Report on Identifying Viable Long-Term Investment Opportunities
By Yin Kwong John Lee

1 Motivation

Inflation is an inevitable, yet defensible economic phenomenon, when prices of goods and services increase over time. For example, the amount required to buy a cell phone today may only best afford a calculator twenty years later. Fortunately, the stock market allows us to adapt to, but not avoid the effect of rising prices in the long-run. To adapt to inflation, careful investments must be made so that the amount invested now can be sufficiently compounded to offset the inflation rate in the long-term future. This can be done by first and foremost selecting viable businesses to invest in. The metrics introduced in this study will evaluate how each stock treats its investors, and with the help of computation techniques such as Spark and SparkSQL, suggest the viability of investments on both state and corporate levels. To summarize, this study aims to inspect the viability of long-term investments in the United States over the past 19 years. I chose to conduct this study because I want to help people with limited accessibility to financial resources make informative investment decisions. I do this by shedding light to the following three questions:
1) How do viable investment opportunities change at different locations throughout the years?
2) How are investment opportunities compared on a state and corporate level?
3) Are there any good indicators for identifying viable long-term investments?

2 Data Sources

2.1 Company Information Dataset (http://rankandfiled.com/#/data/tickers)

The company information dataset is the subset of the original data obtained from the link because some non-relevant columns are drops from the original data (e.g. medium of exchange). There are 13737 observations with 3 variables (described below). This data is stored in a CSV (Comma-Separated Format) file under the filename “company_info.csv”. Here are the important variables:

- *ticker* : symbol representing a company’s stock
- *company_name* : official name under which a company files its reports
- *location* : state where a company operates in

2.2 Company Sections Dataset (https://simfin.com/data/bulk)

The company sections dataset is combined from 3 separate datasets- income statements (“market_income.csv”), balance sheets (“market_balance.csv”), and cash flow statements (“market_cashflow.csv”), which reflect the earning powers reserve abilities, and cash handling skills of the listed companies respectively from 2001 to 2019. These separate datasets are downloaded independently from the link above, and then joined together on *ticker* and *fiscal_year* to summarize the companies' important measures in one bigger dataset. This was done using the merge function from pandas dataframe in python. There are 16493 observations with 15 variables (described below). This data is stored in a CSV (Comma-Separated Format) file under the filename "company_sections.csv". Here are the important variables:

- *ticker* : symbol representing a company's stock
- *fiscal_year* : year which a company files its annual report (income statement, balance sheets, and cash flow statements are contained in the annual report)
- *num_shs_outstand* : number of shares traded in the stock market
- *revenue* : the amount that a company earns from sales
- *pretax_income* : a company's profit before tax
- *net_income* : a company's profit after tax
- *total _current_assets* : the amount of assets that can be converted into cash quickly
- *total_current_liabilities* : short-term financial obligations due
- *total_assets* : the amount of assets for the year
- *total_liabilities* : financial obligations due for the year
- *st_debt* : debt to be paid off within a year
- *lt_debt* : debt that matures more than one year
- *retained_earnings* : income left after shareholders are paid dividends
- *cash_operating _activities* : cash flow generated from operations
- *dividends_paid* : the portion of profit paid to shareholders

3 Data Manipulation and Methods

The source code for this section is stored inside the path "./lykjohn-part-1/data_manipulation.py"

3.1 Joining the Two Datasets

The company information and company sections datasets are joined on *ticker* such that the pair *ticker* and *fiscal_year* becomes a unique identifier for the new dataset. This means any record can be queried by specifying the company ticker and the year of the record. There are 15072 observations with 17 variables in the joined data.

3.2 Handling Missing Data

The newly joined dataset may contain missing values for an arbitrary record. For example, Amazon's total revenue may be missing in 2019. Since revenue and other financial measurements are fundamental for building our metrics later on, they must not be speculated with inexact values like zeros. Therefore, I dropped the rows that contain at least one null value in the financial variables, variables from the company sections data except *ticker*, *fiscal_year*, and *dividents_paid*. Note that records with missing dividend payouts are not dropped, but rather set to zero because dividends are paid only by well-established businesses, and assuming the worst case (zero dividend) when dividends are unknown could yield more conservative results, thus more well-established businesses to be included in the list of viable businesses. The missing *location* values are set to “Unknown” because states, although trivial for our metrics calculations, are important for comparing the viability of investments. Some invalid state values (e.g. A0) are also set to “Unknown”. The dataset has 9273 observations left with 17 variables at this point.

