**ETL Project Report**

Yanelly Nunez

Ricardo Calderon

Date: 1/02/20

**Introduction**

Most investors understand the importance of diversification through asset allocation within a portfolio, which is meant to balance out the risk and reward trade-off between equity, debt and cash holdings. Under the equity or stock positions within a portfolio, however, investors have a wide array of investment options, each with different characteristics, advantages and disadvantages.

Publicly traded companies are classified into three categories: large-cap ($10 billion +), mid-cap ($2-$10 billion), and small-cap (< $2 billion). Most small capitalization firms are startups or relatively young companies with high-growth potential.

 A large-cap stock investment, defined as the stock of a company with a market capitalization of $10 billion or more, is common among both growth and value investors as a portion of an overall asset allocation. Large-cap, or big-cap, stocks have unique advantages for investors, including stability in size and tenure, steady dividend payouts to shareholders and clarity in valuations.

The greatest advantage to adding large-cap stocks to an investment portfolio is the stability they can provide. Companies that are considered small- or mid-cap do not have the same level of stability, and therefore carry a greater degree of risk than large-cap investments.

The small firm effect theory holds that smaller companies have a greater amount of growth opportunities than larger companies. Small-cap companies also tend to have a more volatile business environment, and the correction of problems—such as the correction of a funding deficiency—can lead to a large price appreciation.

General wisdom says then that large-cap companies will produce a smaller return than small-cap companies, because smaller companies have a greater amount of growth opportunities than larger companies. However, if an investor follows that wisdom, the investor may miss those large-cap and strong companies that are still growing, that have large revenues, and which are giving investors not only good dividends but also great returns for holding the stock. Our analysis tries to collecte what are those companies within the top 100 companies by revenue so that investors do not miss those great investment opportunities.

**Extract**

Data from the following websites were used:

* https://www.kaggle.com/
* https://www.bloomberg.com/

From Kaggle, we got dataset from 2004 to 2018 that contains 200+ financial indicators, that are commonly found in the 10-K filings each publicly traded company releases yearly, for a plethora of US stocks. On average there are over 4,000 stocks listed in the dataset. We used the most recent data which is a file that contains information from companies in 2018. The dataset was saved into a file in .csv format. (File name: 2018\_financial\_data 2.csv)

From Bloomberg, we got stocks that are listed in the Standard & Poor’s 500 and from the NASDAQ, where the largest companies in the US are listed (large-cap). We got prices from those stocks from December 29th, 2017 to December 31st, 2018, a one-year period (2018). We calculated the 2018 return for each stock. The data set was downloaded and saved into a file in .csv format. (File name: Return.csv)

**Transform**

The data transformation was performed in Python within Jupyter notebook using panda’s library: The following data transformations were done:

* The first file, 2018\_financial\_data 2.csv, contains over 200 financial indicators, each one located in each column of the data set. We read the data from the csv file (pd.read was used) converted the data into a panda dataframe (df).
* We selected some of the data that we considered important when decided in which company to invest: ticker, revenue, net income, earnings per share (EPS), dividend per share, Return on Equity (ROE), debt ratio and market capitalization and we renamed the columns for a friendly reading (df.rename method was used).
* For the second file, return.csv, we did something similar. We read the data (pd.read was used) and converted into a pandas dataframe (df) and renamed the columns for a friendly reading (df.rename method was used).
* As we are interested in the largest companies per revenue, we sorted the financial data dataframe by the largest revenue. (.sort\_values(by=’revenue’, ascending =False, method to sort was used).
* In order to keep the index of the columns in the right order, we reset the dataframe by index (.reset\_index(drop=True , reset) index method was used).
* As we are interested in the top 100 largest companies by revenue, we reduced the data frame to the top 100 ([0,101], was used).
* We merged the two data sets using the ticker name as it was the column in common between the two datasets (pd.merge on=ticker method was used. When using the way to merge, if how = “outer” we get the whole list of companies, over 4000. Since we are interested in the top US companies, how=”inner” was used).
* A formatting of the data was done in each column for a friendly reading, adding “$” (.map(“${:,0f}”.format) and “%” (.map(“{:.2f}%”.format) when appropriate.
* The final data set is name final\_financial\_return\_data.

**Load**

Postgres (SQL data base) was used to upload the data. SQL databased was preferred as there is a predefined schema and it is a table based databased. There is no need for a hierarchical data storage in this project where a NOSQL database like Mongo would be better. So a structured query language was favored over a NOSQL database.

A table called financial\_data was created in Postgress with the following columns: Ticker, Revenue, Net\_income, EPS, Dividend\_per\_share, ROE, Debt\_Ratio, Market\_cap, Company\_name, Stock\_price \_as \_of 12292017 and Stock\_prce\_as\_of\_12312018.

A confirmation of the databased was uploaded was performed in order to make sure the database was properly stored (pd.read\_sql\_query method was used)

In the output the top 100 US largest US companies and its stock return in 2018. Those with highest return in the list are the companies that investors should not miss to look when thinking about investment opportunities.