Faculty of Economics

The Use of Python & Machine Learning to Optimise a Portfolio of U.S. Small-Cap Companies

Author: Alfonso Rosa Sánchez
Director: Mahmoud Aymo

MADRID | June 2020

Table of Contents

Figure Index

Abstract

This paper deals with the disruption of data analytics and machine learning in the investment management industry. In particular the small cap industry, which will be deeply analysed and studied in order to find interesting investment opportunities. This mentioned opportunities will be found from the benchmarking index of the small cap universe , the S&P 600 Small cap index which will also be fully studied in the paper. Once all of the 600 components of the index are analysed the portfolio constructing process will allow the creation of a diversified portfolio which will be able to beat the market.

The prementioned process is a combination of the most important theories in the portfolio management history, which will be combined with a fundamental analysis to help the portfolio maximise returns. The process will start by using the k-mean algorithm to find 3 clusters of well diversified stocks. Once the groups are clear the stocks will be analysed using the Fama & French 3-Factor Model to deeply understand the nature of the success of the portfolio. Then Joel Greenblatt´s famous magic formula will be adapted in order to fit in the small cap investment universe. Finally , when the magic formula has selected the 30 stocks which will compose the portfolio, Markowitz´s efficient frontier model will identify the most efficient distribution of the 30 stocks in the portfolio, in order to find the portfolio which maximises Sharpe ratio thus the reward investors receive for the undertaken risk in the portfolio.

Once the portfolio has been constructed the results will be compared and contrasted with the index performance as a check of the well-functioning of the new magic formula

Key Words: Small Caps, Python, Data Analytics, Machine Learning, K-means, Clustering, Fundamental Analysis, Portfolio Management, Efficient Frontier, Sharpe Ratio, Fundamental Analysis, Joel Greenblatt, Fama & French and Markowitz

1. Introduction

This paper being presented will focus on developing a disruptive strategy to outperform the small cap market index by using data analytics and machine learning in order to analyse, rank and select the appropriate stocks for the investment portfolio. The new methodology will provide the investor with a simple strategy which includes a fundamental analysis of the quality of the selected companies, so that only the most successful business are included in the portfolio. Therefore, as a brief overview, the main goal of the paper is to construction of an investment portfolio of small cap companies, which uses fundamental analysis to select the top-rated stocks. The analysis performed by data analytics will reduce the hypothetical management fee costs needed to pay analysts and therefore will be able to compete with ETFs at the same cost. This last statement will require the construction of an efficient portfolio which maximises Sharpe Ratio, so that investors undertaken risk is correctly rewarded.

In order to achieve this efficient portfolio, the benchmark index of the small cap universe, the S&P 600, will be fully analysed to select only in the stock which can offer the maximum returns in the long run. This will require a thorough process, first of all the k-means algorithm will be used to find 3 groups of stocks which are well diversified as a strategy to hedge against possible bear markets. Then once the groups are identified the Fama & French 3-Factor model will be used in order to understand which fundamentals should be used to rank the stocks from the 3 different clusters. Once, the fundamentals to use in the analysis are identified, Joel Greenblatt´s famous magic formula will we adapted to the small cap universe in order to rank the stocks and only the top 10 stocks from each cluster will be used for the investment portfolio. Finally, in order to determine the weight of each stock in the portfolio Markowitz´s efficient frontier theory will be applied in order to select the portfolio which achieves the maximum Sharpe ratio.

Once the portfolio has been created, a statistical study of it will be compared to the benchmarking index to assess the efficiency of the portfolio and the effectiveness of the proposed investment methodology.

The creation of this methodology thanks to the programming language python will increase the efficiency of the portfolio construction process, this cut in both cost and time will mean generate a more attractive product which will have extremely low management fees but will contribute to the alpha creation of the portfolio. ((ETFGI, 2016) According to the firm ETGI, in 2019, $4.569 trillion was invested in passive ETFs, this big market means that there is the opportunity to create a fund which despite being passively managed it is able to create alpha and therefore outperform the market regularly.

2. Small Cap Investment Universe Analysis

2.1 Small Cap Companies

Defining small caps stocks can be quite ambiguous as it has been modified several times across history. Currently, small cap companies are known as companies with a market cap , share price x nº of shares between $3,000 million and $2 billion. This group of stocks approximately amounts to 10% of the total stocks in the US markets, this percentage may vary because of market conditions and constant shifts in share prices which alter companies market cap. figure 1 shows the evolution of small caps companies in the US.

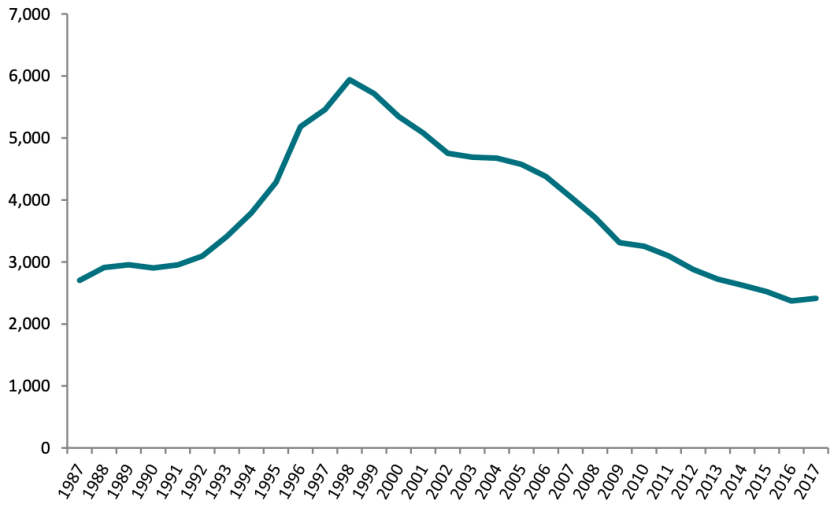


Figure 1 Nº of Small Cap Companies in the US 1987-2001.

