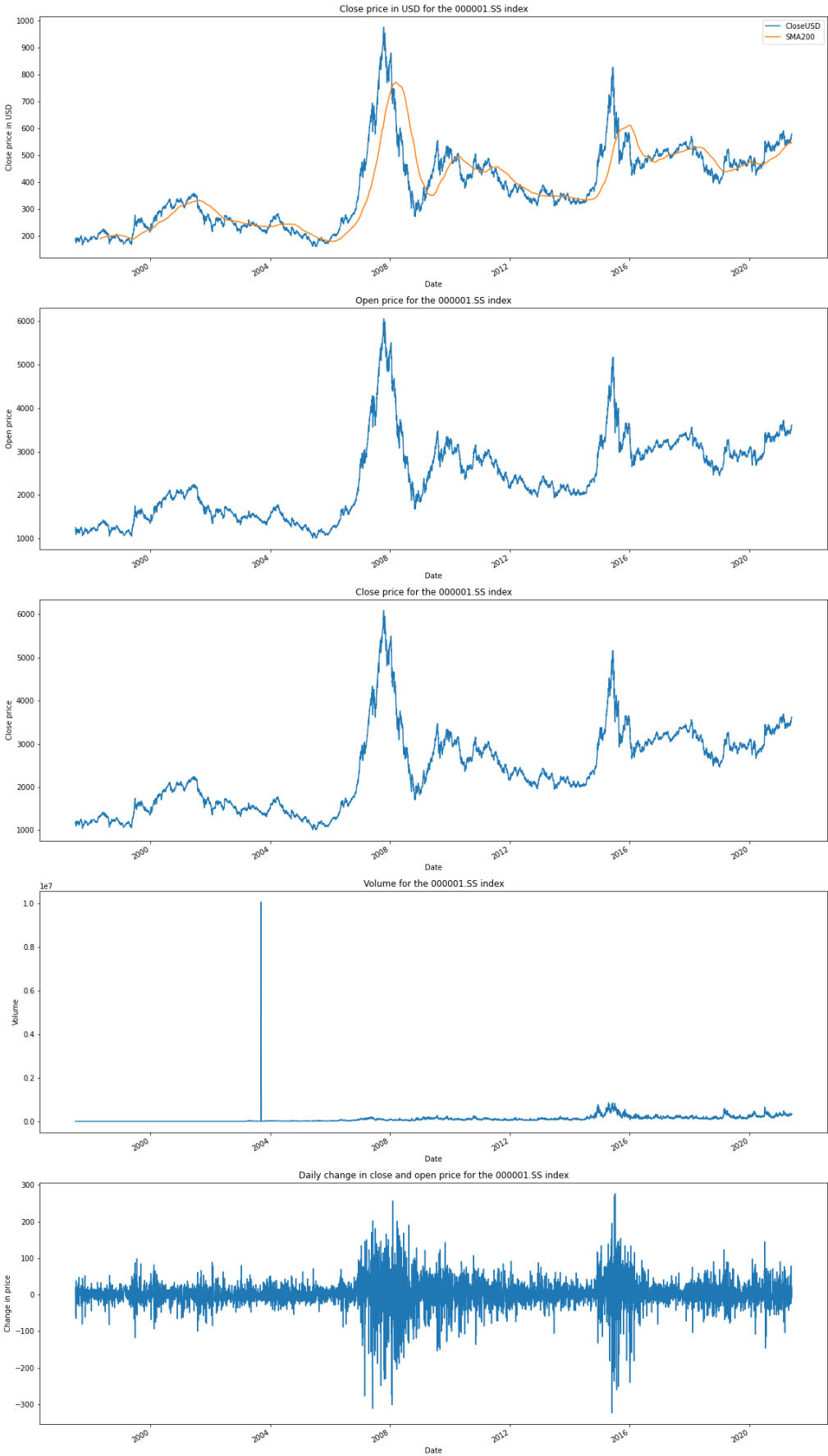
Attribute	Description
Region	Country the exchange is located in
Exchange	Name of the stock exchange
Index	Ticker of exchange index
Currency	Currency the index is quoted in

The processed dataset is the primary dataset for this project, due to there being no null entries but more importantly the addition of the CloseUSD attribute. Before any cleaning, there are 104224 entries but also a varying number of entries for each index. For example, the J203.JO index has 2346 entries while NYA has 13947 entries. This could prove to be a major problem in fitting the indexes. Furthermore, for all indexes the volume initially remains at zero for years since what is presumably the exclusivity of trading at the time. Also, due to unknown reasons a substantial portion of data for each index has the same opening, high, low, close and adj close value. However, just from speculating, it is possible that the attributes were not recorded at that time.

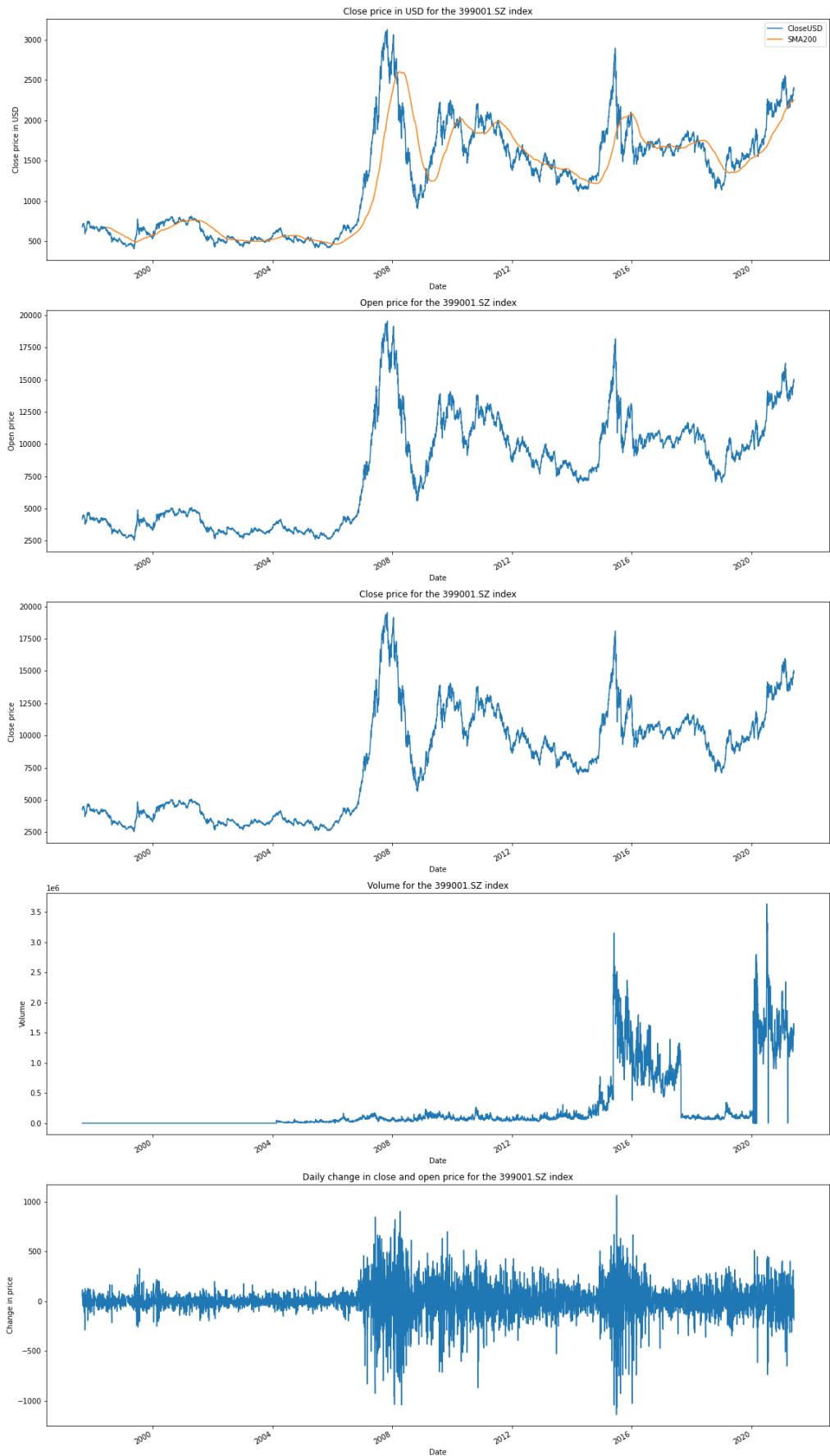
It is evident that the model would predicting prices using the dates. Using attributes where for a sizeable proportion of which are zero will negatively affect the model. Thus, no other attributes will fit with the model. To prepare for modelling, I use a new attribute, days since epoch. This takes the dates as a datetime subtracted with the Unix epoch time, 1<sup>st</sup> January 1970, converted into days. Now I have a quantitative representation of the date rather than the datetime object. In addition, a new attribute, change, completes the data. Change is the difference between the close and open attribute, representing the daily fluctuations in index prices. Finally, simple moving average (SMA) is a widely used technical indicator for trends in finance (Hayes, 2022). Therefore, the SMA of the past two hundred data points (not necessarily days) is found and plotted alongside the CloseUSD. At this point, an exploratory data analysis (EDA) of the data, more specifically, for each index is ran.



