Mid-Project: The Trend Is Your Friend?

# Introduction

This project will look at historical stock and ETF data to answer the following:

1. What is the probability of a short-term price fluctuation (trend) continuing another day. For example, if a stock price rises one day, what does historical data show is the probability that it will rise the next day?
2. Are these probabilities true for ETFs as well as for stocks?
3. How does the number of trend days affect the probability?
4. How does the total trend rise affect the probability?

# Dataset

Kaggle’s Stock Market Dataset (534 MB)

This dataset contains data for stocks and ETFs traded on US markets from 2.1.1964 – 1.4.2020.

Only high-volume stocks (average volume > 1 million trades) will be used, and only for the last two years of data.

## Data

Data for stocks is in 5884 csv files, size 98 B – 1.7 MB. Total size: 2.331 GB.

Data for ETFs is in 2165 csv files, size 1321 B – 736 KB. Total size: 417 MB.

File names are the stock or ETF symbol (1-4 letters), as set by the company.

Structure (both file sets)

**Column Type, Format**

**Date** Date, dd/mm/yyyy

**Open** Decimal (4,15)

**High** Decimal (4,15)

**Low** Decimal (4,15)

**Close** Decimal (4,15)

**Adj Close** Decimal (4,15)

**Volume** Int

## Metadata

symbols\_valid\_meta.csv includes the symbols and metadata for stocks and ETFs in the data files.

Not needed for this project.

Structure

**Column Type, Format**

Nasdaq Traded Y/N

Symbol Char(4)

Security Name VarChar(254)

Listing Exchange Char(1)

Market Category Char(1)

ETF Y/N

Round Lot Size TinyInt (1, 10, 100)

Test Issue Y/N

Financial Status Char(1)

CQS Symbol Char(4)

NASDAQ Symbol Char(4)

NextShares Y/N

# Variables

The input variables are:

1. Number of trend days
2. Probability (percent) of trend continuation in next day
3. Stock or ETF

# Analysis outputs

1. Number of trend days
2. Total rise (percent)
3. Probability of rise on next day
4. Stocks/ETFs

# ETL Process

1. Create PostGreSQL database (Tools: Docker, Python, Psycopg)
2. Extract csv data into JSON (Tools: Python, csv module, json)
3. Transform JSON data to PostGreSQL-compatible data ? (Tools: Python, json)
4. Load into PostGreSQL tables (Tools: Python, Psycopg)

# Analysis

## Data observation & verification

Load PostGreSQL data into JSON

Observe it visually (Tools: Python, Psycopg, Pandas)

## Data cleaning

Delete incomplete data (e.g. CEF.csv has no Open price data; EXPC has days with Volume = 0) (Tools: Python, Psycopg)

## Manipulation

(Tools: Python, Psycopg, )

### Indexing

### Modification

Calculate intermediate results (total days per trend-length?)

### Grouping

By number of trend days

### Filtering

By Market Category, Exchange, Stock/ETF

### Aggregate

Total number of post-trend continuations

## Storage

Into PostgreSQL tables from JSON (Tools: Python, json, Psycopg)

## Presentation

(Tools: Python, Plotly or Matplotlib)

X axis: Number of trend days {n}

Y axis: Percent occurrences (number of trends{n+1} divided by number of trends{n}

Legend: Population description: Stocks/ETFs, Market Category, etc.

If time permits: show for user-input Stocks/ETFs, Market Category, minimum price increase

# Notes

If time permits, the analysis will also be done for price falls.