3.3 Appending Company Metrics

Now, 8 financial metrics are calculated based on the financial variables (e.g. *revenue*, *net_income*) and appended to the cleaned dataset. There are now 9273 observations left with 25 variables ready for computation. The formulation and description of each metric is tabulated below:

Metric	**Formulation**	**Viable Condition**
Return on Equity: The return that a company generates through the use of equity	ROE= pretax_income/ (total_assets + total_liabilities)	ROE> 0.12
Return on Total Capital: The return that a company generates through the use of its capital structure(equity + debt)	ROE= pretax_income/ (total_assets + total_liabilities+ st_debt+ lt_debt)	ROTC> 0.12
Net Profit Margin: The portion of revenue earned as income before tax	NPM= pretax_income/ revenue	NPM> 0.12
Current Ratio: Measures whether a company’s current assets is sufficient to cover its current liabilities in a given year	CURRENT_RATIO= total_current_assets/ total_current_liabilities	CURRENT_RATIO> 1
Debt-to-Earnings Before Income Tax Ratio: The number of years estimated for a company to	DEBT_EBIT= (st_debt+lt_debt)/ pretax_income	DEBT_EBIT< 3

pay off its total debts		
Dividend Payout Ratio: The portion of net income paid to investors for a fiscal year	DIV_PAYOUT_RATIO= dividends_paid/ net_income	0.35< DIV_PAYOUT_RATIO< 0.55
Earnings Per Share: The amount investors can expect the company to earn for each share they own	EPS= net_income/ num_shs_oustand	N/A
Retained Earnings Per Share: The amount a company keeps for possible future developments for each share they own	REPS= retained_earnings/ num_shs_oustand	N/A

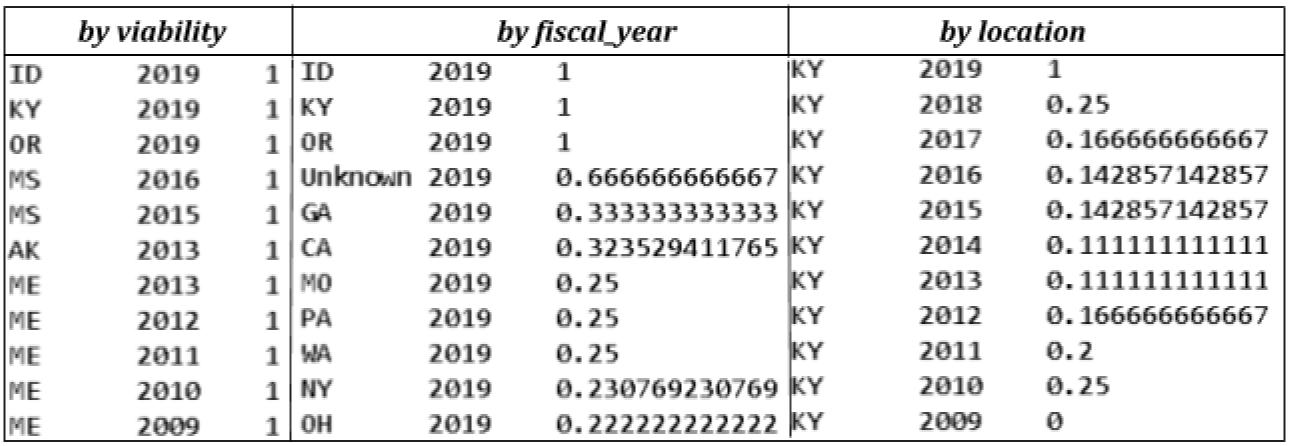
Remark: A company's stock is viable if its financial metrics in the "Metric" column satisfy all of the conditions in the "Viable Condition" column in the above table.

