**COGS9 : Introduction to Data Science**
FInal Project

**Group Member Info**

| First Name | Last Name |
|---|---|
| Jacinth | Fang |
| Darwin | Park |
| Benjamin | Xue |
| Irene | Jiang |
| Hamlet | Torosian |

**Question**

Is there a relationship between return on equity and the US stock price's growth rate for companies within the consumer goods industry in 2019?

**Hypothesis**

There is a positive correlation between the return on equity and the US stock price's growth rate for companies within the consumer goods industry in 2019.

**Justification**

Return on equity (ROE) provides information on how much a company has profited (their net income) compared to the total equity the company's shareholders have. Positive assumptions such as money efficiency can be made when a company has a high ROE. Moreover, higher return on equity usually results in an increase in stock prices, which attracts more investors to buy the company.

**Background information**

Since the financial metric "Return on Equity (ROE)" is a highly deterministic factor for the profitability of a company, we are interested in finding out if there is a relationship between the ROE percentage and the increase in the stock price. Important financial metrics such as ROE and others have been researched by many others data scientists. Martina Rut Utami and Arif Darmawan have analyzed a similar topic on the effect of ROE on the stock price in Sharia Indonesian Stock (https://core.ac.uk/download/pdf/229850617.pdf). They used the purposive sampling method with 53 companies in the sample, and they conclude that there is no statistically significant relationship between the ROE and the stock price.

However, another study done by Agung Fajar Ilmiyono yielded different results. In his research paper, “The Effect of ROE, ROA, and EPS toward Stock Prices in Companie sub Sektor Construction and Buildings Listed in Exchange Indonesia Effect”, Ilmiyono was able to hypothesize and conclude that an increase or decrease in ROE will influence the stock price in a similar manner (http://www.ijlemr.com/papers/volume4-issue8/4-IJLEMR-44183.pdf). This study also used the sampling method but with only 9 companies. ROE is definitely not the only factor to look at when determining which stocks will be profitable, but it would be interesting to see whether or not the ROE of a company is an influential factor of their stock price considering the two different research results.

**Data**

Our ideal dataset should have all the companies within the consumer goods industry. However, it is impossible to determine the exact number of observations as there is no information about the number of companies in the consumer goods industry. Therefore we will take samples of 150 companies to perform the research. The two variables we would be measuring are ROE and growth rate/decay. For each company, we would record their average return on equity and stock price growth rate.

Since we are working with stocks, we need a couple of data sets in order to answer our question. To calculate the ROE (return on equity), we would need an API that lists a company’s income statements and balance sheets. Then we would need to obtain the closing stock prices of companies throughout the year of 2019 and calculate their rate of growth.

We will obtain stock price data from yfinance, an open-source library compiled with financial data from Yahoo Finance. More than 3,000 companies’ data are reachable via yfinance. Yfinance provides stock price in a CSV format allowing undemanding data wrangling. Furthermore, the library frequently updates and provides the data at various intervals ( 1 min, 5 min, 10 min, etc ). Depending on the variation of the interval, the number of observations fluctuates. The library provides five variables: open, high, low, close, adj close, and volume. The “close” column will be isolated and used to calculate the daily growth or decay rate.

| Date | Open | High | Low | Close | Adj Close | Volume |
|---|---|---|---|---|---|---|
| 1980-12-12 | 0.128348 | 0.128906 | 0.128348 | 0.128348 | 0.101261 | 469033600 |
| 1980-12-15 | 0.122210 | 0.122210 | 0.121652 | 0.121652 | 0.095978 | 175884800 |
| 1980-12-16 | 0.113281 | 0.113281 | 0.112723 | 0.112723 | 0.088934 | 105728000 |
| 1980-12-17 | 0.115513 | 0.116071 | 0.115513 | 0.115513 | 0.091135 | 86441600 |
| 1980-12-18 | 0.118862 | 0.119420 | 0.118862 | 0.118862 | 0.093777 | 73449600 |
| ... | ... | ... | ... | ... | ... | ... |
| 2020-10-28 | 115.050003 | 115.430000 | 111.099998 | 111.199997 | 111.199997 | 143937800 |
| 2020-10-29 | 112.370003 | 116.930000 | 112.199997 | 115.320000 | 115.320000 | 146129200 |
| 2020-10-30 | 111.059998 | 111.989998 | 107.720001 | 108.860001 | 108.860001 | 190272600 |
| 2020-11-02 | 109.110001 | 110.680000 | 107.320000 | 108.769997 | 108.769997 | 122866900 |
| 2020-11-03 | 109.660004 | 111.489998 | 108.730003 | 110.440002 | 110.440002 | 107020000 |