Source: BGF Asset Management

The benchmark used to measure the performance of small caps is the SP 600 Small Caps index which will be deeply analysed later on in this paper.

Small-Caps stocks are a special asset class and it is characterised by the following features:

⇒ **High Returns:** Historically small-cap shares have been a really attractive asset class for portfolio managers world-wide. This kind of companies offer a huge upside growth potential which translates also into huge profits for investors, reflected in the low P/BV of companies in this sector.

⇒ **High Risk**: An investor seeking high returns is also assuming a greater risk, therefore the great growth potential of small-cap stocks comes in hand with a great downside risk which could be translated in big loses.

⇒ **Long-term Investments**: Companies growth is slowly reflected on the company's market value. First of all, new upcoming strategies are slow to be implemented and therefore be reflected in earning of the company, besides, market value also reflects investors' appetite in this specific business and in order to grow this appetite amongst investors companies have to be well stablished and have strong and regular earnings.

2.2 Small Cap Companies V.S Large Cap Companies

In order to analyse and quantify the veracity of this factors it is essential to quantify and compare them to another benchmark, therefore this factors will be analysed numerically and compared to large-cap stocks, a data-set with the largest 20 stocks by market-cap and the smallest 20 stocks by market-cap, will be compared.

2.2.1 Returns

Returns are calculated as the percentage difference between prices each month, small-caps will account to the lowest 20 and large-cap to the largest 20:

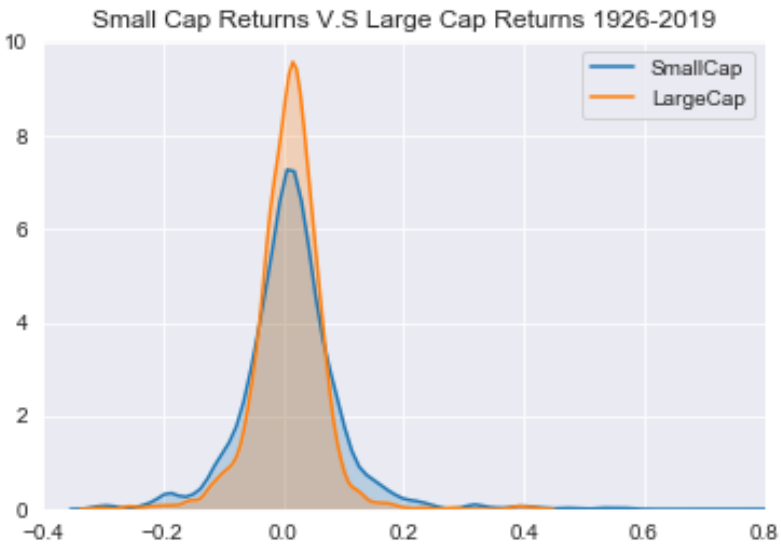


Figure 2 Small Cap Returns V.s Large Cap Returns

Source: FactSet

	Annualized Return	Annualized Vol	Skewness	Kurtosis	Cornish-Fisher VaR (5%)	Max Drawdown	Sharpe Ratio
SmallCap	0.151977	0.336701	3.629829	38.285414	-0.049569	-0.867228	0.352273
LargeCap	0.098490	0.195116	0.345472	11.847243	0.067464	-0.858502	0.341408

Figure 3 Small Cap Returns V.s Large Cap Returns

Source: FactSet

Returns had to be annualised in order to easy comparability amongst each other and the results are clear, the historical average returns from 1926 to 2019 of small caps is substantially higher than the returns from Large Cap stocks (15.20% against 9.85%). Besides the greater average returns from small-cap companies there are also two key parameters to be interpreted: skewness and kurtosis.

2.2.2 Skewness

This statistical parameter is widely used to see how a certain stock or portfolio is preforming. It basically is the degree of distortion of the set of returns from the symmetrical bell curve or the normal distribution.

The normal distribution curve is characterised by having a zero skew, therefore values higher than zero will mean a positive skew and vice versa. In a normal distribution, mean, median and mode are in the same place, whereas in positive skews means are greater than the median, this results in having a fatter tail tilted to the right, which returns-wise means having grater positive returns than negative. This sense of fatter tails can be seen in figure 2 where tails are thicker towards the right (positive returns) than they are on the left side of the distribution curve.

Skewness goes beyond focusing on averages and median values, it examines at the extreme value of returns, therefore the fact that SC have a higher skewness than LC, shows that extreme values are positively tilted and have more positive results than negative results, which proofs that historically returns from SC haves been more consistent than from LC.

2.2.3 Kurtosis

Another statistical figure used to describe and helps to understand the distribution of the returns of the two examined portfolios. On one hand skew talked about how tilted towards the right or left where the extreme values, in other words if tail was fatter towards the right or to the left; kurtosis examines both tails extreme values and how many standard deviations they differ from normality. There are three type of kurtosis

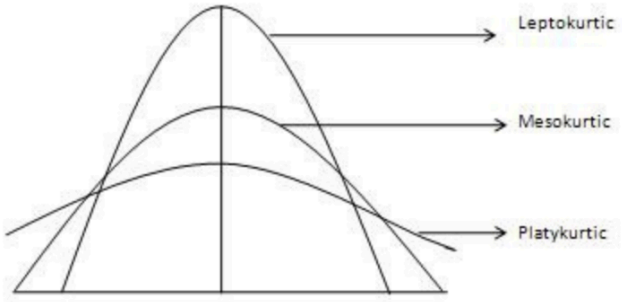


Figure 4 Kurtosis.

