Analyzing stock markets in different development groups

by SangHyun Han, Jonathan Harper, Jongwon Lim

The Efficient Market Theory

The accepted view of the efficient market theory is that when new information arises, this information spreads quickly and is incorporated into the prices of the stock market without delay. Therefore, any technical analysis, including data mining, cannot help investors select 'undervalued' stocks.

This is associated with the idea of a 'random walk,' which notes that tomorrow's price change will be dependent on tomorrow's news and will be independent of the price change today.

(source: 'The Efficient Market Hypothesis and Its Critics' by Burton G. Malkiel, Princeton University)

Motivation

It is impossible to buy an undervalued stock or to sell an inflated stock in an efficient market, assuming the validity of the efficient market theory, and any factors that could be considered as 'predictive' become useless.

However, this is only true in an efficient market. In a less efficient market, or a market in a less developed country, information may spread slower and technical analyses could result in the purchases or sales of mispriced stocks.

Objective:

Through this analysis, we would like to answer the following question:

Are there factors that make an investment in a company in a less developed country profitable, where a similar investment in a more developed country would not be profitable?

Click here to view our code

Data Sources

Primary data set: Compustat Global

- **Description**: Compustat Global provides standardized global financial statements (Fundamentals Annual) and market data (Security Daily) for over 80,000 active and inactive publicly traded companies. Financial professionals have relied on these statements for over 50 years. In this project, financial statements and market data from **10** countries from the years 2000 to 2021 were used.
- **Size**: 17,383,574 records of market data & 73,969 records of financial statements
- **Variables**: Country Name, Exchange Code, Stock Price, Trading Volume, Dividend per Share, Assets, Revenue, Income, Cash Flow, etc. Additional information on slide 3.
- **Format**: API (Python Library "wrds")
- **Access & Locations**: Wharton Research Data Services (WRDS). Needs account approval from University of Michigan's WRDS Representatives. Follow instructions here to sign up.
- **Time Frame**: 2000 - 2021

(source: Compustat description provided in WRDS

Secondary Data Source: The World Bank's Economy data set

- **Description**: The data set shows purchasing power parity for 266 countries across 62 years. Purchasing power parity (PPP) measures a country's gross national income using an international dollar rate and is a strong indicator for how developed a country is, as it accounts for the overall welfare of a country by accounting for the prices of goods that aren't traded across countries. The data from this data set are in current international dollars and based on the 2011 International Comparison Program round.
- **Size**: 192 KB
- **Variables**: Country Name, Country Code, PPP (per year), PPP (average - calculated during data manipulation)
- **Format**: XLS
- **Access & Locations**: The World Bank's Economy data set; View and download here
- **Time Frame**: 1960 - 2021

(source: Economy description provided by the World Bank)

Data Background

Primary Data Source - Source Code Found Here, Notebook: Factor Significance, File: milestone.py

Primary Data Source: Compustat Global

Goal: Use market data and financial statements from Compustat Global to create variables (a.k.a. factors) that can explain future returns.

A new data set is created for each country that is being analyzed, and statistical relationships between factors and future returns will be computed individually by country. Time-series market data and financial statements will be transformed to cross-sectional data with "year" and "company" being unique identifiers. This means that each row in the data set represents "company X at year Y".

"Factor" is a term used in finance to define a quantifiable firm characteristic that can explain differences in stock returns. In the data manipulation phase, we create multiple factors that inform our target variable, "annual return."

All rows are timed for the first business day of each year.

Feature Engineering - How Each Factor Contributes

There are 8 factors being used to calculate future annual returns. A full list of variables can be found here. We will explain what they are below and explain how they were created and manipulated on the following slide.

1. Momentum: Uses past stock price trends to show the direction in which the stock is moving.
2. Volatility: Shows how rapidly and unpredictably a stock is moving.
3. Liquidity: Shows how easily a security can be bought and sold.
4. Size: aka market capitalization, the total value of all of a publicly traded company's shares
5. Yield: The income returned on an investment
6. Quality: Refers to the performance of a company, determined by a number of values on the company's financial statements. We are using the following factors in this analysis: assets, revenue, operating income, earnings before interest, and cash_flow.
7. Growth: Annual percentage change in quality factors
8. Value: Shows if a company is being overvalued or undervalued in the market compared to their financial performances

Data Manipulation Methods

Primary Data Source - Source Code Found Here, Notebook: Factor Significance, File: milestone.py

Feature Engineering - How Each Factor Was Created

Factors were pulled from the primary data set's API through SQL scripting. Please reference the source code for more specific details.