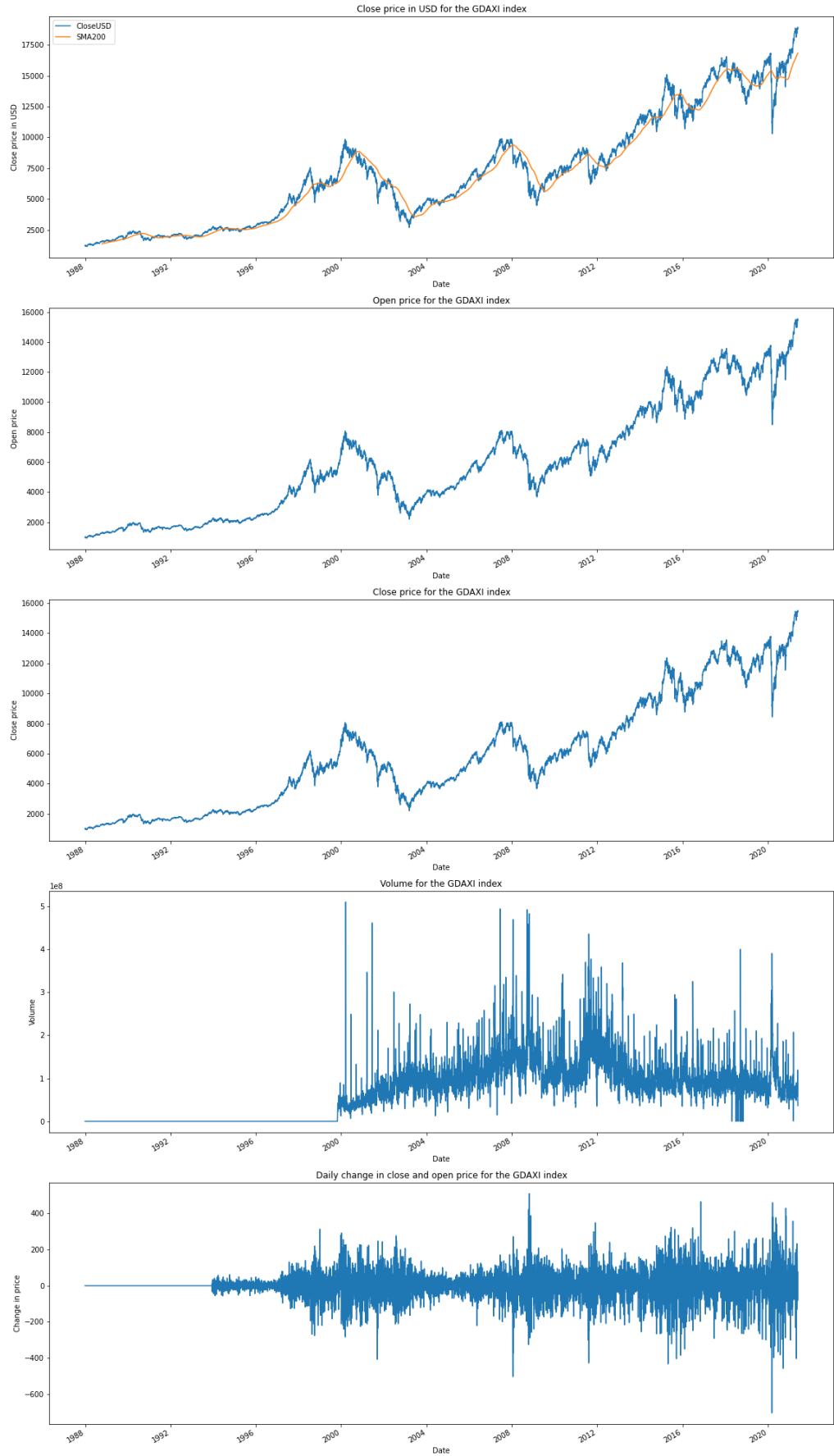
000001.SS



399001.SZ



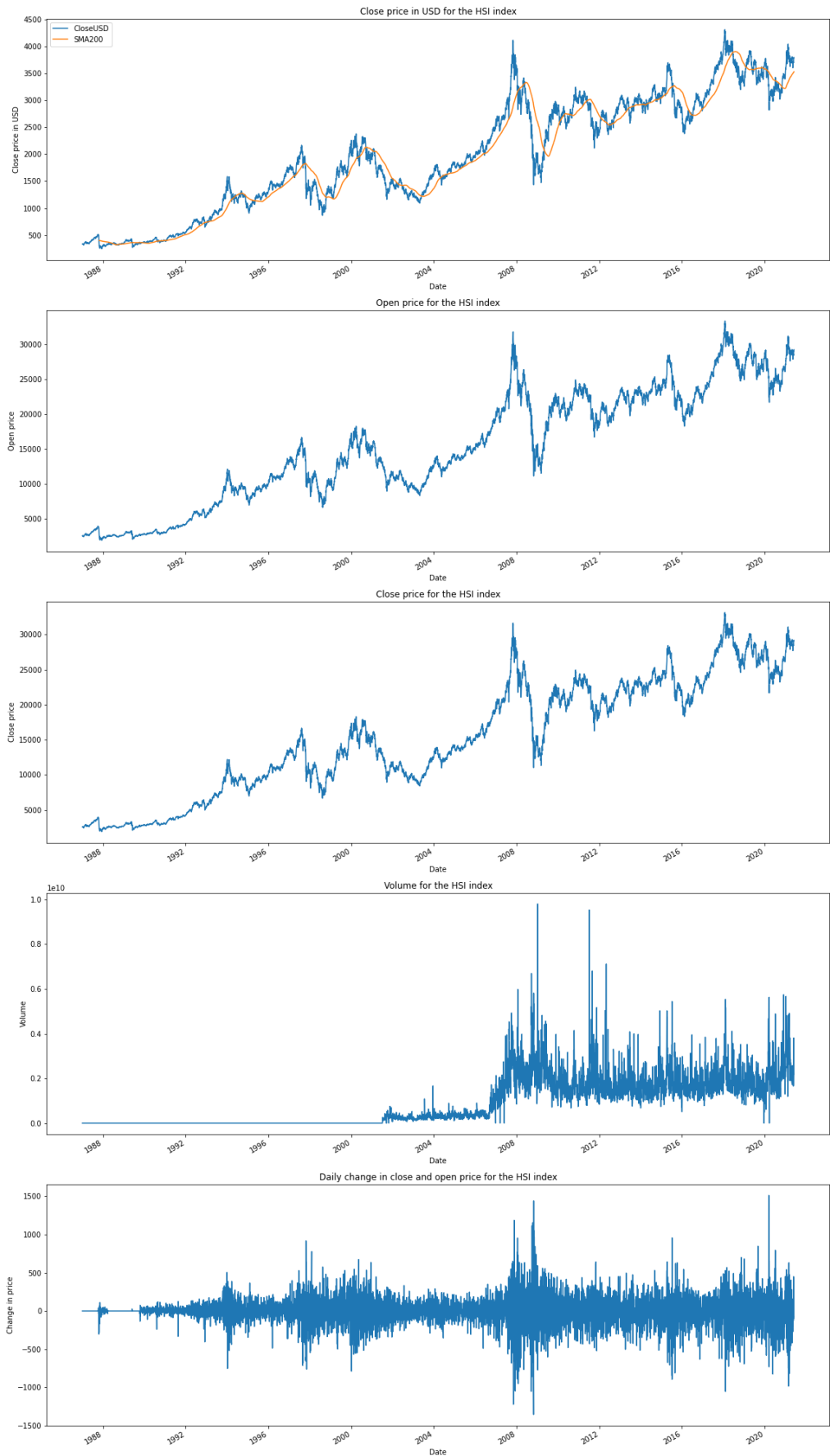
## GDAXI



GSPTSE

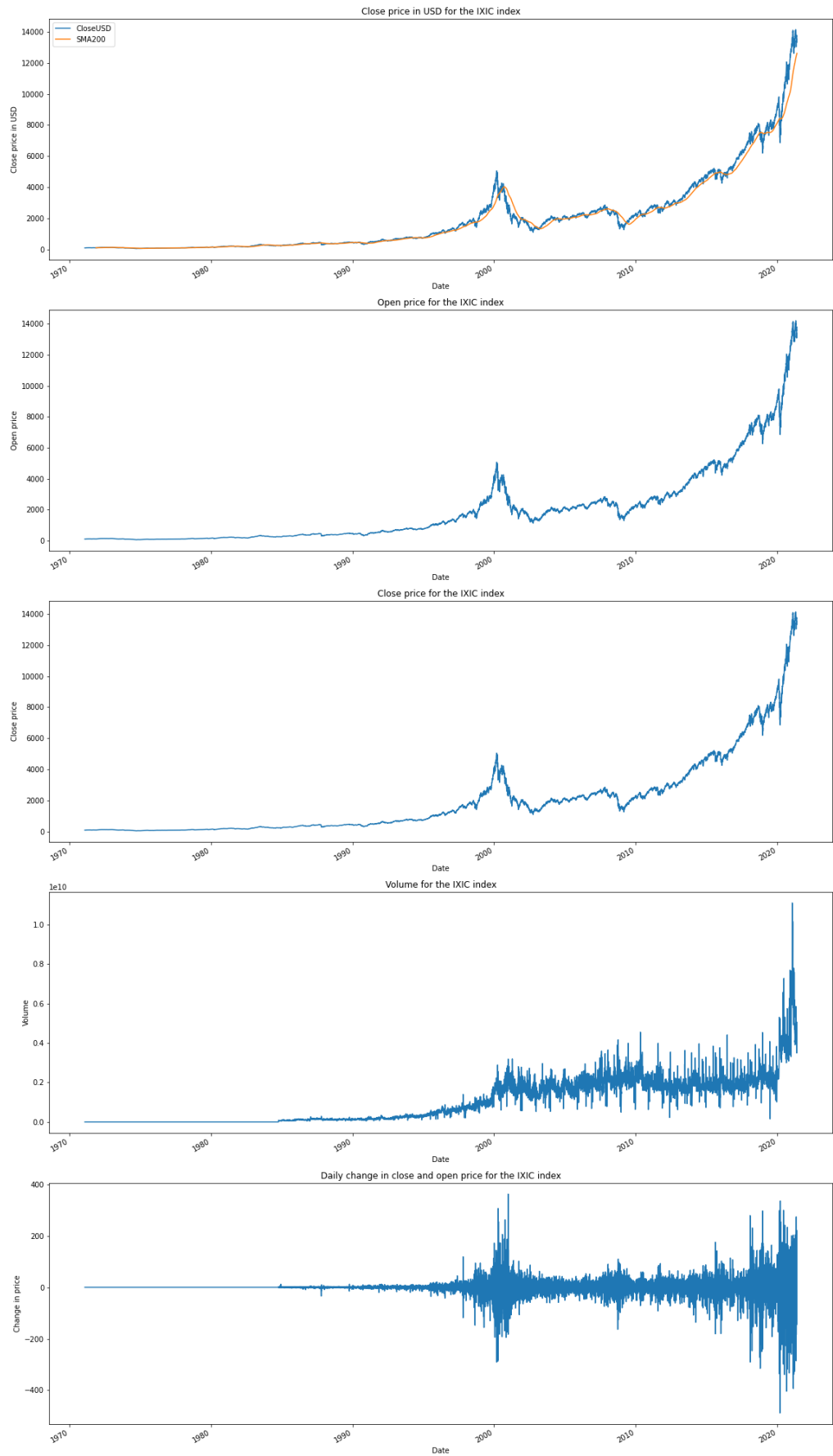


HSI

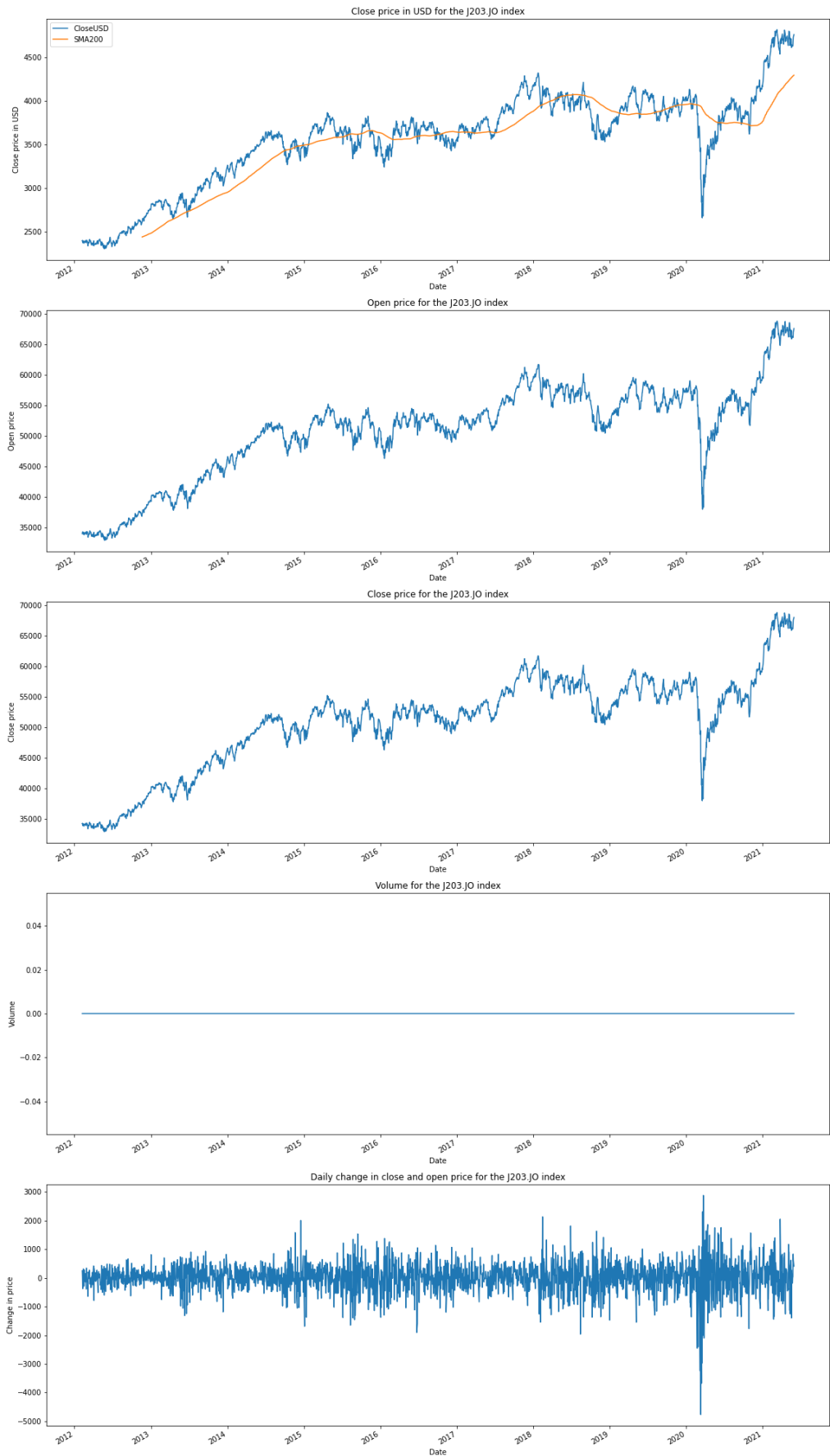




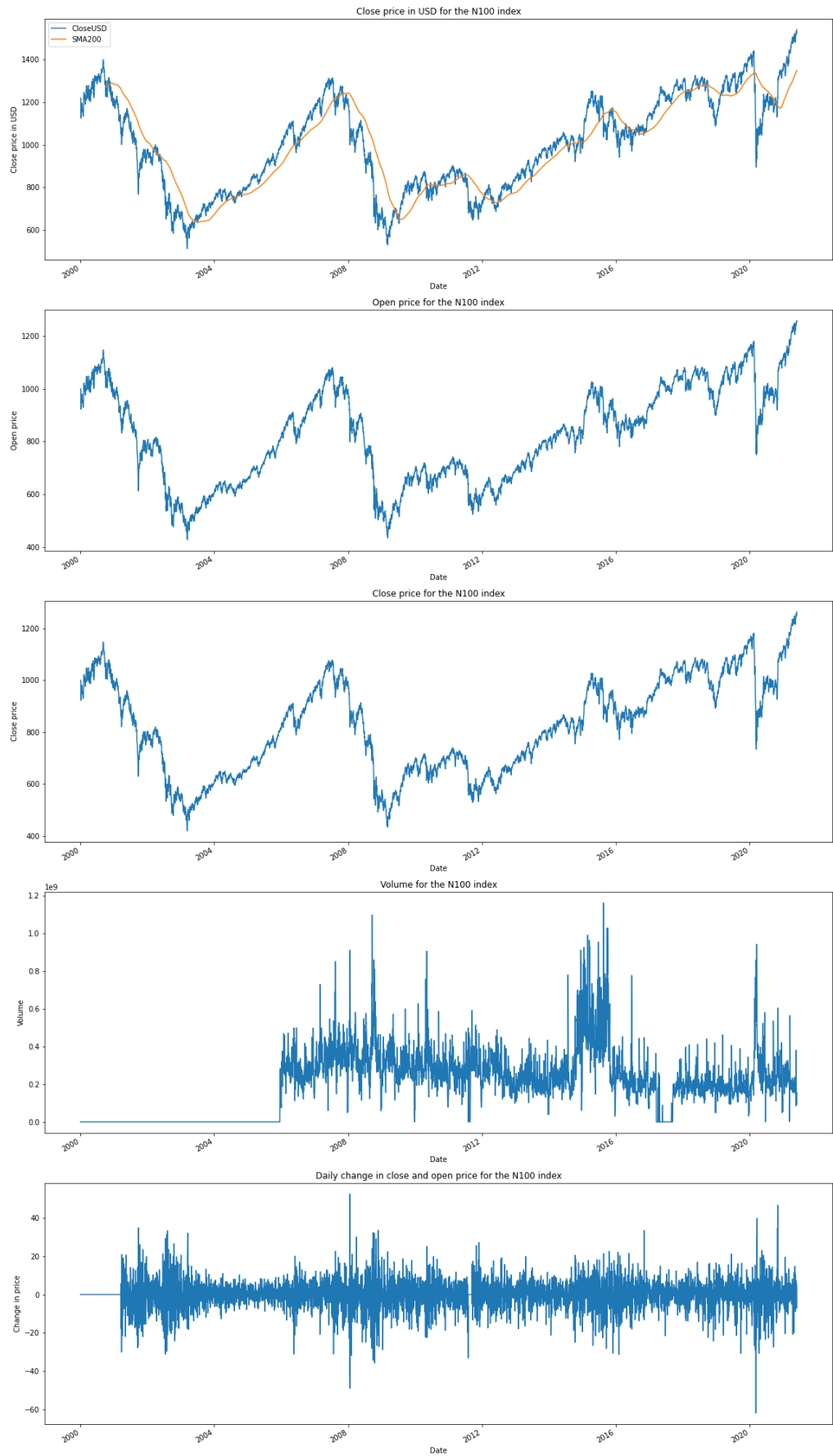
IXIC



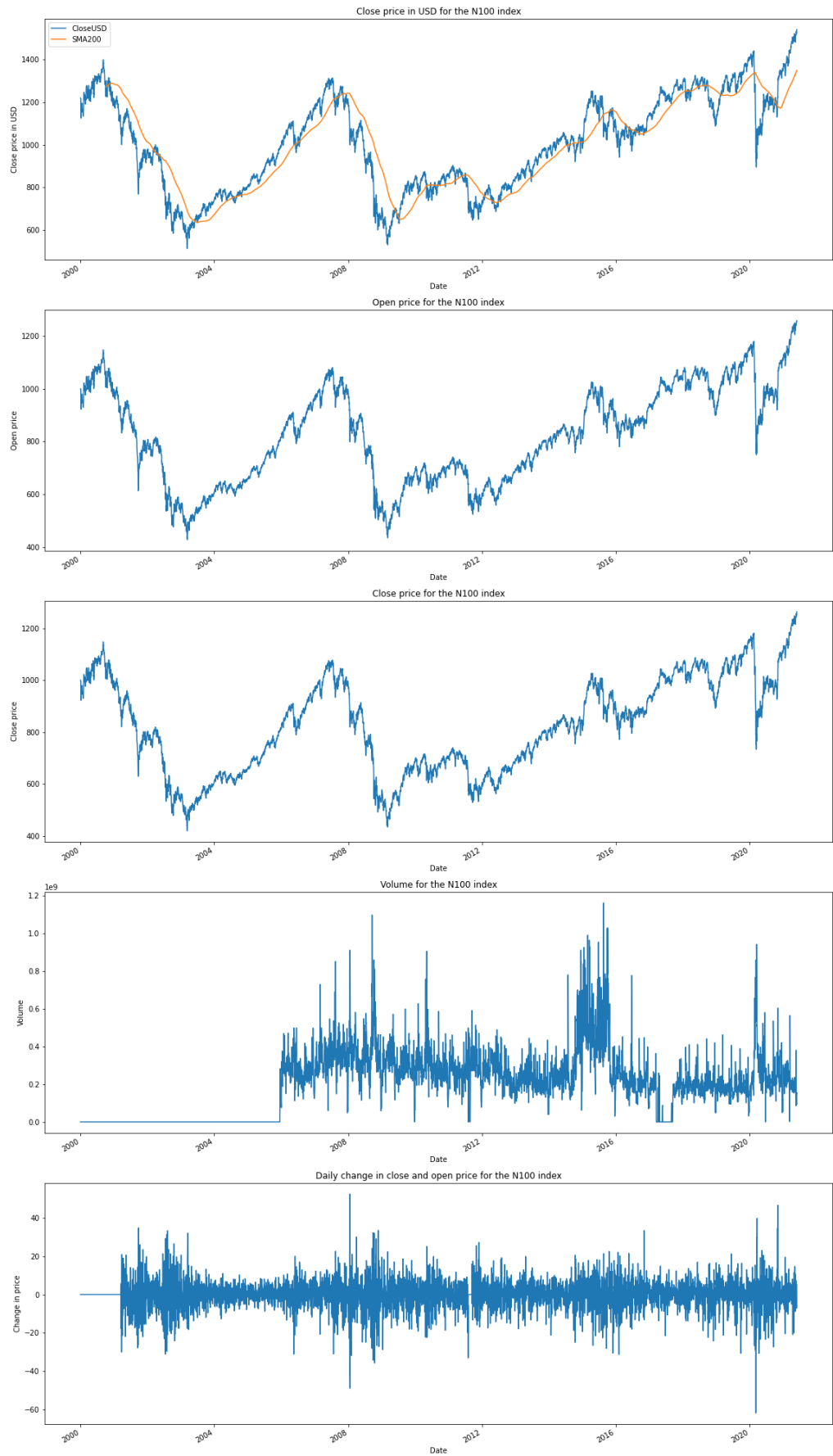
J203.JO



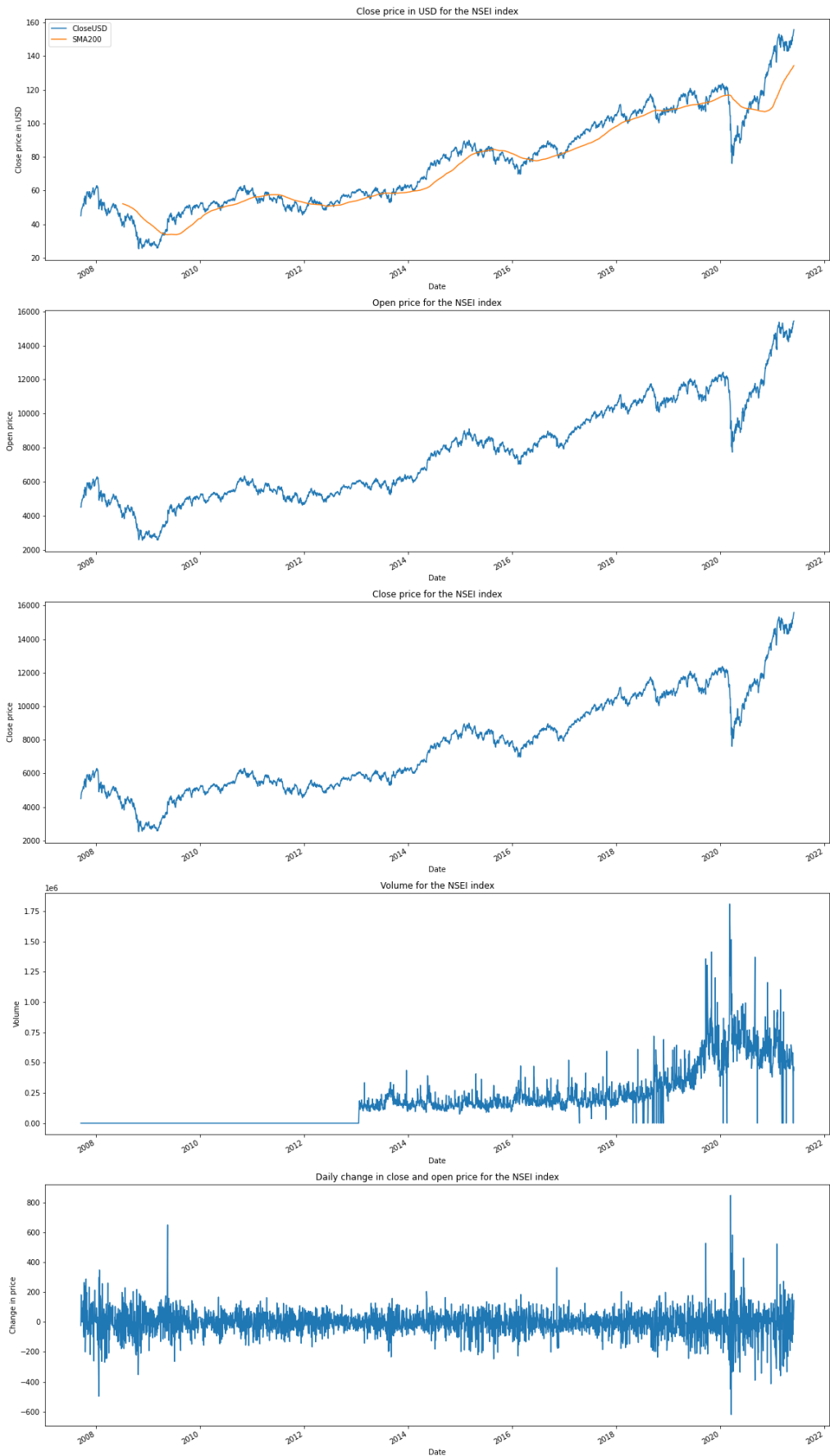
N100