Source: CFA

Mesokurtic, where returns will have approximately a normal distribution with few outliers and a total value of kurtosis of 3. The next will be platykurtic(< 3), with a reduced number of outliers and therefore shorter tails. This type of kurtosis can be find in fixed income investments where result don't deviate from normality and therefore are easy to predict. Finally, leptokurtic distributions (3 <), will have a greater number of extreme values hence fatter tails and complicating the prediction of this stocks, high volatile stocks tend to have greater kurtosis.

In this case, both SC and LC have leptokurtic results due to their asset class and volatility, however it means that result from LCs are concentrated towards the mean of 9.85% annually, however SCs have greater number of extreme values, this tends to be a negative aspect for stocks, nevertheless, having in mind that skewness of SC´s was greater than from LCs and extremely positive , shows that despite there is a greater number of outliers; there is a greater chance of this outliers being positive.

2.2.4 Risk

Another key feature to analyse from stocks is the risk associated with each specific company, it is complicated to quantify the risk associate with each company as each business model requires special fundamentals to develop their business activity, different debt to equity values, lower working capitals or even less profits than others. Therefore, the risk associated to each stock, known as its volatility, is calculated by computing the difference between the returns from normality over a period of time, or in other words its standard deviation. As it was to expect the greater uncertainty of small caps companies which still have to grow their business models translates into higher volatility, whereas better stablished LCs have a lower risk associated with them thanks to their established business models and their regular cash flows generation, stabilising share prices.

	Annualized Volatility
SmallCap	0.336701
LargeCap	0.195116

Figure 5 Volatility

2.2.5 VaR

Volatility is the main value when referring to risk, but in order to find out how risky a certain asset or portfolio of assets is value at risk (VaR) is extremely useful. VaR widely used amongst the financial industry to measure the potential maximum loss of the portfolio of assets and the probability of that maximum loss occurring over a certain time frame. For the purpose of this paper a probability level of 5% will be used. This is widely used amongst portfolio managers in the industry and allows to quantify the investment risk of the whole portfolio even when there is a downturn in the financial markets. Figure II shows the results for both LCs and SCs.

As it is proven in the results, SCs have a VaR of -4.957%, this means that the worst possible return of a portfolio of small caps 5% of the times will be this figure, on the other hand 6.75% of the LCs is much higher an even positive proving that the consistency of the returns of the LCs has been high throughout history. This is due to their consistent and well stablished business models allowing each company to generate recurrent cash flows which increases the stability of their share price and therefore their lower annualized volatility, making them ideal for investor with smaller risk appetites and willing to achieve consistent returns on their portfolios.

2.2.6 Maximum-Drawdown

The final parameter to compute the risk of a pool of assets is to calculate the maximum drawdown, which computes the maximum loss an investor would have experienced if the money was invested at the peak and had sold the portfolio at the through. It is true that it is not efficient to capture the consistency of the returns of the portfolio, but it allows to portray how would the worst outcome the portfolio will experience in case there is a severe collapse in the financial markets.

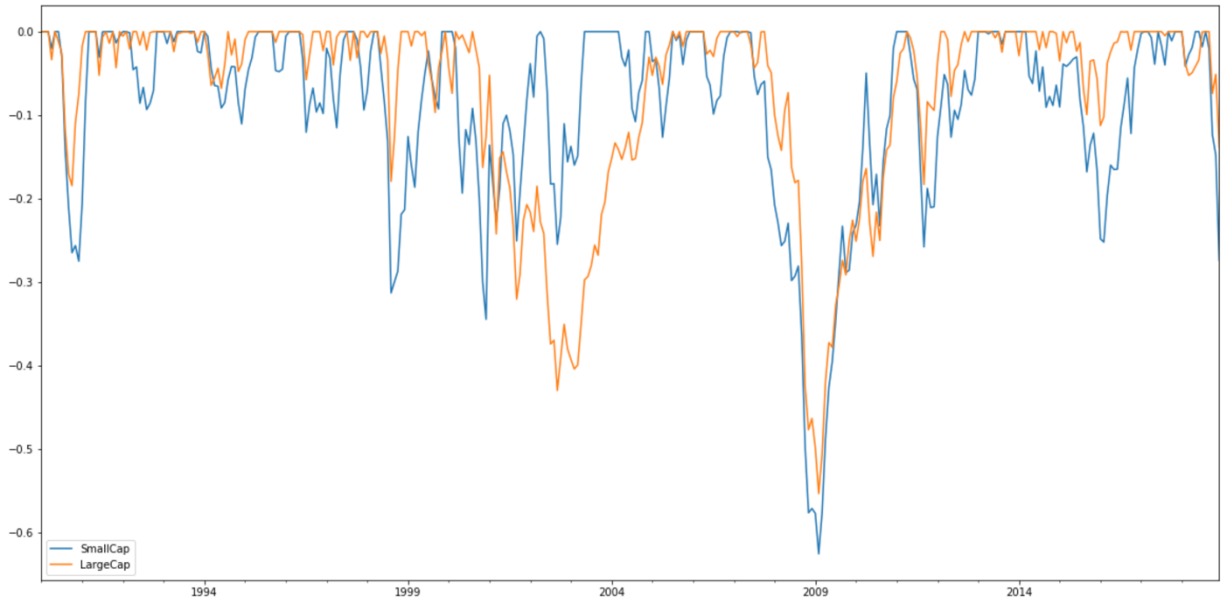


Figure 6 Max-Drawdown

Figure V represents the drawdowns the portfolios have suffered from 1990 onwards, it is proven that there is a great correlation amongst each assets class, nevertheless, it is evident that during bear markets normally the SCs had a greater drop in their price, however that hasn´t been always the case as in the dot com bubble at the beginning of the 21st century, large caps dropped nearly double than what SCs did. Finally, despite their differences the values figure I prove that the difference in their maximum drawdowns is minimal, being the value of SCs 86.7% and of LCs 85.9%, the difference of 0.8 clearly states that when fighting against a market contraction both assets behaved in an extremely similar way.

