Financial Data Science - Value Investing Stock Identifier

FIN42110: Financial Data Science
April 20th 2022

Individual Project

Barry Crosse – 21207024

UCD Michael Smurfit Graduate Business School

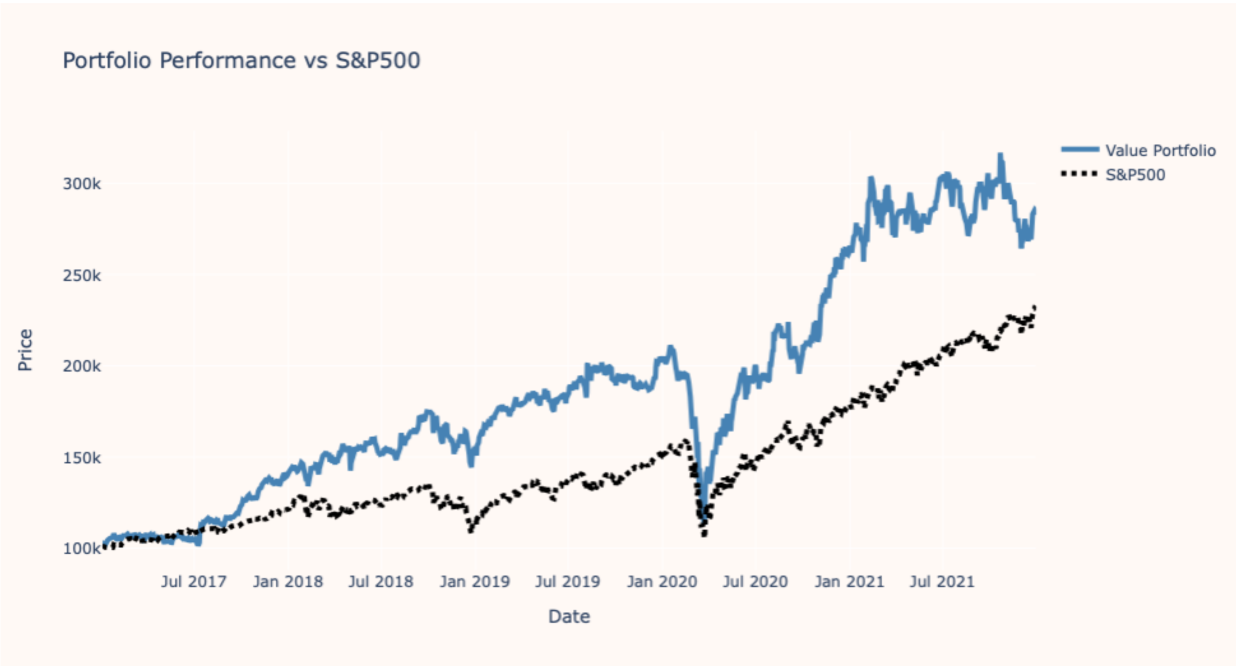
Table of Contents

Executive Summary

Value Investing can be thought of originating from Benjamin Graham's (1949) book 'The Intelligent Investor'. Some of the world's most famous value investors such as Warren Buffet, Charlie Munger and Michael Lee-Chin have amassed great fortunes from following the principles of value investing. Simply, value investing is a strategy that aims to pick stocks that are trading below their intrinsic value, or their book value. Value investors believe that they can buy shares when they are cheap and sell when they are overpriced.

Traditionally, value investors spent hours each day reading financial statements and newspapers to gather information on price fundamentals and public sentiment towards companies. The aim of this project is to create a tool that not only searches for key information about the value of each stock in the S&P500, but also to analyse the public's mood towards each company, analyse trends in past prices and to select stocks whose prices are predicted to increase in the long run. The output is a tool where no user input is required, it is a fully automated python script that identifies stocks that are currently undervalued, have positive sentiment on Twitter and whose prices are predicted to increase.

The tool was back tested over the five-year period from 01/01/2017 – 31/12/2021 and provided the following results:



Asset	Starting Balance ($)	End Balance ($)	Overall Return (%)	Y.O.Y. Return (%)	Daily Standard Deviation (%)
Portfolio	100,000	285,851	185.85	24.28	25.25
S&P500	100,000	230,980	130.98	18.46	18.88

The stocks identified by the python script outperformed the S&P500 in terms of returns by an average of 5.82% per year, yielding an additional $54,871 on a $100,000 one-off investment. The details of the stock identifier tool will be discussed in this report.

Introduction

Value Investing can be described as "buying shares for less than their intrinsic value so as to make a profit when the share price moves towards or above its intrinsic value in the future" (Villalta, 2015). Simply, it is the idea that one can buy shares when they are cheap and sell when they are overpriced. Value investors have the belief that they can profit from buying shares that they deem are undervalued. However, just because a stock is cheap does not mean that it is undervalued. If the business' performance and productivity drops, the business will not increase in value. Its intrinsic value will fall so that the market value of the stock is deemed as its fair value.

Benjamin Graham's (1949) book 'The Intelligent Investor' is one of the most popular investment books of all time. Warren Buffet, the world's most notable value investor, writes the preface for 'The Intelligent Investor' and calls it "by far the best book about investing ever written". The book is a great read for all those interested in investing, but it is a tough task with over 600 pages of content. Modern investors are less interested in spending hours reading books, with platforms such as YouTube and Investopedia providing short and to the point descriptions of all topics one can think of. As an example, The Swedish Investor's (2018) 13-minute summary of 'The Intelligent Investor' has 2.1 million views. Today's investors want to dedicate less time to learning new information and are demanding the latest new at all times.

Warren Buffet is famous for publicly stating that he reads newspapers and company reports for five to six hours per day. Can any of this be automated? The modern investor has no intention of spending their time reading annual reports in an attempt to find a select few undervalued stocks. It would take over a month to read the annual report for each stock in the S&P500, with a conservative estimate of two hours per report.