4 Analysis and Visualization

The dataset resulting from the above manipulation procedures is stored as "company_metrics.json" and "company_metrics.csv" inside the "./data" folder. The JSON file will be used for analysis while the CSV file will be used for visual interpretation. The source code for this section is stored inside the path "./lykjohn-part-1/data_computation.py".

4.1 Question 1: How do viable investment opportunities change at different locations throughout the years?

This question can be answered using the "company_metrics.json" dataset under the PySpark framework. After each line of the JSON file was loaded into a resilient distributed data (RDD) object, the object was passed into a function called *viable_by_loc_year()* to assign a binary pair (1,0) to the state for which its stock is a viable investment in a particular year and (0,1) otherwise. For instance, ((CA, 2019), (1,0)) is assigned to a record having California as the state and 2019 as the fiscal period, given that the corresponding company's stock qualifies to be a viable investment. The company ticker doesn't matter because I am interested in how each state performs. The resulting list of objects are then reduced by their keys to contain the total number of viable and non-viable stocks. The instance ((CA, 2019), (50, 150)) means that In 2019, California has 50 viable stocks and 150 non-viable ones. The percentage of viable stocks for a state in a given period is mapped to each RDD object. In other words, in the same example, a resulting tuple of (CA, 2019, 0.25) may be interpreted as 25% of California- based companies are viable for long-term investment in 2019. The percentage may also be referred to as the *viability* of long-term stock investments for a state in a particular year. The same procedure is repeated for each state in a given year, hence a list of tuples with unique *location* and *fiscal_year* parameters, along with their viability results are expected. The results are ordered by *viability, fiscal_year,* and *location*, and are stored in the files named "viable_prop_year_loc.tsv", "viable_year_prop_loc.tcv", and "viable_loc_year_prop.tsv", respectively. Here are some excerpts from the outputted text files:



by viability			*by fiscal_year*			*by location*		
ID	2019	1	ID	2019	1	KY	2019	1
KY	2019	1	KY	2019	1	KY	2018	0.25
OR	2019	1	OR	2019	1	KY	2017	0.166666666667
MS	2016	1	Unknown	2019	0.666666666667	KY	2016	0.142857142857
MS	2015	1	GA	2019	0.333333333333	KY	2015	0.142857142857
AK	2013	1	CA	2019	0.323529411765	KY	2014	0.111111111111
ME	2013	1	MO	2019	0.25	KY	2013	0.111111111111
ME	2012	1	PA	2019	0.25	KY	2012	0.166666666667
ME	2011	1	WA	2019	0.25	KY	2011	0.2
ME	2010	1	NY	2019	0.230769230769	KY	2010	0.25
ME	2009	1	OH	2019	0.222222222222	KY	2009	0

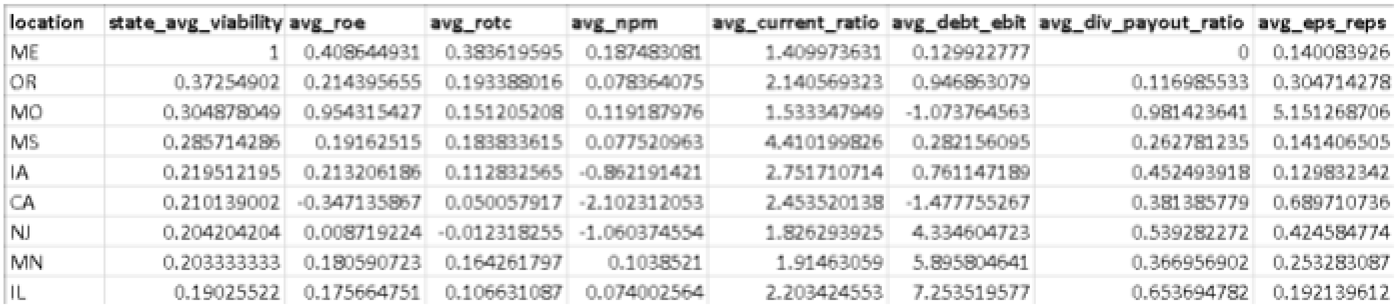
Based on the above excerpts, Kentucky state, denoted by "KY", in 2019 has one of the highest viability of long-term investments of all time. This can be seen by ranking all the states by *viability*. This means that in 2019, 100% of the publicly listed businesses from Kentucky are qualified for viable stock investments. We then order by *fiscal_year* to observe that the only other states that have 100% long-term investment viabilities in 2019 are Idaho ("ID") and Oregon ("OR"). This means that Ketucky is one of the 3 out of 52 states to consider when making investment decisions. We look at its viability trend after sorting the data by *location* and observe that Kentucky has been a state of improving investment viabilities, from 16.67% in 2012 after a dip in 2011 to 100% in 2019. Another thing to note is Kentucky is a state with high investment viability because there are not as many businesses as in other bigger states such as California ("CA") and New York ("NY"), which in such case may result in a lower viability. The 32.35% in CA and 23.08% in NY accounts for the proportions of viable stocks among a more extensive pool of business ranging from BigTech companies to financial institutions.