2.3 Risk Adjusted Returns

Once both risk and returns have been detailed it is essential to compute whether the level of risk assumed by the investors is worth the returns. In order to compute this, William F Sharpe, one of the greatest economists in world´s history wrote *The Journal of Portfolio Management,* here he explained the Sharpe Ratio. *This ratio computes the excess return rewarded to shareholders over the risk-free rate and per unit of volatility or risk assumed*, figure VI shows the formula

$$S = \left(\frac{R_p - R_f}{\sigma_p} \right)$$

Referencing figure III, SCs have a Sharpe ratio of 0.35 whereas LCs of 0.34. This shows that the portfolio of SCs is more efficient as it is able to acquire a greater return for the risk assumed by itself. Therefore, this creates an opportunity as if portfolio of SCs is managed efficiently acknowledging each stocks risk it would be able to reward investors with higher returns than if they invested in a LCs portfolio, therefore this is one of the main reasons to focus on this asset class to create the portfolio.

2.4 The Wealth Index

In portfolio management it is a priority to compute the parameters previously described in order to allocate assets in a successful manner, it is essential to control risks and hedge against possible downturns. Nevertheless, what the most important figure is the result and how much money the portfolio holder ends up with. figure VII shows the evolution of $ 1,000 the 1st of January of 1990 and had cashed out on 2019.

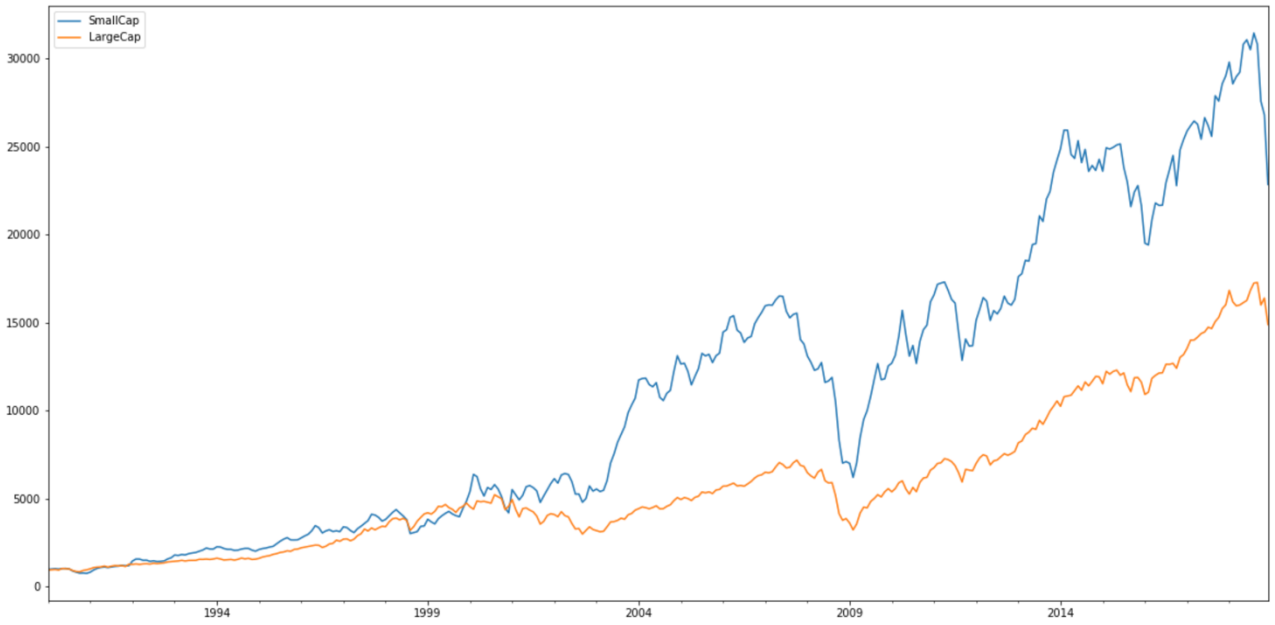


Figure 7 Wealth Index SCs & LCs (1990-2019)

The previous graph clearly shows the spread between the results achieved by LCs which managed to obtain $14,886 from 1990-2019 whereas SCs managed to achieve $22,854, this means that SCs where able to achieve 154% times more than a portfolio of LCs. This clear spread amongst asset classes reflects a clear investment opportunity as if there is a thorough analysis of the companies inside the investment universe, there is a great change to outperform the market and achieve outstanding results. Therefore, this paper will analyse and investigate the core of companies by analysing fundamentals and pick the stocks which have the most upside potential and trade at bargain prices.

3. Benchmark Index

3.1 Index Overview

Once the investment universe is clear and deeply analysed it is essential to focus on a benchmark to track the performance of this particular market. Despite the previous analysis of the 20 largest and 20 smallest companies trading in the US market, was useful to represent the differences and highlight the investment opportunities inside this market segment, a more representative index which englobes a higher number of the population of stocks in this precise market. This is why in this paper the portfolio the S&P 600 SC Index. The index tracks the performance of 600 stocks with a market-cap between $600 million and $2.4 billion, this index was chosen over its competitor the Russell 2000 index as it has a more faithful representation of the SC investment universe due to the fact that the Russell 2000 tracks the performance of 2000 stocks ranging from $169 million to $4 billion, this wide spread of market-caps would not be a faithful representation of the market and hence not useful for the study.