Dremen (1998) showed that companies with low price-to-earnings (P/E) ratios and low price-to-book (P/B) ratios outperform the market in the long run. As such, these are the two metrics used to identify stocks whose market values are below their intrinsic values. These are the so-called value stocks.

The aim of this project is to create a tool that not only searches for key information about the value of each stock in the S&P500, but also to analyse the public's mood towards each company, analyse trends in past prices and to select stocks whose prices are predicted to increase in the long run. There is no user input needed, it is a fully automated python script that identifies stocks that are currently undervalued, have positive sentiment on Twitter and whose price is predicted to increase.

The process of the stock identifier tool is as follows:

1. The list of stocks in the S&P500 is identified
2. P/E and P/B ratios for these stocks are collected
3. Stocks with both the lowest 10% P/E and P/B ratios are selected
4. Sentiment analysis is performed on these stocks using data from Twitter
5. The Prophet model is used to forecast those companies with greater positive sentiment than negative.

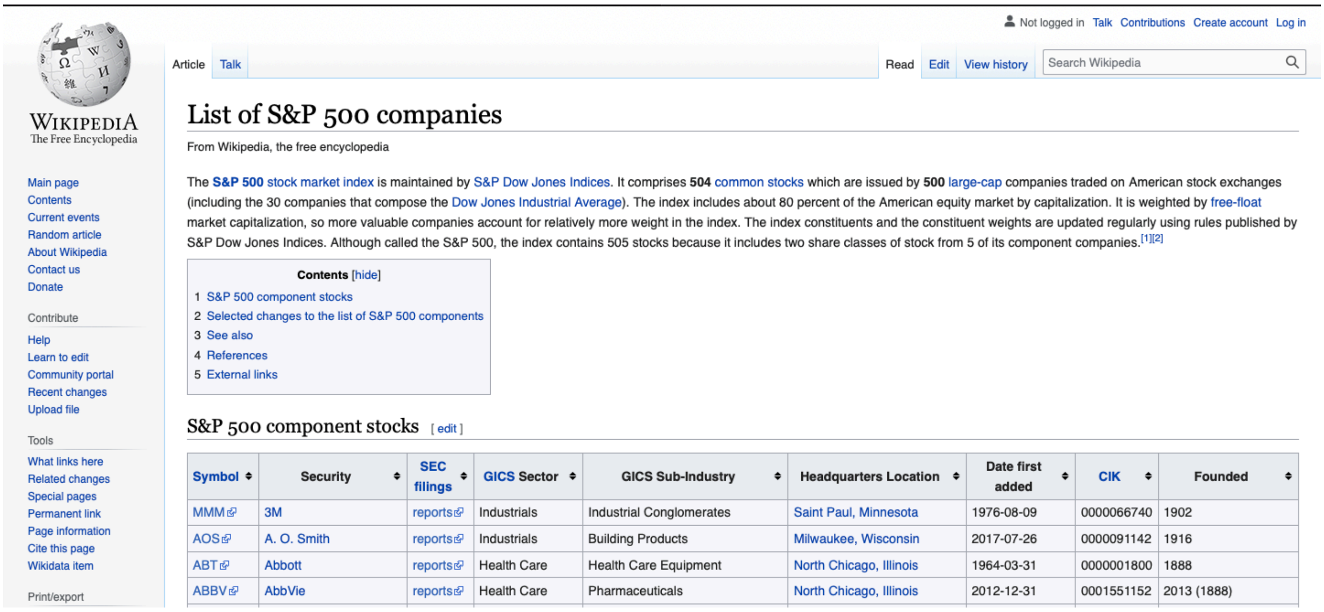
6. Those stocks whose prices are predicted to increase are added to the list of value stocks that are recommended as value stocks to be purchased.

Each step is a filtering phase from step 3 on, with the number of stocks passing through each phase reducing. During testing an average of 5 – 15 stocks make it to the end of the process and are added to the final list of value stocks.

Section 1 – Novel Data Set Collection

Web Scraping S&P500 Stocks

The first step was to get a list of all the stocks in the S&P500. Wikipedia (2022) was used to source the list. Wikipedia was used solely to source the list of stocks; no financial information was sourced from this site. It is common practice in industry to use this site to scrape the list of stocks in the S&P500, as it appeared in many examples online. The function 'pandas.read_html' was used to read the tables on the site into a list of DataFrame objects.
A snippet of the website can be seen below. The 'S&P500 component stocks' table was the table of interest.



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export

List of S&P 500 companies

From Wikipedia, the free encyclopedia

The S&P 500 stock market index is maintained by S&P Dow Jones Indices. It comprises 504 common stocks which are issued by 500 large-cap companies traded on American stock exchanges (including the 30 companies that compose the Dow Jones Industrial Average). The index includes about 80 percent of the American equity market by capitalization. It is weighted by free-float market capitalization, so more valuable companies account for relatively more weight in the index. The index constituents and the constituent weights are updated regularly using rules published by S&P Dow Jones Indices. Although called the S&P 500, the index contains 505 stocks because it includes two share classes of stock from 5 of its component companies.[1][2]

Contents [hide]

1 S&P 500 component stocks
2 Selected changes to the list of S&P 500 components
3 See also
4 References
5 External links

S&P 500 component stocks [edit]