4.2 Question 2: How are investment opportunities compared on a state and corporate level?

To answer this question, I converted the RDDs, in particular ones from the "viable_loc_year_prop.tsv"into a SparkSQL dataframe object and called it *viable_stocks*.

Similarly, I imported the cleaned dataset from “company_metrics.json”into another SparkSQL data frame object and called it *stocks_df* .

The first goal is to construct a data frame that contains the average metrics measurements and viabilities from a state level. This was done by joining the two data frames on their common variables, *location* and *fiscal_year*, then averaging the *viable_prop*, *ROE, ROTC, NPM, CURRENT_RATIO, DEBT_EBIT, DIV_PAYOUT_RATIO*, and *EPS/REPS* throughout the years. *viable_prop* is the third variable from the “viable_loc_year_prop.tsv” data that represents the percentage of viable investments in a state at a certain point in time. *EPS/REPS* is the quantity describing the amount that contributes to a company’s earnings for every dollar it retains for possible future developments. Typically, the higher this quantity, the more value added to a stock, the better an investor is treated. The data is sorted in descending order by the average of *viable_prop*, denoted by *state_avg_viability*. The eventual data frame contains 47 observations with 9 variables, and is written into the file “state_avg_viability.csv”. Here is an excerpt:

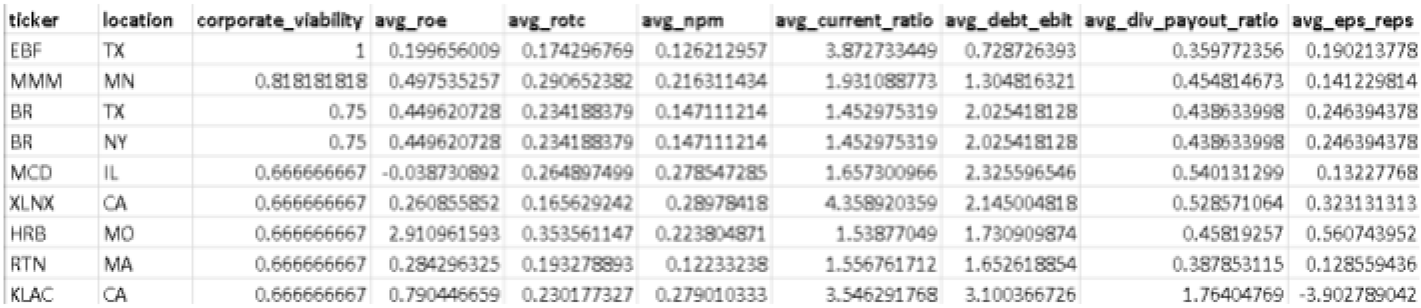


location	state_avg_viability	avg_roe	avg_rotc	avg_npm	avg_current_ratio	avg_debt_ebit	avg_div_payout_ratio	avg_eps_reps
ME	1	0.408644931	0.383619595	0.187483081	1.409973631	0.129922777	0	0.140083926
OR	0.37254902	0.214395655	0.193388016	0.078364075	2.140569323	0.946863079	0.116985533	0.304714278
MO	0.304878049	0.954315427	0.151205208	0.119187976	1.533347949	-1.073764563	0.981423641	5.151268706
MS	0.285714286	0.19162515	0.183833615	0.077520963	4.410199826	0.282156095	0.262781235	0.141406505
IA	0.219512195	0.213206186	0.112832565	-0.862191421	2.751710714	0.761147189	0.452493918	0.129832342
CA	0.210139002	-0.347135867	0.050057917	-2.102312053	2.453520138	-1.477755267	0.381385779	0.689710736
NJ	0.204204204	0.008719224	-0.012318255	-1.060374554	1.826293925	4.334604723	0.539282272	0.424584774
MN	0.203333333	0.180590723	0.164261797	0.1038521	1.91463059	5.895804641	0.366956902	0.253283087
IL	0.19025522	0.175664751	0.106631087	0.074002564	2.203424553	7.253519577	0.653694782	0.192139612