3.2 Index Criteria

In order to analyse the performance of the index and to create a portfolio to beat this index it is essential to know how it is created and how stocks are weighted amid the index. The stock selection was clear and previously stated, choosing from liquid companies trading in the US market with markets caps or what is the same the current price of all of the outstanding shares in the market ranging between $600 million to $2.4 billion, once this filter is done , in order for a company to be eligible for the index, the following criteria must be met:

⇒ ***Public Float***: Refers to the number of shares that the company has readily outstanding in the market or in other words the number of shares available for investors immediately. This value must be over $300 million, this is a liquidity restriction which ensures that there is liquidity for investors in case they are willing to sell their stake inside the company, this generates a security for investors and makes that companies with poor liquid situations are kept away from the analysis. Reducing in the long run the risk and hence volatility of investing in this index. (S&P Dow Jones Index, 2020)

⇒ ***Financial Viability :*** Companies must have reported positive earnings in the last quarter, as well as having positive earnings reported in the sum of the last 4 quarters (full year). This again is a defence measure for investors as it only includes companies who generate positive earnings and therefore are able to meet debtors and hence producing cash flows to shareholder in the form of dividends. This will not only create value for shareholders but will assure a high share price in the long run and also it will assure assuring that companies have a high-quality profitable business model, again reducing the risk of the portfolio. (S&P Dow Jones Index, 2020; S&P Dow Jones Indices, 2020)

⇒ ***Adequate Liquidity and Reasonable Price:*** Using composite pricing and volume, the ration of annual dollar value traded (defined as average closing price over the period multiplied by historical volume) to float-adjusted market capitalization should be at least of 1.00, and the stock should trade a minimum of 250,000 shares in each of the six months leading up to the quarter evaluation date. Again another measure to reassure even more that companies have a high liquidity and prices reflect true value of companies which will decrease the price volatility thus risk in the index. (S&P Dow Jones Index , 2020)

⇒ ***Sector Representation :*** In order to have a faithful representation of the whole market segment, the index diversifies its components in 11 industries (GICS sectors). The weighting of the sectors, shown in figure 8, shows the portfolio is well diversified the having no more than 17.7% invested on a single sector, the

diversification, a strategy developed by Henry Markowitz, allows the portfolio to reduce the overall market risk which reduces the volatility of the index. Obviously the index segment leaders are Industrials, Technology and Financial Services which compute 48.3% of the whole index, despite it might seem quite of an overweighting in this three sectors, it is a trend in most of the major indices like the S& 500 where this three sectors are dominant, this due to the fact that these areas concentrate the most number of companies

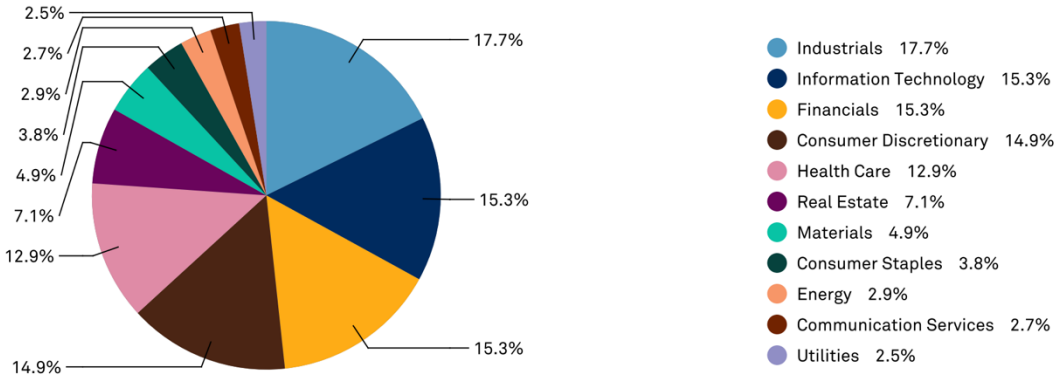


Figure 8 SP 600 Sector Weighting.

Source: S&P Dow Jones Indexes

⇒ ***Company Type:*** Despite many companies are inside the constraints established of size, it is clear that the index includes only financially stable which have long term growth projections. However, it is also essential for the fund to choose from companies which really reflect the evolution of this particular sector therefore, it is selecting certain companies inside the whole range of assets is key for this particular goal. Hence only US common equities listed in the US are eligible, inside this big investment universe there is a special type of listed

companies known as real estate investment trusts (REITS) , which are mutual funds composed of a pool of investments in real estate projects. Due to the fact that real estate generates $ 1.15 trillion (2018) , or in other words composes 6.2% of the whole US GDP, REITs are a tool to reflect this industry in the index. Excluded from the eligibility are closed-end funds, ETFs, ADRs and ADS are ineligible for inclusion.