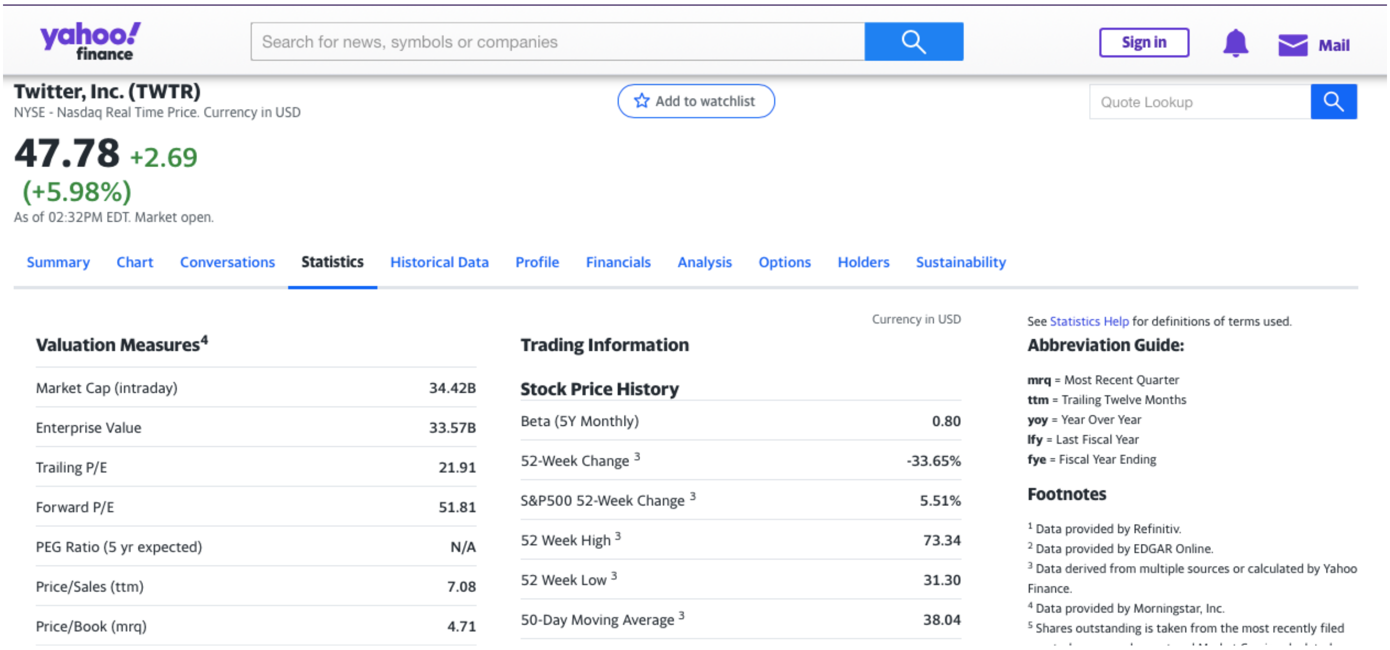
Symbol	Security	SEC filings	GICS Sector	GICS Sub-Industry	Headquarters Location	Date first added	CIK	Founded
MMM	3M	reports	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1976-08-09	0000066740	1902
AOS	A. O. Smith	reports	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	0000091142	1916
ABT	Abbott	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800	1888
ABBV	AbbVie	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)

Yahoo Finance & Beautiful Soup

One of the most difficult processes of the project was to collect financial data from Yahoo Finance. This task required collecting the P/E ratio and P/B ratio for each of the 500+ stocks. Beautiful Soup was used to scrape the ratios from Yahoo Finance. Beautiful Soup is a python library for scraping data from HTML, XML, or other markup languages. It allows users to pull data from webpages that do not offer downloads, removes the HTML markup, and saves the information requested by the user. It is an extremely beneficial tool for collecting large amounts of data, but one should be aware of the limits placed by website owners on web scraping their sites.

During the data set collection Yahoo Finance repeatedly returned 'N/A' values for stock ratios as it did not allow for as many requests from the system as was attempted. To combat this the code was run in blocks of 10 stocks, with a 1-minute sleep timer after each block was executed. It lengthened the collection process significantly, approx. an extra 50 minutes, but ensured the financial data required was collected and no rate limits were breached. The 'Trailing P/E' and 'Price/Book (mrq)' were the information scraped by

Beautiful Soup for each of the 500+ stocks. The Twitter (TWTR) statistics page can be seen below as an example. It would be completely inefficient to manually search each stock's statistics page and write down the ratios required. Beautiful Soup allows for effective automation of this process.



Whilst Yahoo does not appear to explicitly prohibit web scraping, it does not allow for unlimited requests to access data. Inspecting the HTML code and finding the exact sections required is a common topic on Stack Overflow, with many others experiencing similar issues. Kesely (2021) provided the example code basics to be used to combat these issues. Following significant adaptation for use in this project, the code was implemented to collect the financial information.

As part of the business analysis section, historical P/E and P/B ratios were scraped from Macrotrends.net using Beautiful Soup. This will be discussed in detail in the relevant section.

Section 2 – Database Creation and Querying

SQLite3

SQLite3 is a self-contained, file-based SQL database. It can be used in python within any script a user is working on. It allows the user to create a connection to a SQLite3 database, create, edit, read, and modify tables all in python.

As discussed in the previous section, the data collection took a significant amount of time to run, with sleep timers implemented to ensure no rate limits were breached. As such, SQLite3 was extremely beneficial as this data could be saved into a database after it was created. The data collection process did not have to be re-run each time the project was worked on. The data could be taken from the saved database, which is a very quick process, and then used immediately.

The data stored in the database was initially a pandas DataFrame with 3 columns for the ticker, P/E ratio and P/B ratio. This DataFrame was converted in a list of tuples before being added to the database. The function 'add_many()' was created that enabled an efficient process of adding records to the database.

The creation of this database allowed for effective filtering of the data, where stocks with negative ratios were removed and the database then saved. This saved database was then converted back to a pandas DataFrame using the 'pandas.read_sql_query' function, which could then be used for the remainder of the project. Whilst the database was useful for filtering and querying data, its main benefit was in preventing the need to run the web scraping process again. It saved hours of work that was better spent on the following analysis.