**Figure1**: Apple’s data received from yfinance open source library.

Another principal data to our research is the return on equity (ROE). ROE is the calculation of values listed on the income statements and balance sheet of companies. FMPcloud(https://fmpcloud.io/) provides the required financial documents as API. The data is available in both CSV and JSON format. The documents downloaded from API contain variables including revenue, gross profit, expense, EPS, net income, and many more. FMPcloud provides financial documents from over 3000 companies. We will calculate the ROE using the net income and shareholder's equity data. The companies provide financial documents quarterly.

aapl_income_statement_full_quarter

| date | operatingIncome | operatingIncomeRatio | totalOtherIncomeExpensesNet | incomeBeforeTax | incomeBeforeTaxRatio | incomeTaxExpense | netIncome |
|---|---|---|---|---|---|---|---|
| 2020-09-26 | 1.4775E+10 | 0.2283687285542060 | 1.26E+08 | 1.4901E+10 | 0.23031623852360200 | 2.228E+09 | 1.2673E+10 |
| 2020-06-27 | 1.3091E+10 | 0.219335 | -1.58E+08 | 1.3137E+10 | 0.220106 | 1.884E+09 | 1.1253E+10 |
| 2020-03-28 | 1.2853E+10 | 0.220414 | -1E+07 | 1.3135E+10 | 0.22525 | 1.886E+09 | 1.1249E+10 |
| 2019-12-28 | 2.5569E+10 | 0.278472 | 8.9E+07 | 2.5918E+10 | 0.282273 | 3.682E+09 | 2.2236E+10 |
| 2019-09-28 | 1.5625E+10 | 0.243988 | 2.06E+08 | 1.6127E+10 | 0.251827 | 2.441E+09 | 1.3686E+10 |
| 2019-06-29 | 1.1544E+10 | 0.214537 | 4.3E+07 | 1.1911E+10 | 0.221357 | 1.867E+09 | 1.0044E+10 |
| 2019-03-30 | 1.3415E+10 | 0.231233 | 3E+07 | 1.3793E+10 | 0.237749 | 2.232E+09 | 1.1561E+10 |
| 2018-12-29 | 2.3346E+10 | 0.276907 | 1.43E+08 | 2.3906E+10 | 0.283549 | 3.941E+09 | 1.9965E+10 |
| 2018-09-29 | 1.6118E+10 | 0.256248 | -1.4E+08 | 1.6421E+10 | 0.261065 | 3.396E+09 | 1.4125E+10 |
| 2018-06-30 | 1.2612E+10 | 0.236778 | 1E+08 | 1.3284E+10 | 0.249395 | 1.765E+09 | 1.1519E+10 |
| 2018-03-31 | 1.5894E+10 | 0.259974 | -4.39E+08 | 1.6168E+10 | 0.264455 | 2.346E+09 | 1.3822E+10 |
| 2017-12-30 | 2.6274E+10 | 0.297577 | 3.8E+07 | 2.703E+10 | 0.30614 | 4.365E+09 | 2.0065E+10 |
| 2017-09-30 | 1.312E+10 | 0.249529 | 9.5E+07 | 1.3917E+10 | 0.264687 | 3.203E+09 | 1.0714E+10 |
| 2017-07-01 | 1.0768E+10 | 0.237139 | -1.85E+08 | 1.1308E+10 | 0.249031 | 2.591E+09 | 8.717E+09 |
| 2017-04-01 | 1.4097E+10 | 0.266504 | -1.65E+08 | 1.4684E+10 | 0.277601 | 3.655E+09 | 1.1029E+10 |
| 2016-12-31 | 2.3359E+10 | 0.298133 | 1.22E+08 | 2.418E+10 | 0.308611 | 6.289E+09 | 1.7891E+10 |
| 2016-09-24 | 1.1761E+10 | 0.251025 | -1.59E+08 | 1.2188E+10 | 0.260138 | 3.174E+09 | 9.014E+09 |
| 2016-06-25 | 1.0105E+10 | 0.238562 | -2.63E+08 | 1.0469E+10 | 0.247155 | 2.673E+09 | 7.796E+09 |
| 2016-03-26 | 1.3987E+10 | 0.276658 | -5.1E+08 | 1.4142E+10 | 0.279724 | 3.626E+09 | 1.0516E+10 |