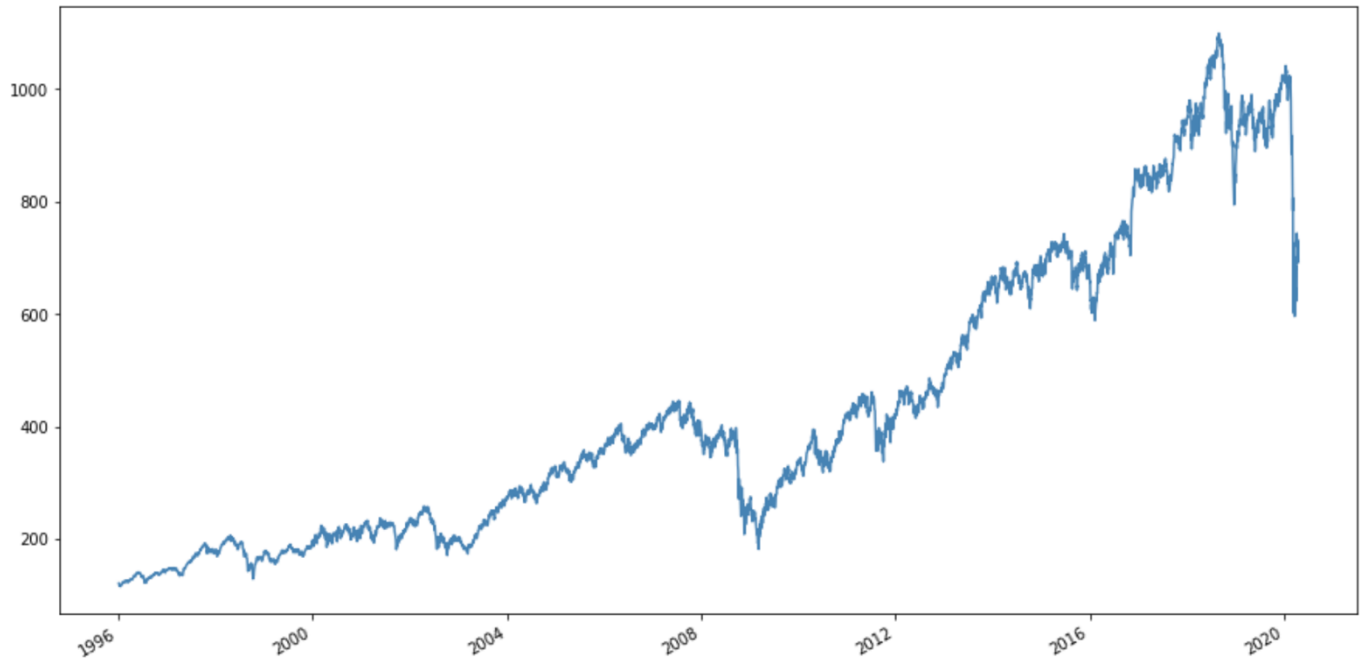
3.3 Weighting Criteria

After the companies which are under the constraints established by S&P, the question comes on how much each stock should weigh in the index. Not every single stock is as important as other and therefore each stock doesn´t have the same market share as others. In order to try and reflect this on the index S&P use a float adjusted market cap weighting methodology basically, adjusts the weight of each stock regarding their market capitalization, for example, an index composed of two companies A with $8 Billion of market cap and B with $2 Billion, A will have a weight of 80% whereas B will have is 20%, this will make that fluctuations in A have a greater impact on the whole performance of the index, for example an increase in the price of A of 10%, will mean an increase in the fund of 8%, whereas an increase of B in 10% will only impact the index by 2%. Despite overconcentration can bias the fund in only focusing on the biggest and usually most successful companies, it is just a true representation of the importance each company has on the whole economy.

Due to the market´s volatility and the day to day fluctuations of stock prices which amongst SCs stock is much higher than over other LCs present in other indexes like the S&P 500, the rebalancing of the fund is carried out quarterly in order to update the weighting of each component and have a faithful and clear image of how the performance and fluctuations in the whole market are being affected.

3.4 Index Performance

The index was created the 28th of October 1994, having its inception the 30th of December 1994, since then it has experienced two severe stock market crisis and it is currently immersed in one of the worst market crashes in the history of the stock exchange. It has been able to survive the dot-com bubble burst of the beginning of the 20th century, where the Nasdaq (tech company index of the US) crashed by 78% in October 2002. It has also been able to survive the great financial crisis of 2008 where the financial system had a turning point and due to the speculation of many funds and an evident deregulation of the system markets fell drastically. This extremely damaged this index, where it suffered a drop of 53% (figure 10), being its worst drop ever. Since then, as seen in figure 9 it has been a complete rally, rising from 150 at the end of 2009 and even reaching 1000 at the beginning of 2020, nevertheless, it is evident that this recession has extremely damaged the performance of the fund this year having a complete drop of 44%. Nevertheless, markets are quite uncertain in the near future due to the questions and insecurity the outbreak of the pandemic has caused amongst investors and portfolio managers, however a positive resolution of the COVID crisis could result in a rally in the last quarter of 2020.



Once the performance of the fund is analysed it is essential to study the statistical figures of this performance in the lifetime of the index. For this purpose, the same procedure as in the analysis of the SC and LCs was carried out. The data is represented in figures 10,11 and 12.