Section 3 – Cleaning, Checking & Organising the Data

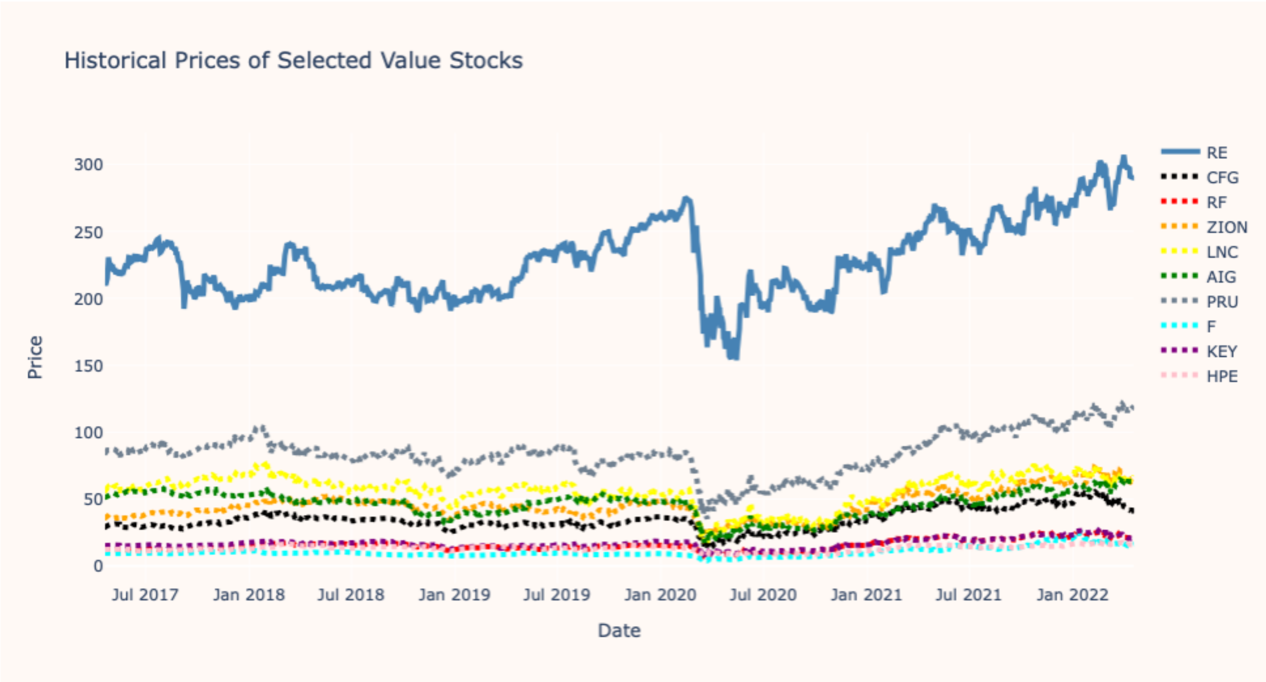
The data collected contained missing values and some incorrect data types. After the initial web scraping of financial data from yahoo Finance, some of the ratios were missing. These figures had to be calculated in an alternative way. To calculate missing P/E ratios, the 'yfinance' package was used. Yfinance allows users to download historical market data from the Yahoo Finance API. It is extremely simple to use and allows for a wide range of data to be downloaded. The only negative is that it can be slow if large amounts of data are requested. However, for this project it was only used to retrieve data for a small number of stocks, so this was not an issue.

Yfinance was used to download share price and earnings per share (EPS) information for some stocks. P/E ratios are calculated by dividing share price by EPS. The same method was used to calculate missing P/B ratios, with book value per share used instead of EPS.

Next, some of the ratios were extremely large. Yahoo finance uses 'k' to represent 1000 for large numbers, e.g. 2,500 is 2.5k on the Yahoo Finance statistics page. As any stock with a P/B or P/E ratio greater than 1000 is not an undervalued stock, the value was changed to negative one, negative ratios would be disregarded later in the filtering process.

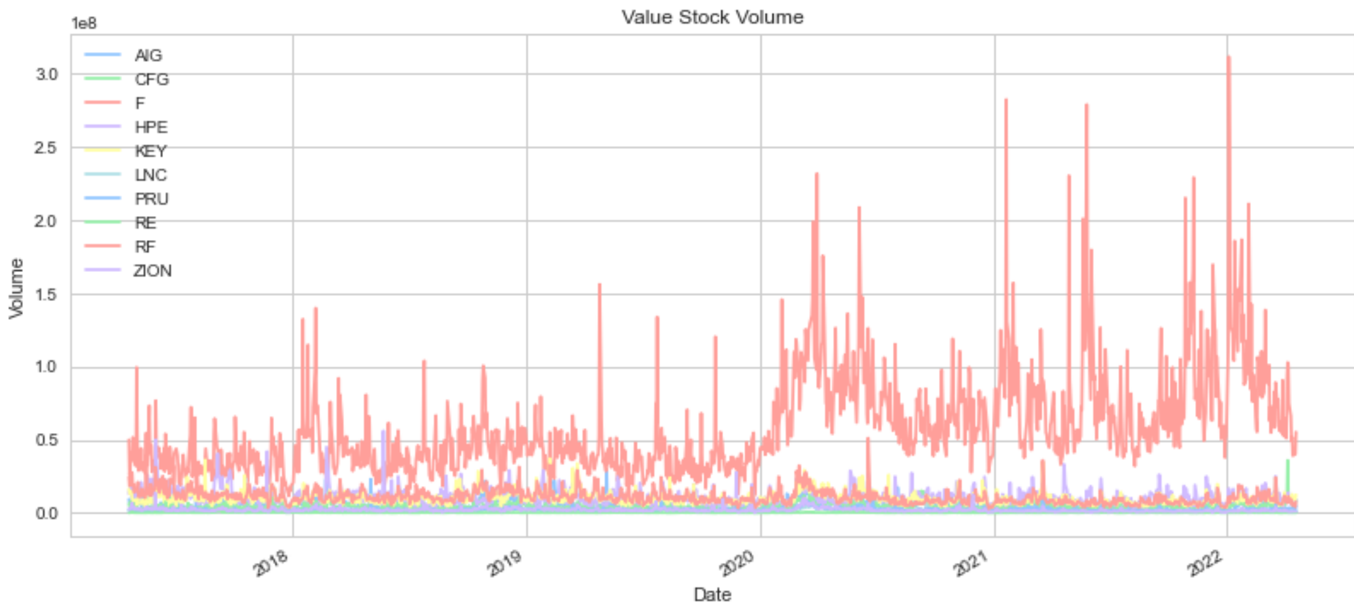
After the database had been created, filtered, saved, and converted back to a pandas DataFrame, the undervalued stocks were identified. The 10% of stocks with the lowest P/E ratios were identified, and the process repeated to find the 10% of stocks with the lowest P/B ratios. Any stocks that fell under both categories were classified as value stocks. A list containing the tickers of these stocks was created, with further analysis of these companies to be undertaken. This list contained 24 stocks to be analysed further in the next stage.

