# EDA

* Data recorded in **6 years**, from Feb, 2018 to March, 2023
* **HPG closing price** ranging from 7,411 VND to 43,895 VND, mean = 19,249 VND and standard deviation = 9,780 VND
* **Trading volume** ranging from 2mil/day to 100mil/day, mean = 24 mill/day and standard deviation = 15 mill/day



**Explaining price & volume:**

* HPG starts at 12,000 VND and is in a downtrend, entering a **bullish falling wedge pattern** from 2018 to March 2020, which suggests a **potential upside breakout**. This is backed up by **rising RSI lows**, which suggests **increasing buying strength** despite the falling price.
* HPG eventually breaks out of the falling wedge's upper trendline, entering an uptrend. Both moving averages of 50 and 200 days act as supporting levels. MACD showing bullish divergence (rising highs) and RSI cross 70 numerous times, often stays over 50 and rarely dip to 30. This confirms the trend reversal into a bullish phase, from March 2020 to March 2021.
* March 2021 to May 2022 shows **a head and shoulders chart pattern**, which is a bearish reversal pattern. Once the price broke below the neckline (32,000 VND in May 2022), a **downtrend followed**. MACD show bearish divergence, confirming this trend reversal.
* 2023: Post the significant downtrend, the stock price has stabilized, showing less volatility.



* Trading Volume:
  + **High Volatility Periods**: There are significant spikes in trading volume at various points, indicating periods of high activity. Notable peaks can be observed around mid-2018, early 2021, and mid-2023.
  + **Low Volatility Periods**: Conversely, there are stretches where the trading volume is relatively low, such as the latter part of 2019 and early 2020.
  + **Overall Trend**: The trading volume appears to have an overall increasing trend with more frequent and higher peaks towards the later years.
* Bollinger Bands:
  + **Wider Bands**: Periods where the bands are wider, such as from 2020 to 2022, indicate higher price volatility.
  + **Narrower Bands**: Periods where the bands are narrower, such as before 2019, indicate lower price volatility.

# 2. Models to forecast the stock price

## 2.1. ARIMA

First, we test for stationary and the result is non-stationary. After differencing the first time data becomes stationary, suggested d=1. Then we identify MA(q) and AR(p), we get q=2, p=2. But the output is apparently not a good fit for prediction. Therefore we apply hypertune.

| Model | RMSE | R^2 | MAPE |
| --- | --- | --- | --- |
| SARIMAX(2, 1, 2)(2, 1, 2, 5) | 19336.4394927638 | -8.2787198191 | 0.9673484896 |
| SARIMAX(2, 1, 2)(2, 1, 2, 21) | 19462.7078439175 | -8.4002969065 | 0.9720506899 |
| SARIMAX(2, 1, 2)(2, 1, 2, 63) | 20569.5394066321 | -9.4998761912 | 1.0292536953 |

Second, we apply Hyper-parameter Tuning for most optimal set of parameters which result in the best model being SARIMAX(2, 1, 2)x(2, 0, 0, 5) Model: p=2, d=1, q=2.

Third, we fit a SARIMAX(2, 1, 2)x(2, 0, 0, 5) Model with Walk-forward Validation (Rolling Fit Approach) to get better predictions.

## 2.2. Long Short - Term Memory (LSTM)

First we build the LSTM model structure and allows the specification of various hyperparameters

Second, we apply Hyper-parameter Tuning for most optimal set of parameters which result in the best parameters being **'optimizer': 'rmsprop', 'units': 150, 'dropout\_rate': 0.2, 'batch\_size': 16, 'epochs': 30** and getting this output:

| Model | RMSE | R - Squared | MAPE |
| --- | --- | --- | --- |
| LSTM | 0.0177639970228678 | 0.9430008569843317 | 0.0534893001933293 |

## 2.3. Comparison

| Model | Metrics | | |
| --- | --- | --- | --- |
| R^2 | MAPE | RMSE |
| ARIMA (rolling\_fit approach) | 0.9910210639 | 0.0216267379 | 601.5128609415 |
| LSTM | 0.9430008569843317 | 0.0534893001933293 | 0.0177639970228678 |

=> Although R^2 and MAPE of ARIMA perform better, RMSE (which indicates the predictory accuracy of the model) of ARIMA is not as good as LSTM. Therefore we choose LSTM