The second goal is to construct a dataframe that contains the same types of information, but from a corporate level. This was done by averaging the same financial variables, *ROE, ROTC, NPM, CURRENT_RATIO, DEBT_EBIT, DIV_PAYOUT_RATIO*, and *EPS/REPS*. The challenging part was to query the *viable_prop* for each company in each state because the data frame converted from RDD just gives us the *viable_prop* for every state. To do this, I repeated the procedure from 4.1, but in a SQL format. The key query here is:

```
SUM(CASE
         WHEN ROE>0.12 and ROTC>0.12 and  NPM>0.12 and CURRENT_RATIC
and DEBT_EBIT<3 and DIV_PAYOUT_RATIO>0.35 and DIV_PAYOUT_RATIO<0.55
         THEN 1
         ELSE 0
    END)/(count(*) * 1.0)
```

This works because the observations that meet our viability criteria are assigned the value 1, and otherwise 0. A sum is taken to record the number of observations that meet the criteria. Divide this over the total number of observations to get the proportion that suits for viable stock investments. This proportion is denoted by *corporate viability* in the eventual data frame. Lastly, group the proportions by *ticker* and *location* to obtain the viability for each company's stock in each state. The *location* variable is included for us to compare companies' financial measures with the state's financial averages. The same grouping was applied to the rest of the financial variables to obtain a complete data frame that records the average metrics measurements and viabilities from a corporate level. The data is ordered by *corporate_viability* in a descending manner, consists of 1263 observations with 10 variables, and is written into the file "corporate_viability.csv". Here is an excerpt:



ticker	location	corporate_viability	avg_roe	avg_rotc	avg_npm	avg_current_ratio	avg_debt_ebit	avg_div_payout_ratio	avg_eps_reps
EBF	TX	1	0.199656009	0.174296769	0.126212957	3.872733449	0.728726393	0.359772356	0.190213778
MMM	MN	0.818181818	0.497535257	0.290652382	0.216311434	1.931088773	1.304816321	0.454814673	0.141229814
BR	TX	0.75	0.449620728	0.234188379	0.147111214	1.452975319	2.025418128	0.438633998	0.246394378
BR	NY	0.75	0.449620728	0.234188379	0.147111214	1.452975319	2.025418128	0.438633998	0.246394378
MCD	IL	0.666666667	-0.038730892	0.264897499	0.278547285	1.657300966	2.325596546	0.540131299	0.13227768
XLNX	CA	0.666666667	0.260855852	0.165629242	0.28978418	4.358920359	2.145004818	0.528571064	0.323131313
HRB	MO	0.666666667	2.910961593	0.353561147	0.223804871	1.53877049	1.730909874	0.45819257	0.560743952
RTN	MA	0.666666667	0.284296325	0.193278893	0.12233238	1.556761712	1.652618854	0.387853115	0.128559436
KLAC	CA	0.666666667	0.790446659	0.230177327	0.279010333	3.546291768	3.100366726	1.76404769	-3.902789042