Section 4 – Data Visualisation

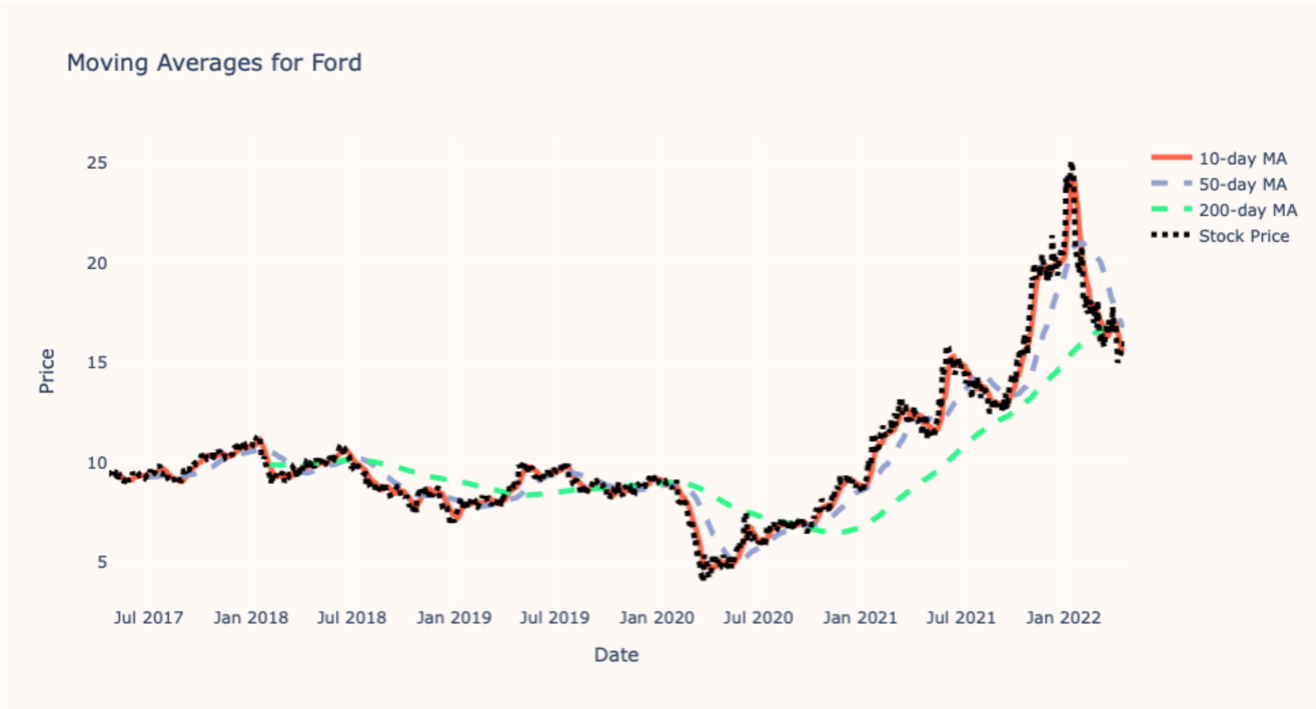


The historical prices of the final selected value stocks (from Section 7) can be seen in the chart above. It is interesting to note that only one of the stocks has a price consistently above $100. Stocks whose value soar to large values such as $1000+ do not appear to have great value. It is possible that stocks whose prices increase dramatically do not increase business performance at the same rate. These value stocks identified can be seen as companies that reflect a bargain purchase.

Looking at the volume, it is clear there is one stock whose shares are traded the most often:



Ford Motor Company is the by far the most actively traded company in the identified list of value stocks. As such, this stock may be bringing large amounts of volatility to the portfolio. Investor sentiment surrounding this particular company is crucially important. Ford Motor Company's historical performance over the last five years can be seen below:



These moving averages could be used to identify trends and deploy trading strategies. A trader may generate signals when asset prices cross over their moving averages. However, as this research focuses on investing, not trading, trading strategies using moving averages will not be investigated further.

Further charts, graphs and images will be shown throughout the next sections as is relevant.

Section 5 – Twitter Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique that determine whether a piece of text is positive, negative, or neutral. Twitter has been used to perform sentiment analysis for stock price predictions for many years. As Gordan Gekko famously quoted in the Wall Street in 1987, "The most valuable commodity I know of is information". The finance sector is quick to utilise all forms of information in an attempt to gain an advantage and to generate alpha.

Pagolu et al. (2016) performed sentiment analysis to check for a correlation between "the rise/fall in stock prices of a company and the public opinions or emotions about that company expressed on Twitter". They classify the tweets as positive, neutral, or negative, and show that there is a strong correlation between sentiment and stock price.