N225



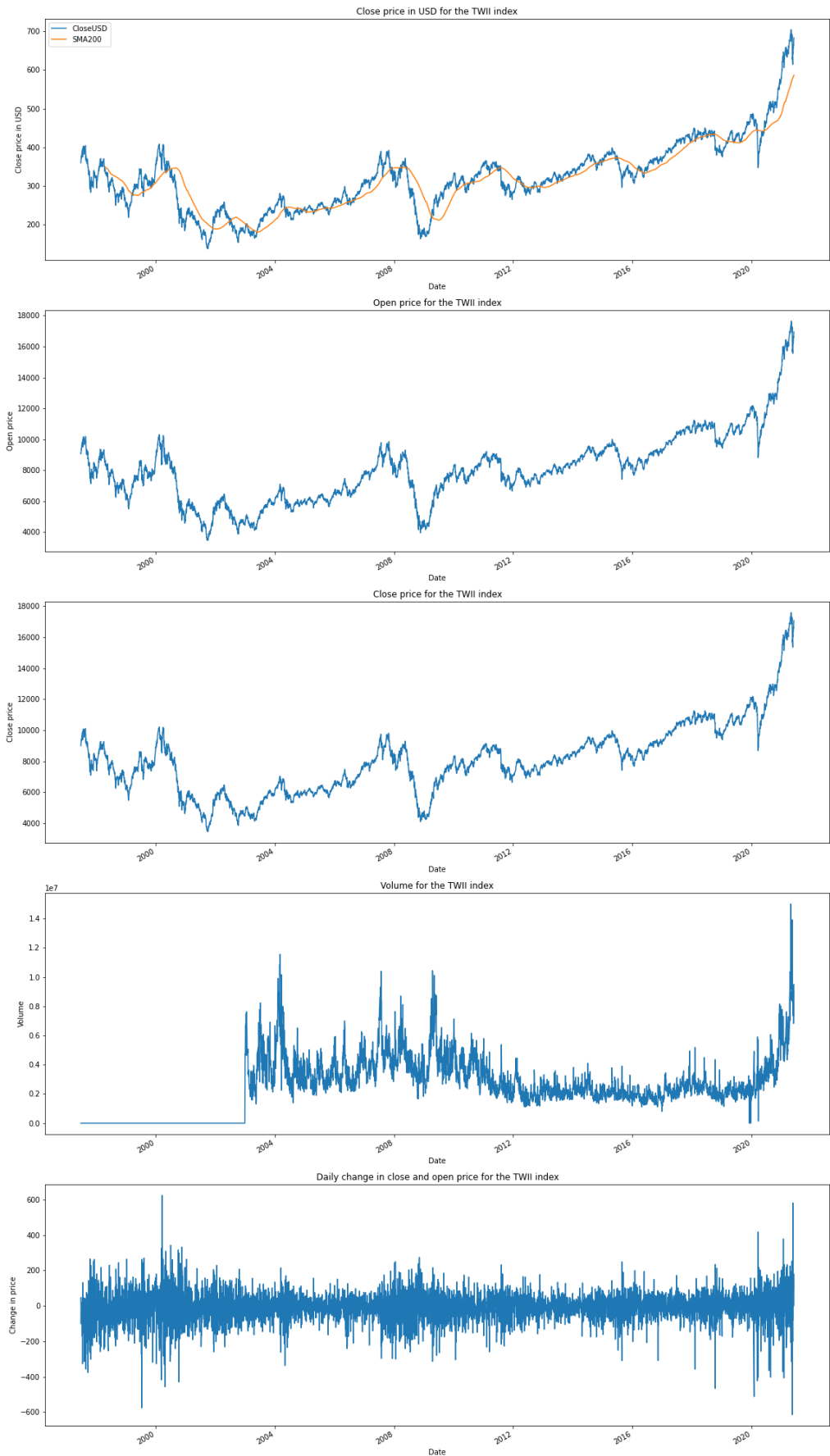
NSEI



SSMI



TWII



The regression models used are linear regression (LR) and polynomial regression (PR). At this point no further pre-processing is required, since the x and y data required have already been defined or calculated beforehand.

Three metrics assist in evaluating a model. These include; R-squared score ( $R^2$ ), mean absolute error (MAE) and mean squared error (MSE). For simplicity, only  $R^2$  determines how well a model has performed. This is essential, especially in PR where the hyper parameter, degree, must be tuned.  $R^2$  represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model (Fernando, 2021). Simply put, a higher  $R^2$  score is a better model, anything above 70% (Fernando, 2021) in finance is acceptable.

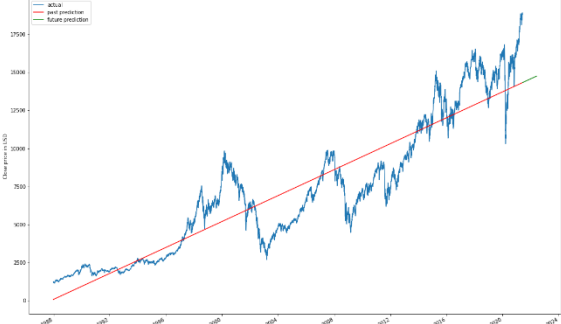
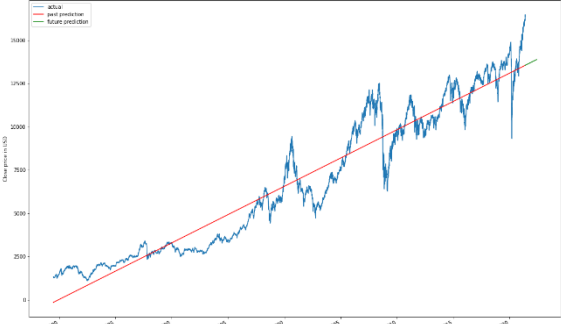
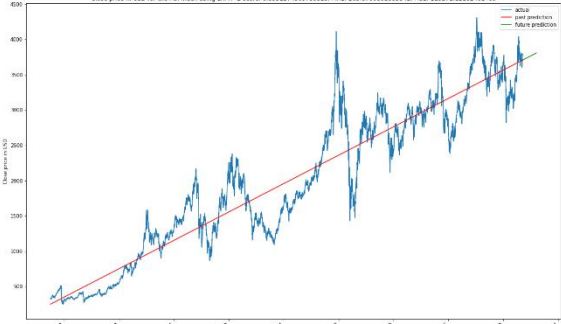
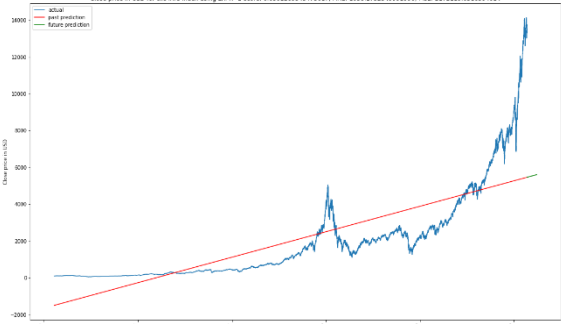
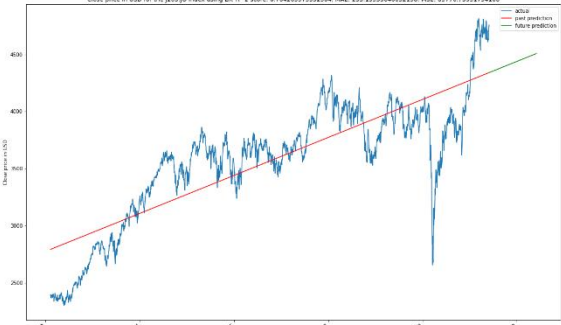
## Results




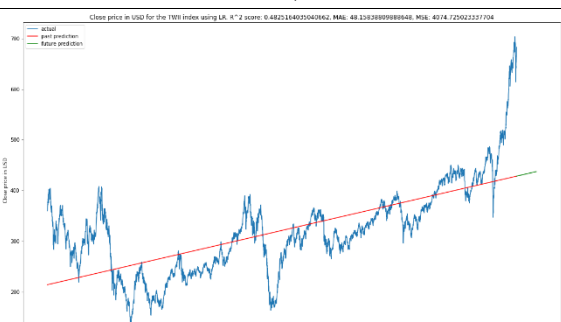
### Linear regression

The first model to be trained on is the LR. More specifically simple LR since only one variable is used. Initially, the LR model was more specifically multiple LR as the data to fit into the model where three independent variables as day, month and year. However, this did not improve the model in terms of  $R^2$ , therefore, the simple LR model, where x is the days since epoch, was favoured more.