**Figure 2**: Apple's income statement downloaded from FMPcloud.

Our final dataset would have the date, name of the companies, calculated growth or decay rate, and calculated ROE as variables. ROE is calculated by dividing a company's net income by the shareholder's equity. Since ROE is given out quarterly, we will average out the four values and that will be the company's average ROE for the year. We will store our data in a table where the rows will represent the date and company's name and the columns being the growth/decay rate and the averaged out ROE. Because it is difficult to pinpoint the exact number of companies in the consumer goods industry as it varies all the time, the dataset will have more than a hundred companies in the desired sector.

## Ethical Considerations

### Data collection

There are always ethical considerations when dealing with this specific data science question that involves the extent of data collection that is carried out with this project. We will be getting our stock price data from the yfinance module, a tool that assists in scraping and

downloading web data from Yahoo finance. Additionally, a company's ROE can be obtained through fmp cloud, an API that provides us with access to net income and a shareholder's equity, which can be used to calculate the company's ROE. This can cause some ethical issues with our project, as web scraping can be considered highly unethical when poorly conducted. To address these issues, we will make sure to only scrape the data we need, and be sure to only send a reasonable number of requests to the owner of the data. Since our only source is Yahoo Finance, it's a possibility that some of the numerical information may be off compared to other websites and there are underlying biases. This will result in inaccurate data. One resolution to ensure our data is consistent, is to conduct another experiment from a different source. Using the exact same companies and finding the same information, we can see if the data collected from Yahoo Finance is accurate or if it has any inconsistencies. Lastly, We will not be required to receive consent from any people because we are not using anyone's personal information as our data.

**Data storage**

In data storage, there are ethical concerns over the privacy of our storage. In order to keep our data secure and prevent potential hackers from accessing it, we will store it in a database that only us and the user can access. The Right to be forgotten will also not be an issue since there will be no personal names being used or gathered during our experiment. In order to only use useful data, we will need to extract the data we think we need and go over the data again before we start our analysis. As we go over the data, our group would need to carefully consider which data points can answer our question and which points can not. The data points that will not be used are deleted so it does not interfere with our relevant data. Unnecessary data points may skew our results.

**Analysis**

We must ensure that there are no outside sources of bias when we report our observations. When analyzing the data, we will only base our findings and results off what the data represents with no individual opinions or assumptions. This is critical in making sure we can make predictions and obtain knowledge from the data we obtained. Additionally, we must be able to understand what potential sources of bias led to the data that was collected, and try to ensure a fair sampling of each company. Also, as mentioned in Data Collection, any underlying biases in Yahoo finance will need to be reinforced through reproducing the analysis with another source (ex: google finance). In order to ensure transparency in our analysis, we will display our data in ways that help the reader understand what each variable and data point represents. This will give the user a fundamental understanding of our different financial metrics. When representing our analysis, our group will carefully ensure our visualizations and statistics truly represent the data by interpreting the data first, then asking a third party.

### Modeling

When modeling our data, we will remove any biases that influence the results of the data analysis. When choosing which stocks to look at in the consumer goods industry, we will do a random sample. This way, all stocks within the consumer goods industry have a fair chance of getting chosen, and data wouldn't be manipulated. Also, before gathering or analyzing anything, our group will explain why the variables ROE and stock price rates were chosen to be in our study and why we decided to model our experiment this way. Explaining our experiment and model can help with transparency and allows the audience to understand why we made our model the way it is. Similarly to our analysis, our model will strictly consist of the statistics we need such as stock price, net income, and shareholder's equity. As mentioned in the Data section of our proposal, many companies are coming and going in the stock market. This is unfortunately a shortcoming to our model that will need to be explained to relevant stakeholders since this may affect the results of our sampling or data gathering process.