Bollen et al. (2011) show that Twitter can be used to identify the mood of the public toward a company, and that this analysis can be used as part of stock price predictions. They analysed approximately 9.85 million tweets over a 3-week period. The use of Twitter sentiment analysis contributed to an 87.6% accuracy in "predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%". This is a relatively old paper, but it shows that textual analysis has been used to aid in the prediction of stock prices for over 10 years.

Thormann et al. (2021) use TextBlob as part of their 'Stock Price Predictions with LSTM Neural Networks and Twitter Sentiment' research, where they attempt to forecast stock prices 30-minutes and 60-minutes ahead. This short timeframe is not directly applicable to value investing, but it shows that the financial sector is still using Twitter as part of its analysis and predictions.

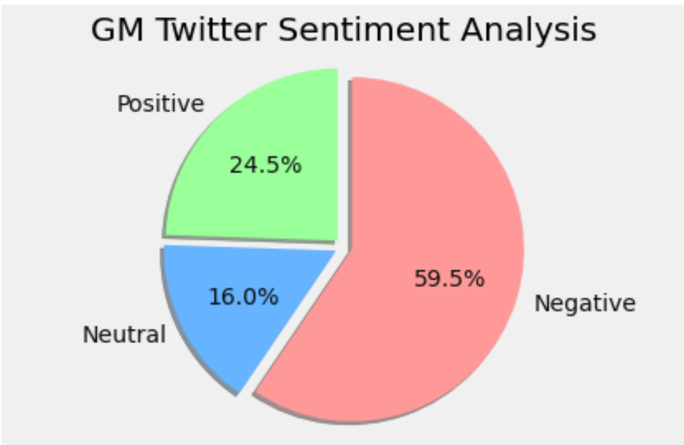
As part of the textual analysis section, this project used TextBlob to perform sentiment analysis of a large number of tweets. TextBlob uses Natural Language ToolKit (NLTK) to perform textual analysis. It outputs polarity and subjectivity scores. Polarity measures the sentiment of tweets, with values ranging from -1 to +1. The greater the score, the more positive the sentiment. Subjectivity measures the level of personal opinion in the text, with values ranging from 0 to 1. The higher the subjectivity score, the more the text resembles personal opinion than factual information.

As part of the textual analysis, Tweepy was used to connect to Twitter's API. Tweepy is a multi-purpose tool that allows users to read tweets, write tweets, send direct messages, and much more. For this project, it was used to collect tweets related to each stock. It was not as simple as simply searching each stock's ticker and reading the results. For example, one of the identified value stocks is Ford, whose ticker is 'F'. If you were to search 'F' using the Twitter API, you would receive irrelevant posts. To combat this, the search term used was '$' + the individual ticker. This allowed for tweets directly related to each ticker to be collected and analysed. Academic access proved extremely beneficial for this research, as over 500,000 tweets were used in total. Tweets from the 1st of March 2022 were collected, allowing for a large number of recent tweets to be analysed. If the search goes back too far it may be analysing sentiment that is no longer relevant.

The sentiment analysis performed on the 24 stocks identified in section 3 was completed using TextBlob. Foote (2021) provided a useful template for performing the analysis. After each tweet was cleaned, removing irrelevant details such as hyperlinks, its sentiment was categorised into positive, negative, or neutral.

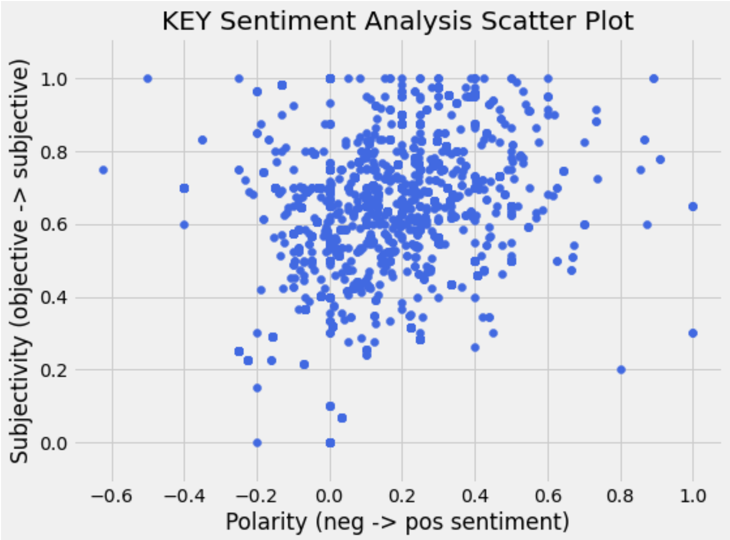
Due to Pagolu et al.'s (2016) findings on the correlation between sentiment and stock price, those companies whose negative sentiment was greater than their positive sentiment were removed from the filtering process.

General Motors (GM) was the only stock to have greater negative sentiment than positive. It was overwhelming negative, with almost 20,000 negative tweets compared to 8,000 positive tweets, and 5000 neutral. Part of this negative sentiment may be due to its poor sales results in Q1 of this year. A news search reveals that its U.S. sales for Q1 this year were down 20% year-on-year compared to 2021 (CNBC, 2022).



This overall performance is captured well in the sentiment analysis. The recommendation here is to avoid this stock, and any future stocks, whose sentiment is overwhelmingly negative. As such, this filters the list of stocks down to 23 different companies.

The analysis also looked at the polarity and subjectivity of the tweets surrounding each stock. The majority of the stocks had a moderately average subjectivity score, but one that stood out was KeyCorp (Ticker: KEY), a company that offers a wide range of financial services and operates under the name of KeyBank. Its performance can be seen below:



KEY had an overwhelming positive polarity score, with the vast majority of these tweets being highly subjective. As such, this positive sentiment may lead to an increase in its stock price, as stated by Pagolu et al.’s (2016) findings.