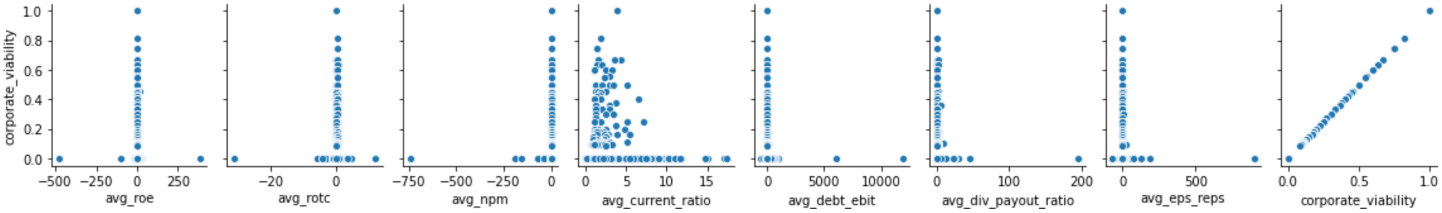
From a state level, California (CA) with 21.01% average viability lies among the top 10 states for having the highest average proportion of businesses viable for long-term investments over the past 19 years. From a corporate level, two California-based companies, namely Xilinx (XLNX) and KLA Corporation (KLAC) rank amongst the top 10 businesses for long-term investments by having stock viabilities of 66.67% over the past 19 years. If we look deeper, the average *ROE, ROTC, NPM, CURRENT_RATIO,* and *DIV_PAYOUT_RATIO* for each of the two businesses are higher than their state averages, which implies that they have above-state average earnings power and reserve ability (Note that the comparisons of *DEBT_EBIT* are dismissed for simplicity purposes). At this point, it seems that XLNX and KLAC can be considered as viable investments. However, if we look at their average *REPS/EPS* quantities, 32.31% for XLNX and -390.28% for KLAC, they seem to stay significantly below that of the state average of 68.97%. This means that average California businesses are retaining more money to fund investors' holdings than XLNX and KLAC. This should be considered when choosing these two stocks to add to one's portfolio.

Moreover, from the above observation, companies ranked highly in their viabilities for long-term investments may seem to have above-state average financial measures. This may be a proper deduction at first glance. However, if we look at McDonalds (MCD), the giant food chain that operates in Illinois with a stock viability of 66.67%, its financial measurements such as average *ROE* and *DIV_PAYOUT_RATIO* are below the Illinois (IL) average. This means that top ranked businesses for long-term investments do not necessarily outperform the whole state in certain aspects.

Similar rationale can be used to draw conclusions for other stocks when comparing from corporate and state levels.

4.3 Question 3: Are there any good indicators for identifying viable long-term investments?

The source code for this section is stored inside the path
"./lykjohn-part-1/data_visualization.py"



I have plotted the pairwise plots for each of the financial features from the "corporate_viability.csv" to see whether the direction of one measure can imply the direction of the viability of companies' stocks. Since no positive or negative correlation is present in any of the plots, it is safe to conclude that none of the averages of financial metrics, *ROE, ROTC, NPM, CURRENT_RATIO, DEBT_EBIT, DIV_PAYOUT_RATIO*, and *EPS/REPS,* serves as a good indicator for the viability of a company's stock. The same conclusion can be drawn by looking at the metrics values around 0%, when the viability of corporate stock varies in many percentages. For example, at the average ROE of 0%, there are stocks with 0% viability to ones with 100% viability, which implies no direct correlation between *ROE* and *corporate_viability*. If the outliers are removed, we may be able to see the trends clearer, but I do not expect for the conclusion to change much.

5 Challenges

Although the three questions can be approached independently, I find it interesting to leverage the information derived from one question and carry it to another. Therefore, the biggest challenge in this study is to apply large-scale computation techniques for one data set and then translating its context to give meaning to another dataset. In particular, when the RDD objects in 4.1 are converted to a SparkSQL data frame to store variables about different states and years, I was struggling to connect this state-level data with the corporate level data to answer the question in 4.2. This was because although I have the average viabilities for long-term investments for each state, I do not have the viability of each company in each state for comparison. I could have repeated procedures in 4.1 to obtain this, but rather, I would like to practice with SQL queries that do the same job. I ended up constructing the key query in 4.2 to deliver the viability of each company in each state. Analysis from then on is easy.

Another notable challenge is to come up with a meaningful question to ask for visualization. Given that there are so many variables in this study, I was overthrown from asking a linear question, such as "How does the viability of long-term investments vary from 2001 to 2019?". Rather I seek to ask a more sophisticated question that is useful even after this study, "Are there any good indicators for identifying viable long-term investments?"