1. Momentum: Six variables were created under the momentum category, all related to time. Momentum is measured over the past 1, 3, 6, 12, 24, and 36 months. To measure momentum on a 3 month interval for company X on January 1, 2020, we will use the stock prices from October 1, 2019 to January 1, 2020 to do so. We created a pandas library to store our momentum values and calculated them using the following formula: (current price - price N months before) / (price N months before)

2. Volatility: Six variables were created under the volatility category, all related to time. Volatility is measured over the past 1, 3, 6, 12, 24, and 36 months. We used the pandas functions 'rolling' and 'std' to calculate rolling standard deviations for each timestamp at each company.

3. Liquidity: Six variables were created under the liquidity category, all related to time. Liquidity is measured over the past 1, 3, 6, 12, 24, and 36 months. We used the pandas functions 'rolling' and 'mean' to calculate rolling averages for each timestamp at each company.

4. Size: We multiplied a company's stock price by the company's total listed shares.

5. Yield: This value was pulled directly from our primary data set and did not require modification.

6. Quality: Five variables were created under the quality category: assets, revenue, operating income, earnings before interest, and cash_flow. These values were pulled directly from our primary data set and did not require modification.

7. Growth: Five variables were created under the growth category. We used the pandas function 'pct_change' on the quality values that were pulled directly from the primary data set to get Assets year over year (YoY), revenue YoY, operating income YoY, earning before interest YoY, and cash flow YoY .

8. Value: Five variables were created under the value category. We divided size by the five quality values that were pulled directly from the primary data set to get PBR, PSR, POR, PER, and PCR.

Data Manipulation Methods

Secondary Data Source - Source Code Found Here, Notebook: PPP

Secondary Data Source: The World Bank's PPP data set

Note: Joined to the primary data set using the Country Name

Goal: Use PPP to define four groups of development: Low development, Medium low development, Medium high development, and High development.

1. Data is not equally available for every country or for every year. We condensed the year-range from 62 years to 17 years, chosen based on the available data points. Any country that did not include PPP values for all 17 years was dropped.
2. We averaged the PPP values across all 17 years and used this average PPP data point to rank each country amongst its peers in terms of development standards.
3. Development groups were created based on the percentiles of the averaged PPP column.
    a. 0 - 25% are 'Low Development' countries
    b. 25 - 50% are 'Medium Low Development' countries
    c. 50 - 75% are 'Medium High Development' countries
    d. 75 - 100% are 'High Development' countries.

Grouping explanation

There are 3 potential ways to group countries into development groups.

1. Use the World Bank's Income Groups
2. Section each country into groups based on their PPP percentiles
3. Use the UN Development Programme's Human Development Report groupings, which relies on Human Development Index scores.

We chose to section each country into groups based on their PPP percentiles. This best defines the welfare of a country and its financial markets.

The World Bank uses GNI per capita (similar to PPP) to place countries into income groups, but they regroup these countries every year. This one-year scope conflicts with our seventeen-year scope and places some countries in groups that don't represent their historical data well.

Non-economic factors are used in the calculation of HDI. We are emphasizing an economic approach and will stick with economic factors..

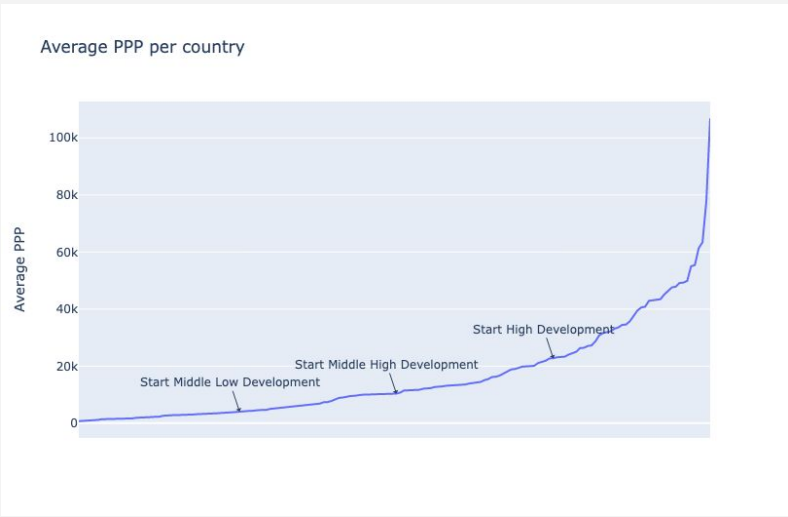
Analysis: Development Groups

Income Group distribution

We originally thought the income groups defined by the World Bank would be the best way to define a country's development. However, when comparing the PPP numbers we had with the income groupings, we saw some countries with lower PPP averages were placed in a higher income group than other countries with higher PPP averages. This is shown through the red and green colored groups in the stacked bar graph. Due to this inconsistency, we chose to split countries by their PPP percentile instead of by their pre-defined income group.