The list of stocks currently contains 23 companies that will be analysed in the next section.

Section 6 – Predictive Modelling

Prophet

Prophet was used to produce a forecast for each of the stocks that had greater positive sentiment than negative sentiment. Prophet is an additive regression model (Meta, 2017), with the following main components:

- A piecewise linear or logistic growth curve trend. It automatically detects changes in trends by selecting changepoints from the data
- A yearly seasonal component modelled using Fourier series
- A weekly seasonal component using dummy variables
- Important holidays

Krieger (2021) shows that Prophet produces a model that is the sum of three functions and an error term:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

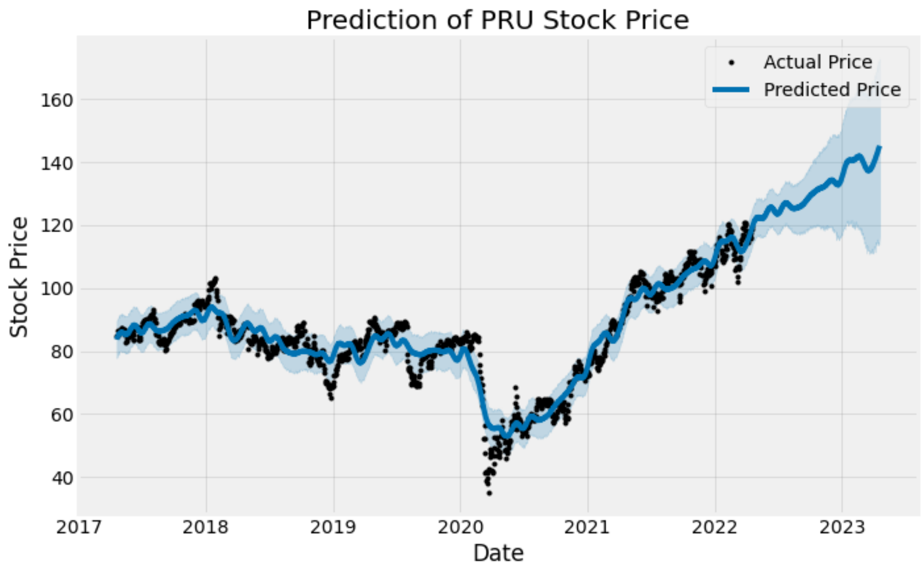
where g(t) = growth, s(t) = seasonality, h(t) = holidays and e_t = error term.

The growth function models the overall trend of the data. It has three options: linear growth, logistic growth, or flat (no growth). The default setting of linear growth was used for this project.

Sheeba et al. (2021) used the Prophet model to predict stock prices for Banco Santander. They noted that "its ability to handle seasonality proved to be extremely fruitful and helped in minimising the uncertainty associated with the prediction of stock market prices". They found that more recent prices have a greater effect on the model than older prices. This is to be expected, as the further a price dates back, the less impact it has on the current price.

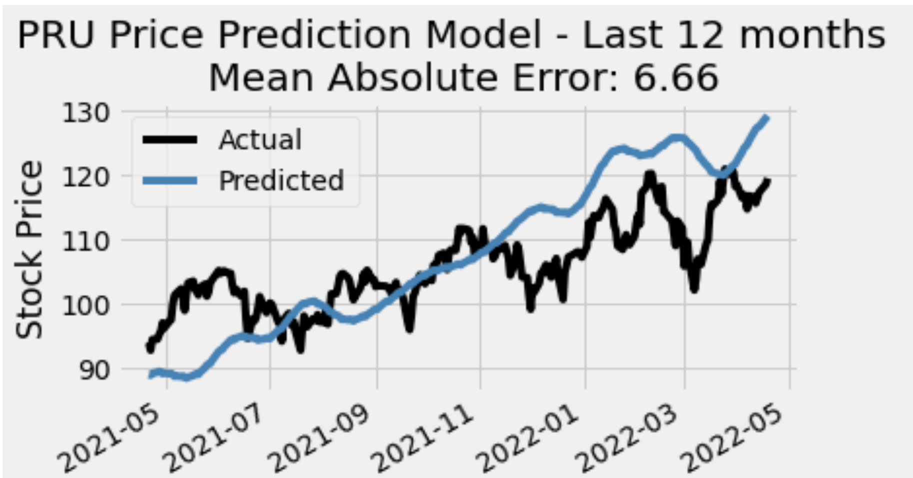
The use of the Prophet model is to identify trends in the past prices of stocks that have made it to this stage of the project. Those stocks who are trending downward and have a predicted price 1-year from now that is lower than their current price will be eliminated. Only stocks whose current price is expected to increase will pass this stage. Stocks that pass this stage of the prices will be the final value stocks, with positive sentiment and a price that is expected to increase. These are the stocks that will be recommended for investment consideration.

The model was trained on the previous five years price data for each stock, and a one-year prediction was made. An example prediction can be seen for Prudential Financial Inc (PRU) below:



s

If it were possible to measure the performance of the model's future predictions, one could exploit this tool and earn arbitrage profits. Alas, this is not possible. As such, a similar model was created using data from April 2017 – April 2022. The first four years of data was used to train the model, with the last year used as a test dataset. This is an 80:20 training-test split. The graph below shows an example of the 23 tests run, with Prudential Financial Inc (PRU) analysed.



The results were positive as a whole with mean absolute error (MAE) scores low. MAE measures the difference between the predicted values and the actual values in general. It is an easily interpretable statistic as it is measured in dollars. In the above example, the MAE score of 6.66 means that the model's predictions were $6.66 off on average. For a stock whose price was in the range of $90 - $125, this score is moderate. Any poor MAE scores do not disqualify a stock from being included in the final basket of stocks, as the aim here is not to exactly predict the prices, but to eliminate stocks which are trending downward.