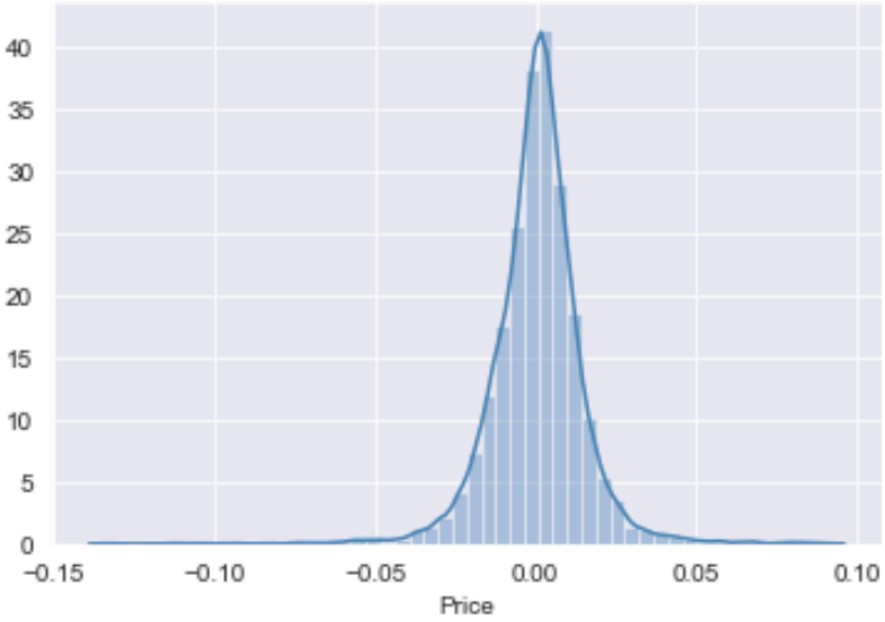


Figure 9 S&P 600 SC Performance 1994-2020

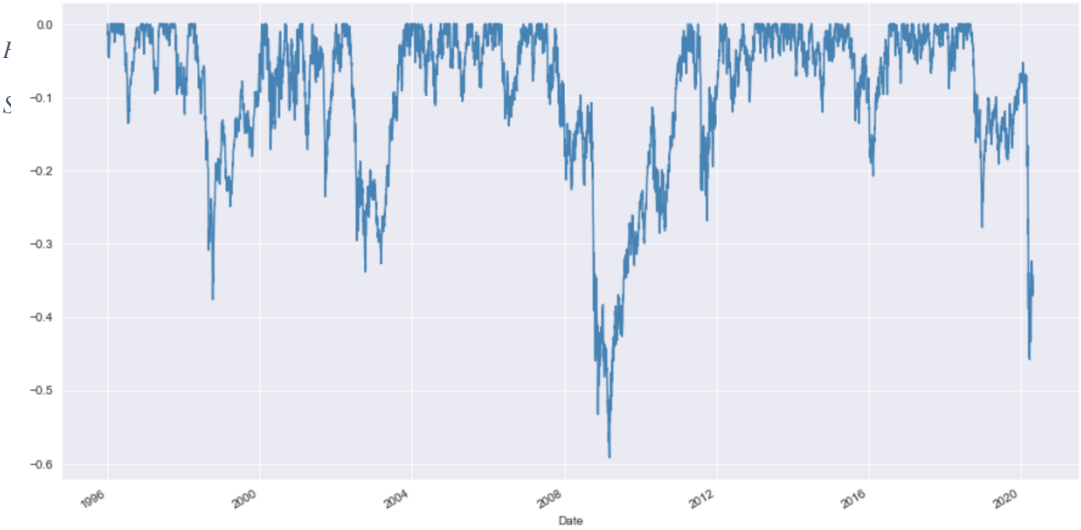


Figure 11 Max-Drawdown S&P 600 SC

Source: FactSet

	Annualized Return	Annualized Vol	Skewness	Kurtosis	Cornish-Fisher VaR (5%)	Max Drawdown	Sharpe Ratio
SP 600 SC	0.003428	0.048997	-0.362737	10.554947	0.022145	-0.591648	-0.52786

Figure 12 Statistical Figures S&P 600

Source: FactSet

⇒ ***Skewness:*** As previously disclosed the skewness pictures the degree of distortion of a set of results, in this case figure 12, shows that the index has a negative skew. This proves that the performance of the stocks at extreme values has a greater tilt towards negative figures than to positive ones , which if compared to the higher skewness of the SC portfolio, proves that if a portfolio was correctly managed, skewness could be improved and therefore extreme values will tilt towards the positive side rather than staying negative. Figure 11 clearly portrays the skewness as the distribution curve of the performance of the index has a "fatter" tail towards the negative side.

⇒ ***Kurtosis:*** Despite having a smaller Kurtosis than both the SC and the LC portfolio, the index has still a leptokurtic distribution curve. This is due to the fact that the portfolio of SCs and LCs was composed of 20 stocks, whereas the index is composed of 600 different companies, which reduces the overall volatility of the portfolio and it is why returns are more concentrated towards the mean in the index portfolio.

⇒ ***VaR:*** Similarly, to what occurs with kurtosis the great level of diversification of the index reduces the overall volatility of the portfolio, as a greater number of assets decreases the market risk exposure of the whole portfolio. This is why, the VaR of the index is positive and higher than the SC portfolio.

⇒ ***Sharpe Ratio :*** In this case the index has a negative Sharpe ratio, this doesn´t directly mean that the performance of the fund is negative, it actually refers to the fact that the index is taking too much risk for the little reward it is obtaining as despite the high level of diversification the stocks in the portfolio are still SCs denoting their high volatility.

If all of these four statistical figures are analysed and compared to the investment opportunities that the SC universe offers (portrayed in figure 3), there is a clear sense that if the stock picking process is improved by going beyond market capitalisation, but also keeping in mind fundamental factors of the companies invested in, extremely high returns could be achieved. This is process of selecting and properly weighing each stock in the portfolio is what will be analysed in the proceeding sections.

4. Portfolio Construction

The goal of this theses is to create a portfolio that beats the index and hence achieves. A better performance than the index, in order to do so, it is essential to carefully select the appropriate stocks and the appropriate number of stocks in order to create a well-diversified portfolio which reduces the market risk amid at the same time achieves to maximise the returns of the portfolio.

4.1 10-year Historic Performance Requirement

The first step which must be carried out in the creation of this new portfolio is to filter the appropriate number of stocks which can be considered to enter the new investment. Therefore, in order to develop a deeper analysis of the correct stocks, some of them have to be removed from it. (Markowitz, 1976) However, taking a look back to the statistical analysis developed in *figure 12* which shows that the index despite the high diversification has a low Sharpe ratio even negative which denotes that the returns of the index are lower than the risk it is assuming. This phenomenon is known as overdiversification which occurs when the number of assets in the portfolio exceeds the point where the marginal loss of the expected return is much greater than the marginal benefit of holding one extra asset in the portfolio.