Difference in purchasing power parity across groupings

The line graph shows that there is a significant gap between the lowest and highest "High Development" country in terms of average PPP. We chose to ignore this gap in favor of consistency within our development group sizes, as our analysis focuses on the individual countries within a grouping and not the group as a whole. In a more complex analysis, it may be wise to omit outlier countries or classify the highest development countries in their own group.



Analysis

Factor Significance in Developing and Developed Countries

We chose to fit linear regression models for five countries in both the lowest and highest percentile groups. Our feature engineering factors are used as independent variables and the country's future annual returns are used as our dependent variable.

We defined countries in the lowest PPP grouping as **developing**. They are:

- Bangladesh, Côte d'Ivoire, Kenya, Zambia, Zimbabwe

We defined countries in the highest PPP grouping as **developed**. They are:

- Australia, Spain, Italy, Singapore, Cyprus

Our regression results show that there are 4 significant factors informing future annual return per developing country, compared to 2.8 significant factors informing future annual return per developed country. **Predictive financial statement factors like quality, growth, and value were seen exclusively in developing countries** and seem to be a significant intelligence gathering area for positive returns in developing countries.

These results support our original hypothesis, stated in the motivation section, that there are factors exclusive to developing countries that may lead to the purchase of undervalued stocks. This is potentially due to the lack of an efficient market.

Significant Factors in Developing Countries

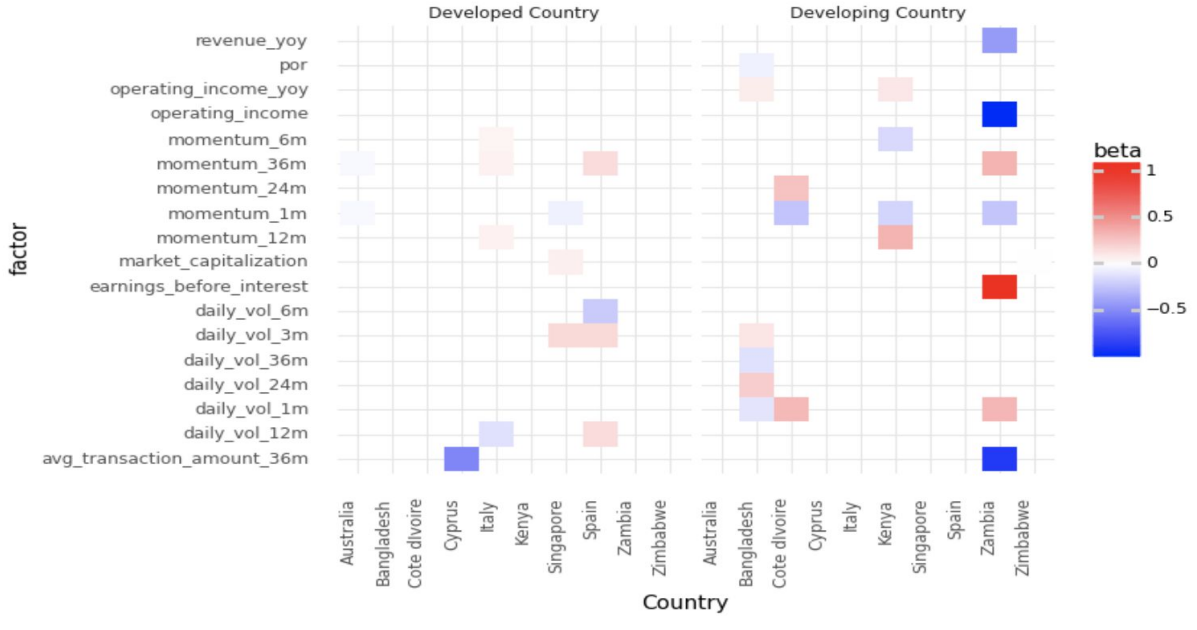
Country	factor	beta	pvalue
Bangladesh	por	-0.060206	0.013099
Bangladesh	daily_vol_1m	-0.109333	0.003569
Bangladesh	daily_vol_3m	0.111891	0.071276
Bangladesh	daily_vol_24m	0.202655	0.020115
Bangladesh	daily_vol_36m	-0.123436	0.077489
Bangladesh	operating_income_yoy	0.079520	0.000387
Cote dIvoire	daily_vol_1m	0.296641	0.000573
Cote dIvoire	momentum_1m	-0.245724	0.001256
Cote dIvoire	momentum_24m	0.251488	0.023905
Kenya	operating_income_yoy	0.099236	0.062283
Kenya	momentum_1m	-0.174095	0.000201
Kenya	momentum_6m	-0.158335	0.021281
Kenya	momentum_12m	0.317678	0.000117
Zambia	operating_income	-0.954534	0.087992
Zambia	earnings_before_interest	1.034127	0.093519
Zambia	daily_vol_1m	0.300734	0.096455
Zambia	avg_transaction_amount_36m	-0.857302	0.043379
Zambia	revenue_yoy	-0.415459	0.022392
Zambia	momentum_1m	-0.241093	0.058503
Zambia	momentum_36m	0.322775	0.068041
Zimbabwe	NaN	NaN	NaN