A popular alternative forecasting model is Long Short-Term Memory networks (LSTM). LSTM models are powerful time-series models. It can produce a forecast up to an arbitrary number of steps into the future, but it is not always the best model (Mushailov, 2021). LSTM models are computationally expensive over long time horizons. As the topic of value of investing is a long-term proposal, LSTM is not the most appropriate model. The full code for this project takes approximately 4 hours to run in total. If an LSTM model was to be created for 20 - 30 stocks it could make the project computationally infeasible. The Prophet model has been shown to perform well and is much less computationally intensive.

23 stocks made it to this final phase of the project. The following ten stocks have passed the rigorous process and should be considered for investment consideration.

Ticker	Company	Price 18/04/22
RE	Everest RE Group	290.07
CFG	Citizens Financial Group	41.23
RF	Regions Financial Corp	20.92
ZION	Zions Bancorporation	63.68
LNC	Lincoln National Corporation	65.33
AIG	American International Group	63.90
PRU	Prudential Financial	118.50
F	Ford Motor Company	15.67
KEY	KeyCorp	20.77
HPE	Hewlett Packard Enterprise	15.71

Section 7 – Business Analysis – Impact of findings

Portfolio Comparison vs S&P500

Stock predictions are notoriously difficult to make. Fund managers who can accurately predict which stocks are going to increase in value are extremely rare, and are often helped by large amounts of good fortune.

It is not possible to directly measure the predictions made, but the value investing tool can be back tested. The code was run to make predictions, but the date of all information collected was for 31st December 2016. Collecting the ratios data for this time was particularly difficult, but the information was gathered from Macrotrends.net. Tweets from December 2016 were analysed, and stock prices from 1st January 2012 to 31st December 2016 were used for the final part of the code. The following seven stocks were chosen by the tool as its output:

['AIG', 'NRG', 'NWSA', 'MKC', 'MRO', 'GM', 'MTCH']

To measure how well these stocks performed, the five-year 1st January 2017 to 31st December 2021 was analysed. An equally weighted portfolio with a starting balance of $100,000 was compared to the performance of $100,000 in the S&P500. The results can be seen below:

Period: 01/01/2017 – 31/12/2021

Asset	Starting Balance ($)	End Balance ($)	Overall Return (%)	Y.O.Y. Return (%)	Daily Standard Deviation (%)
Portfolio	100,000	285,851	185.85	24.28	25.25
S&P500	100,000	230,980	130.98	18.46	18.88

The value investing tool's predictions produced an average yearly return of 24.28%, compared to the 18.46% return of the S&P500 over the same period. This additional 5.88% return equates to an additional $54,871. These gains are significant and would be further amplified by a larger investment. However, this additional reward does not come without additional risk. The daily standard deviation of the value portfolio is approximately 25, whereas the S&P500 had a standard deviation of 19%. it is up to the individual investor if this risk is something they are willing to take on in order to achieve the additional returns.

References

Bollen, J., Mao, H. & Zeng, X.J. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*, vol. 2, no. 1, pp. 1 – 8.

CNBC (2022) 'General Motors' U.S. sales slumped in the first quarter, trailing rival Toyota's', Available at: https://www.cnbc.com/2022/04/01/general-motors-gm-q1-2022-us-sales.html.

Dremen, D. (1998) 'Contrarian Investment Strategies', Simon & Schuster.

Foote, J. (2021) 'Intro-to-Programming-and-Data-Science', Available at: https://github.com/jdfoote/Intro-to-Programming-and-Data-Science/blob/fall2021/extra_topics/twitter_v2_example.ipynb.

Graham, B., & Zweig, J. (1949) 'The intelligent investor: a book of practical counsel', New York, Collins Business Essentials

Kesely, A. (2021) 'Cannot scrape from table in yahoo finance', Available at: https://stackoverflow.com/questions/68263449/cannot-scrape-from-table-in-yahoo-finance.

Krieger, M. (2021) 'Time Series Analysis with Facebook Prophet: How it works and How to use it', Available at: https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a.

Meta (2017) 'Prophet: forecasting at scale', Available at: https://research.facebook.com/blog/2017/02/prophet-forecasting-at-scale/.

Mushailov, J. (2021) 'LSTM Framework for Univariate Time-Series Prediction', Available at: https://towardsdatascience.com/lstm-framework-for-univariate-time-series-prediction-d9e7252699e.

Pagolu, V.S., Challa, K.N.R, Panda, G. & Majhi, B. (2016) 'Sentiment Analysis of Twitter Data for Predicting Stock Market Movements', *2016 International Conference on Signal Processing, Communication, Power, and Embedded System (SCOPES)*.

Sheeba, S.L., Gupta, N., Anirudh, R.R.M. & Divya, D. (2021) 'Time Series Model for Stock Market Prediction Utilising Prophet', *Turkish Journal of Computer and Mathematics Education*, vol.12, no. 6, pp. 4529 – 4534.

The Swedish Investor (2018) 'THE INTELLIGENT INVESTOR SUMMARY (BY BENJAMIN GRAHAM)', Available at: https://www.youtube.com/watch?v=npoyc_X5zO8.

Thormann, M.L., Farchmin, J., Weisser, C., Kruse, R.M., Safken, B. & Silbersdorff, A. (2021) 'Stock Price Predictions with LSTM Neural Networks and Twitter Sentiment', *Statistics, Optimization and Information Computing,* vol. 9, pp. 268 – 287.

Villalta, T. (2015) 'The Large-Cap Portfolio: Value Investing and the Hidden Opportunity in Big Company Stocks', Bloomberg Financial Series, vol.176.

Wikipedia (2022) 'List of S&P500 companies', Available at: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.

Appendix