Index	LR (in LR folder)
NYA	
000001.SS	
399001.SZ	



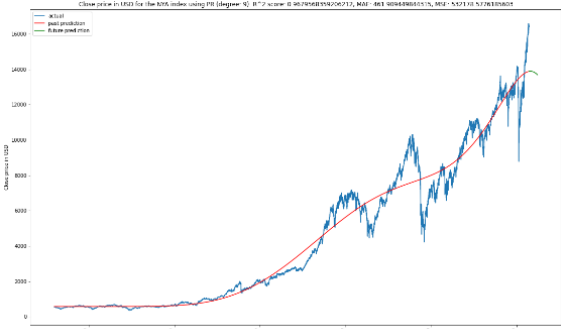
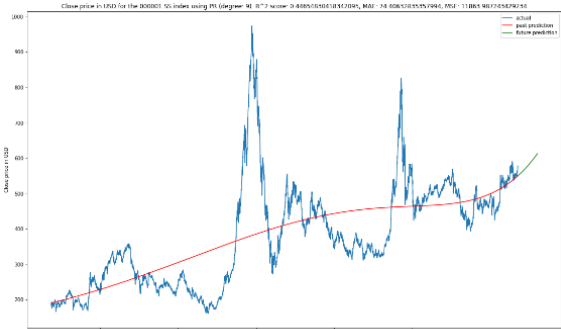
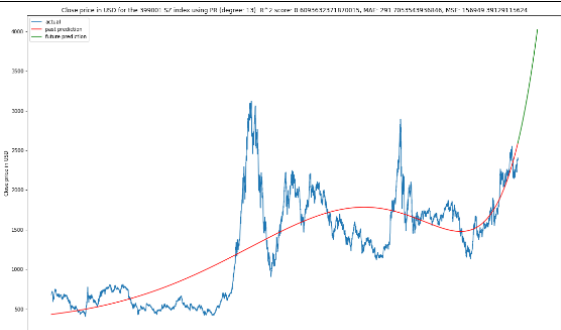
GDAXI	<p>Close price in USD for the GDAXI index using UK R<sup>2</sup> score: 0.85248524851688, MAE: 1395.4512059114226, MSE: 2980529.958299222</p> 
GSPTSE	<p>Close price in USD for the GSPTSE index using UK R<sup>2</sup> score: 0.930903892218296, MAE: 888.6726813822512, MSE: 1179620.549248888</p> 
HSI	<p>Close price in USD for the HSI index using UK R<sup>2</sup> score: 0.892327450979925, MAE: 245.8704818823461, MSE: 113271.2232492489</p> 
IXIC	<p>Close price in USD for the IXIC index using UK R<sup>2</sup> score: 0.825289545472617, MAE: 1028.1752249992995, MSE: 2172119.623279854</p> 
J203.JO	<p>Close price in USD for the J203.JO index using UK R<sup>2</sup> score: 0.10426975252504, MAE: 225.15333640922196, MSE: 82776.7391374338</p> 

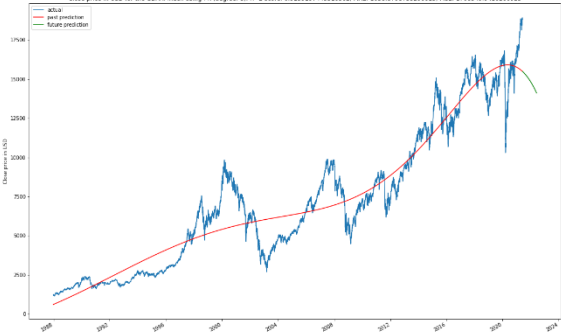
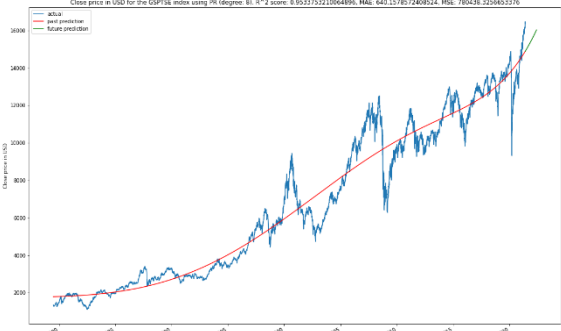
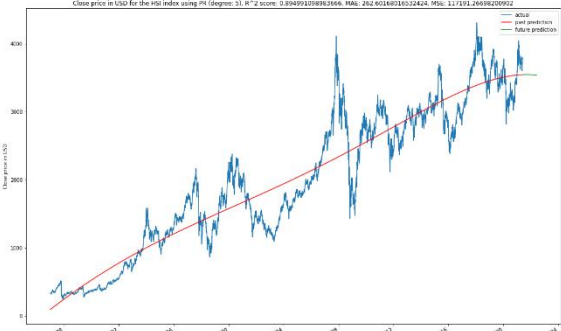
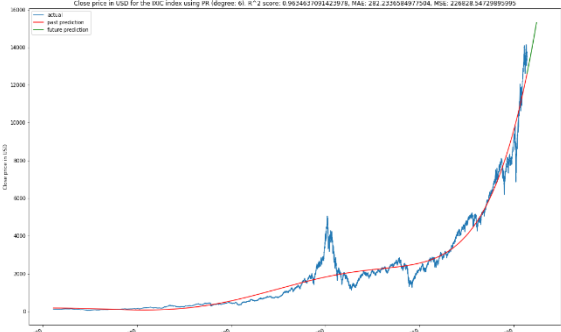
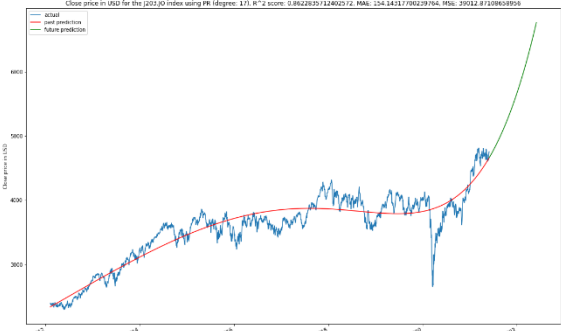
N100	<p>Close price in USD for the N100 index using LR, R<sup>2</sup> score: 0.18992847527782788, MAE: 167.225988078522, MSE: 39686.214080225</p> 
N225	<p>Close price in USD for the N225 index using LR, R<sup>2</sup> score: 0.2426095958603235, MAE: 48.612880701151038, MSE: 4234.911738992843</p> 
NSEI	<p>Close price in USD for the NSEI index using LR, R<sup>2</sup> score: 0.872867707142097, MAE: 7.82968306438705, MSE: 131.0481220488972</p> 
SSMI	<p>Close price in USD for the SSMI index using LR, R<sup>2</sup> score: 0.739052239974449, MAE: 1185.938173851323, MSE: 212498.987823823</p> 
TWII	<p>Close price in USD for the TWII index using LR, R<sup>2</sup> score: 0.4825184055040962, MAE: 48.15828889888848, MSE: 4074.725022327704</p> 

