Scripts/Folders

* **[Step 1: ETL Process]** 0\_map\_algorithm.R: helps with the extraction of daily prices from companies under map files
* **[Step 1: ETL Process]** 0\_sanitise\_data.R: contains all the important utility functions that helps transform raw daily prices to cleaned adjusted daily prices. These adjusted prices are very similar to the ones on other data providers such as Bloomberg and Refinitiv and hence could be trusted.
* **[Step 1: ETL Process]** 1\_separate\_companies.R: contains the algorithm that separates firms into 3 main categories
  + firms that are not in the map file and not in the factor file
  + firms in the map file but not in the factor file
  + firms not in the map file but in the factor file
  + firms in the map file and in the factor file
* **[Step 1: ETL Process]** 0\_prepare\_raw\_data.R: contains the algorithm that applies all the algorithms and consolidates all the information into 1 data table
  + The output of the data table will contain the **unadjusted daily prices** as well as the **dividend factor and split factor** of the companies, if any
  + If the firm is not in the factor file (i.e. dividends or split announcements is non existent for the firm), the dividend\_factor and split\_factor will be set to 1
* **[Step 1: ETL Process]** 2\_prepare\_data.R: contains the algorithm that applies all the algorithms and consolidates all the information into 1 data table
  + The output of the data table will contain the **adjusted daily prices**
  + **In subsequent parts of the workflow, the adjusted daily price data will be utilized.**
* **[Step 2: Feature Engineering]** 3\_preprocess\_data folder: contains 11 R scripts that create new features from adjusted daily price data. Each script serves a specific purpose, aiming to improve forecasting insight and model performance. **0\_consolidation\_features.R utilizes all 11 scripts to create 1 consoldiated data table with all the features for all (S&P500) symbols.** Broadly speaking, these are the main categories of features created
  + Technical indicators
  + Variance tests (Variance Scale Exponent)
  + Stationary tests (Exuber)
  + Time series models and its associated 1 day forecast (TVGARCH, GAS, TsDyn, fractionally differenced ARIMA, wavelet ARIMA)
  + Risk measures: VaR and Expected Shortfall (Quarks, Ufrisk)
  + Generic time series features (Theft, tsfeatures)
* **[Step 3.1: Machine Learning Process]** 4\_MLR\_process\_draft1.R: contains the main machine learning process that creates an ensemble of models (1-layer multi-layer perceptron neural network, xgboost, lightgbm) with the **target variable being the 22-day monthly return [returns\_22]**
* **[Step 3.2: Trading Algorithms]** 4.1\_test\_exuber\_Nico.R: contains a simple trading algorithm using exuber. It normalizes exuber features, sorts stocks, and selects some based on quantiles (lowest and highest) as part of the strategy. It also includes backtesting and performance results.