### Deployment

In order to avoid ethical problems in the deployment stage of our project, we will ensure that company and individual's names or personal information will be removed from any data we used to create our model. If anyone believes their information is used without informed consent, then we would grant them the ability to request for immediate removal of the data.

## Analysis Proposal

### Data Collection

Data will be collected prominently using the Python open-source library module named YFinance and FMPcloud's API. We will first gather all the tickers of companies in the major US stock exchanges(NYSE, NASDAQ, Dow Jones, and S&P 500) by web-scraping the website named "eoddata" (http://eoddata.com/symbols.aspx?AspxAutoDetectCookieSupport=1). The scraped data will have a single column with a variable ticker. Our question concentrates highly on companies within the consumer goods industries in the US stock market. We will isolate consumer goods companies and locations using information about sectors on YFinance. Using the ".info" command on YFinance, it is possible to receive which sector and geological location the company is in. We will narrow down the number of companies for the study by choosing companies with the sector "consumer staples" or "consumer discretionary." However, consumer staples and consumer discretionary sectors contain companies that provide service; we will hand-pick these companies to isolate consumer goods businesses. YFinance provides stock price data in CSV format. We will harvest daily stock data from January 1 to December 31 of 2019 including variables open, high, low, close, adj close, and volume. Furthermore, we will utilize the list of companies in consumer goods industries generated using YFinance to regulate the data of companies we receive from FMPcloud's API. Using the API, we will gather information about operating income, net income, total expenses, and shareholder's equity data, and more from various companies in CSV format, which we could then calculate the ROE based

on these data. We will collect the CSV files from FMPcloud from January 1 to December 31 of 2019 to use for our research.

**Data Wrangling**

Data wrangling for our research is extremely rigorous. The first step of data wrangling lies in eliminating unnecessary columns from both datasets. For YFinance data, we will only use the daily “close” price of stock data. Therefore the rest, including open, high, low, adj close, and volume will be deleted. From FMPcloud’s API, we require net income and shareholder’s equity to calculate ROE. Therefore the remaining variables such as the total income and total expenses will be ruled out. The second step of data wrangling is calculations. Using daily close price data from YFinance, we need to calculate the rate of growth or decay of the company. For example, Day X’s close price will be divided by day X-1’s close price to calculate the growth or decay rate. The calculated growth or decay rate will be included in a new table of data (Figure 3). Furthermore, we need to calculate the ROE of stocks from FMPcloud’s financial documents API. We will divide the net incomes of each company by their shareholder’s equity to obtain the values. These values will be added to the new table of data (Figure 3). There will be tables of data for each company which is later going to be used to calculate the correlation between the growth/decay rate and the ROE.

Nike

| Date | Ticker | Rate | ROE |
|---|---|---|---|
| **2019-01-01** | NKE | 0.03 | 0.01131 |
| **2019-01-02** | NKE | 0.05 | 0.01131 |
| **2019-01-03** | NKE | -0.02 | 0.01131 |
| **2019-01-04** | NKE | 0.03 | 0.01131 |
| **2019-01-05** | NKE | 0.1 | 0.01131 |
| **2019-01-06** | NKE | 0.08 | 0.01131 |
| **2019-01-07** | NKE | 0.07 | 0.01131 |
| **2019-01-08** | NKE | -0.05 | 0.01131 |

**Figure 3 :** Nike’s “new table” with date, ticker, growth/decay rate and ROE. The values are made up.

**Descriptive & Exploratory Data Analysis**

After we have all the data for each consumer goods company’s stock price growth rate and ROE, we would find out the relationship and compare between each company’s stock price growth and ROE in 2019 using a scatter plot and boxplot. For descriptive analysis, we need to find basic statistics such as average ROE and stock price growth rate, and R-squared. Then we

could look at the relationship between ROE and stock price growth in exploratory analysis- if the company with higher ROE, on average, creates a higher stock price growth rate. Also, we would generate a boxplot to explore the relationship between the stock price growth rate and ROE. So, the median, the shape of the distribution, the standard deviation would be observed during descriptive analysis for boxplot. Then, we will compare the median stock price growth rate of companies with different ROE- if the higher median stock price growth rate of a company would come from companies with higher ROEs.