Significant Factors in Developed Countries

Country	factor	beta	pvalue
Australia	momentum_1m	-0.028870	0.005811
Australia	momentum_36m	-0.024894	0.036598
Cyprus	avg_transaction_amount_36m	-0.501323	0.028062
Spain	daily_vol_3m	0.158471	0.015241
Spain	daily_vol_6m	-0.214191	0.003270
Spain	daily_vol_12m	0.146221	0.036153
Spain	momentum_36m	0.146613	0.000193
Italy	daily_vol_12m	-0.121438	0.020041
Italy	momentum_6m	0.044135	0.079279
Italy	momentum_12m	0.057736	0.052956
Italy	momentum_36m	0.063442	0.030163
Singapore	market_capitalization	0.069876	0.019268
Singapore	daily_vol_3m	0.154435	0.003345
Singapore	momentum_1m	-0.060677	0.000138

Visualization

Factor Significance in Developing and Developed Countries



Significant factors for each country are shown in the plot above. Red tiles represent positive correlation of factors with future returns and blue tiles represent negative correlation of factors with future returns. We can see that only developing countries have significant financial statement related factors (Revenue YoY, Income YoY, Price over Operating-income Ratio).

Next Steps

Splitting up the "High Development" grouping

As seen in our Development Groups Analysis, there is a large PPP difference between the lowest and highest "High Development" country. Splitting this up can provide further insight into the differences between the very highest development countries and those lower than it.

U.S. Analysis

One academic paper notes that markets outside of the United States may be less efficient due to a segmented market created by structural cross-country barriers to investment management. We did not include the U.S. market in our analysis, and investigation on this could provide insight into the efficient market theory and how it is impacted by international borders.

P-hacking potential

There is a long-standing debate related to this topic, where many think that analysis like this is anomaly data snooping and qualifies as p-hacking. Others argue that the anomalies found are real to a large extent. Further, more focused and advanced research studying the financial statement factors we found as significant should provide further confidence on their ability to inform future annual returns in developing countries.

(source: 'Anomalies Across the Globe' by Heiko Jacobs and Sebastian Müller, *Journal of Financial Economics*)

Statement of Work

Sang Hyun Han: Collection and manipulation of Compustat Global data. Creating factors and dataset for regression analysis. Interpreting and analyzing regression results for different countries/groupings.
Jonathan Harper: Research to determine economical development characteristics. Collection and manipulation of PPP data set. Creation of development groupings. Primary analysis of PPP data set. Editor for final presentation.
Jongwon Lim: Research to determine economical development characteristics. Collection and manipulation of HDI data set. Primary analysis of HDI data set.

Collaboration Statement: Collaboration went very well, with all three of us contributing significantly to the project's data manipulation, analysis, and presentation. As our project evolved over the two month timeframe, we could have improved our status updates to remove minor confusions and double work on the project's status and findings.

Endnotes

Economy. WDI - Economy. (n.d.). Retrieved September 20, 2022, from datatopics.worldbank.org/world-development-indicators/themes/economy.html

Jacobs, H., & Müller, S. (2020). Anomalies across the globe: Once public, no longer existent? Journal of Financial Economics, 135(1), 213–230. https://doi.org/10.1016/j.jfineco.2019.06.004

Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics (thesis). CEPS.

University of Pennsylvania. (n.d.). Compustat. WRDS: Wharton Research Data Services. Retrieved September 25, 2022, from https://wrds-www.wharton.upenn.edu/pages/get-data/compustat-capital-iq-standard-poors/compustat/