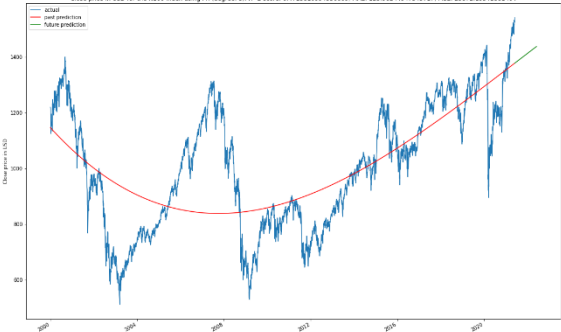
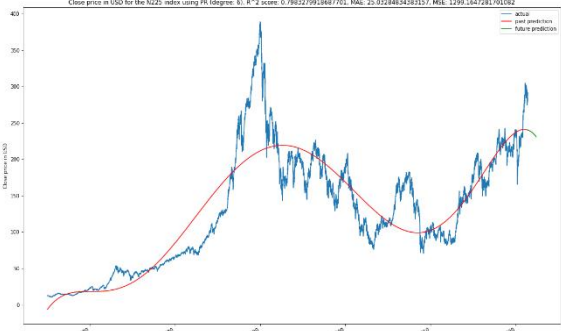
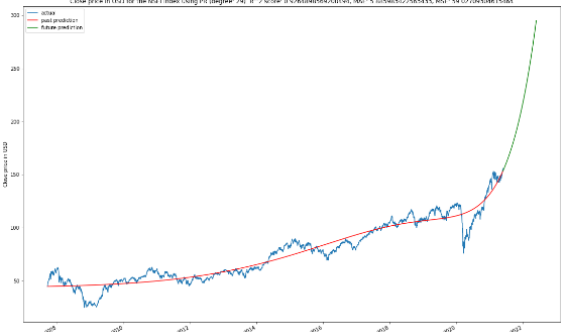
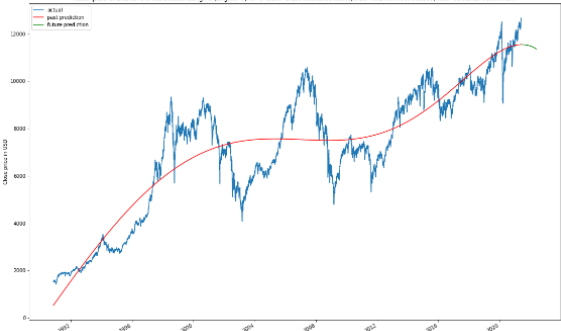
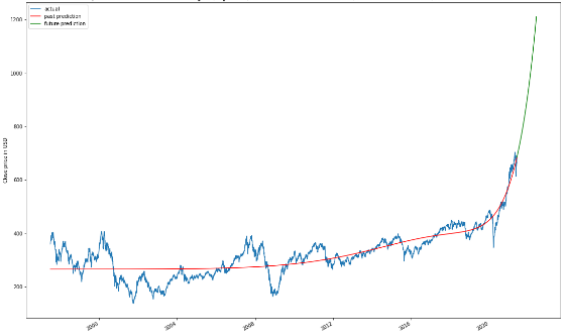
Above is a table of each index and it's respective LR model of date against actual value, historical predicted value and future predicted value. It is clear just from the diagrams, that for the most part indexes do not fit well as a simple LR model. Therefore, 7 of the 13 indexes have an acceptable  $R^2$  score using LR. The best  $R^2$  for simple LR was found in the GSPTSE index, with a score of 93.0%.

### Polynomial regression

The second model to be trained on is the PR. There was an attempt to fit these models with the day, month and year but this began to take far too long especially when degrees higher than twenty were being used. At this point, having multiple independent variables for PR was not considered. Again, the days since epoch was used and this showed great improvements in fitting speed and accuracy.

Index	PR (in PR folder)
NYA	
000001.SS	
399001.SZ	

GDAXI	<p>Close price in USD for the GDAXI index using PH (degree: 0). R<sup>2</sup> score: 0.382089/448233282. MA6: 1036.879873292023. MS6: 1706349.943266923</p> 
GSPTSE	<p>Close price in USD for the GSPTSE index using PH (degree: 0). R<sup>2</sup> score: 0.953752210064896. MA6: 640.15781240934. MS6: 780436.325965376</p> 
HSI	<p>Close price in USD for the HSI index using PH (degree: 3). R<sup>2</sup> score: 0.89499109893466. MA6: 262.60348030532424. MS6: 1.173312468200932</p> 
IXIC	<p>Close price in USD for the IXIC index using PH (degree: 6). R<sup>2</sup> score: 0.9624637001423976. MA6: 282.223584677304. MS6: 228628.3472989395</p> 
J203.JO	<p>Close price in USD for the J203.JO index using PH (degree: 17). R<sup>2</sup> score: 0.8622635712402972. MA6: 154.14137798239764. MS6: 20032.87330638956</p> 

N100	<p>Close price in USD for the N100 index using PM (degree: 3), <math>R^2</math> score: 0.471281099420889, MAE: 123.96244047242717, MSE: 25872.22541881404</p> 
N225	<p>Close price in USD for the N225 index using PM (degree: 3), <math>R^2</math> score: 0.798227993887702, MAE: 21.03288834283237, MSE: 1299.1847281701062</p> 
NSEI	<p>Close price in USD for the NSEI index using PM (degree: 29), <math>R^2</math> score: 0.926888884828164, MAE: 5.037888327384835, MSE: 58.0722848475384</p> 
SSMI	<p>Close price in USD for the SSMI index using PM (degree: 2), <math>R^2</math> score: 0.827857285153354, MAE: 528.788104718811, MSE: 1796281.137382979</p> 
TWII	<p>Close price in USD for the TWII index using PM (degree: 25), <math>R^2</math> score: 0.774888820837104, MAE: 58.0388888888889, MSE: 2758.03888888889</p> 

Above is a table of each index and its respective PR model of date against actual value, historical predicted value and future predicted value. It is clear that for many of the indexes, polynomial fits more closely to the actual data. Therefore, 10 of the 13 indexes have an acceptable  $R^2$  score using PR. The best  $R^2$  score being the NYA index with a score of 96.8%.

### Discussion

Index	LR	PR
NYA	87.5%	96.8%
000001.SS	42.5%	44.7%
399001.SZ	50.3%	60.9%
GDAXI	85.0%	91.4%
GSPTSE	93.0%	95.3%
HSI	89.3%	89.5%
IXIC	65.0%	96.3%
J203.JO	70.4%	86.2%
N100	18.9%	47.1%
N225	34.3%	79.8%
NSEI	87.2%	92.6%
SSMI	70.9%	82.3%
TWII	48.3%	72.6%

Figure 1: Model prediction score based on  $R^2$  score.

From the analysis above, PR performs far better than LR, in all the indexes. The historical predictions made by the PR model fits well with most indexes to an acceptable degree, especially when tuning the hyperparameter degree.

However, this report aims to determine the extents at which regressive models can predict future prices. Therefore, the model has also predicted up to one year from the last historical closing price. Normally, PR does not perform very well in future predictions. This may be due to the nature of the model. PR's susceptibility to outliers. Although, PR could be a good indication for future index prices in the relatively immediate future, like one month maximum into the future. This conclusion has been determined, as going further into the future for example 5+ years tends to lead to a negative index price or an unrealistically high index price. Indexes showing the trend of an unrealistically high index prices include; TWII, NSEI, J203.JO and 399001.SZ.

On the other hand, LR simply cannot accurately predict index data whether it is in the immediate future or further into the future. As seen, LR has an extremely poor  $R^2$  score for most of the indexes, with only a handful achieving a score of 85% or more as seen in Figure 1. But there may be a usage of LR in determining trends. A model with a high coefficient (gradient) will generally result in a larger return for a person regardless of the amount of time it takes.

To summarise, this report has shown that there is a potential for use of LR and PR to predict future index prices. However, LR should not determine immediate or near future index prices but rather determine the general trend of an index. As with PR, there is evidence it can determine immediate or near future index prices. Although, its usage to determine long term trends should not be considered.

### References

Fernando, J. (2021, September 12). *R-squared formula, regression, and interpretations*. Retrieved from Investopedia: <https://www.investopedia.com/terms/r/r-squared.asp>

Hayes, A. (2022, February 1). *Simple moving average (SMA): What it is and the formula*. Retrieved from Investopedia: <https://www.investopedia.com/terms/s/sma.asp>