**Data Visualization**

We will employ a scatter plot in the google spreadsheet with the company's 2019 average ROE on the x-axis and 2019 average price growth rate on the y-axis. In the scatter plot, the relationship between each company's average stock price growth rate and its average ROE will be clear. Then we would divide companies into 3 groups- group1 with ROE less than 10%, group 2 with ROE between 10 to 20%, and group 3 with ROE >20% to generate the boxplot. The x-axis would be groups and the y-axis would be stock price growth rate. From the boxplot, one can clearly compare the median, 25 percentile, and 75 percentile. We could also detect any outliers in the boxplot.

## Discussion

The goal of our project was to determine whether there is a relationship between a US company's return on equity and growth rate, indicated by their stock price. After gathering, wrangling, and visualizing our data, we would look for a correlation between the two variables using the scatterplot generated for the entire sample of companies. The slope of the scattered points could determine if there is a significant positive or negative relationship between the average ROE and stock growth rate for US companies in 2019. We could also calculate the correlation coefficient and the R-squared value to give us a clearer idea of the relationship. For example, an r coefficient of .97 would indicate a strong positive correlation between ROE and stock price growth rate.This means a higher ROE would correlate to a higher growth rate.

To further analyze the data, we then split the companies up into 3 groups based on their ROE (<10%, 10-20%, >20%) and generated a box plot for each group. Using the data points, these plots would help us determine how these groups performed with regard to their stock price growth rate in 2019. By comparing medians and quartiles, we could identify whether there is a certain trend going from the lowest ROE group to the highest ROE group and vise versa, or if there is no trend at all. We can also easily identify outliers and groups that had more or less outliers than expected.

A general limitation of our experiment would be that companies frequently leave and enter the stock market, and therefore some inconsistencies with our data will come up. When going through the process of data wrangling, we must manually find missing values of companies that left the stock market, and find discrepancies that were caused from the instability of the stock market. Our biggest source of bias would likely come from sampling bias, since we are only

collecting data from 150 companies. This is a relatively small sample size, and sampling errors can very easily occur to where the entire consumer goods industry is not represented accurately in our specific sample. To address this problem, we will sample companies randomly based on our calculations of certain attributes of a desired sample, accounting for revenue, size(number of consumers and employees), and other key quantitative facts about each company. We will rank these companies based on the average ranks of each category, and split them up into 5 different groups from these rankings. Random sampling will then be done where we select 30 companies from each group, so that our desired sample will have very similar attributes to the averages of all companies across the consumer goods industry.

Since ROE is only one of the many factors that can potentially contribute to changing stock prices, confounds can be found. One of the confounders include external factors such as economic recessions that aren't accounted for in collecting data. For example, running this experiment in the year 2008 during the Great Recession, where most stock prices and ROE of certain companies were largely affected, would lead us to finding an inaccurate relationship between ROE and stock prices. Since we are using data strictly from 2019, we will not have to worry about recessions, as there were no recessions significant enough in 2019. Despite this, we still must keep track of unusual behavior and outliers found in our data and implement ways to account for recessions and other economic factors that serve as confounders in our model.

When dealing with ethics involving the U.S. stock market data, we have to be aware of issues over misrepresenting stock data. We must not only make sure that our data is accurate and reflective of the stock market, but also ensure that the companies we are using to measure aren't being harmed unintentionally. These two components are especially important because misinformation regarding stock data can lead to societal issues such as unintentionally influencing the way that investors invest in the stock market. To avoid this, we will ensure all of our data is transparent to the audience and write a warning that investors should not invest solely based on our study's result because confounding variables are influencing our results. It is critical that our calculated r-squared values are verified, so that we would know if our result is valid or not. As previously mentioned, we will not include any personal information and the company names will be replaced by identifiers in our data.