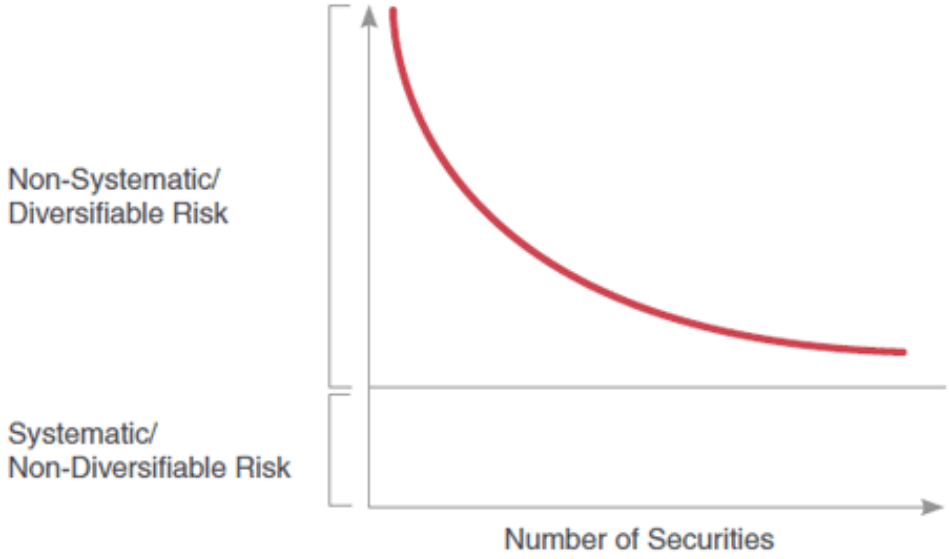


Figure 13 Overdiversification

Source : CFA Level 1 Portfolio Management

In order to solve the overdiversification of the new portfolio , the number of stocks under consideration will be reduced. The first filter in order to eliminate stocks will be to only consider companies which have been listed for over 10 years in the stock exchange. Because the index deals with small caps company there are many companies which are eligible to enter the fund once they go public, also known as developing an IPO, in fact, only in 2019, 159 companies went public in the United States of America which creates a really big bias in the index and creates a gap of data performance.

Reducing the number of stocks considered to only those which have been listed for over than 10 years has many benefits for the portfolio creation. First of all, the statistical relevance of the data is much higher, this is because one of the main aspects to analyse from stocks when creating a diversified portfolio is looking at the correlation amid each other (which will be developed further on), therefore if a stock has only a record of 1 year whereas other stocks have been listed and included in the index for over 15 years, creates a really low correlation. This will create a big bias in the stock analysis as even though there was a really correlation amongst these hypothetical stocks the lack of past

performance and evidence will reduce the correlation close to 0, which will generate a bias in the stock selection as according to the analysis these stocks will have a really low correlation but in reality their performances will be really similar, which will cause the portfolio that when there is a drop the drop will be even more severe due to this fake correlation. Nevertheless, it is also true that in case there is a positive boom in the market both assets will contribute to even bigger gains in the portfolio, however this is extremely risky and the main goal of the creation of the portfolio it is not only to generate maximum returns but also reducing the risk of the investment which in other words could be translated to trying to maximise the return for every unit of risk or maximise the Sharpe ratio, which in the long run it is a much more efficient strategy.

Setting aside statistical reliability of the data analysed, the fact that only companies with a relevant history are included in the new portfolio also assures that these companies have a greater quality than those companies who either have not yet proven their efficiency and have also proven that they can survive adverse stock market situations. Regarding the efficiency and hence the quality of the companies invested in , Warren Buffet, one of the best (if not the best) investor and portfolio manager alive, always reinforces the idea that for a company to be eligible for his portfolio at Berkshire Hathaway, the company must have had a historic of 10 years or more, in fact, he recommends investors to look beyond 10 years. The extremely successful performance of Berkshire Hathaway achieving results over 600% in the last 20 years proves that this is an important aspect to consider when selecting stocks for the portfolio. Going beyond the fact that companies must be of a high quality, 10 years of historic performance means that the companies considered have gone through one of the worst crises in the history of financial markets, the great financial crisis of 2008, triggered by the collapse of Lehman Brothers and the bursting of the real estate bubble caused many companies to go bankrupt, therefore being able to survive this extreme situation denotes the strength and the ability of the company to survive complex situations, it could be treated as if these companies have been able to successfully overcome the worst stress test possible. The surpassing of this severe stress test is extremely useful specially for the current situation financial markets are suffering as due to the pandemic outbreak and the

economic lockdowns, it is essential to invest in highly liquid companies who can surpass this extremely tough situation which is not resolved yet.

As a sum up of the first filter in the portfolio construction process, filtering companies will create a new list of 343 which have reliable data to achieve a proper diversification in the new portfolio and therefore the goal of achieving a higher return per unit of risk assumed by the investments will be achieved.

4.2 Cluster Creation

At this step of the portfolio construction process there are still 343 stocks, despite this is 43% less than the total number of the stocks in the index, there is still overdiversification in the portfolio and the optimum number of stocks (which will be further analysed) is still not met. As previously stated in order to maximise the Sharpe ratio of the new portfolio it is essential to have a well-diversified pool of assets where the market risk is reduced to the lowest possible level. To achieve this extremely complex goal, the 343 remaining stocks will be analysed, and a correlation matrix or heat map will be produced in order to create 3 groups of stocks with a common correlation amongst each other in order to have 3 well diversified group of assets. This process will be carried out thanks to machine learning and the K-mean algorithm which will allow to create diversified group of assets automatically adjusting for changes in the input data, which will allow to rebalance the portfolio in a much more efficient, hence increasing diversification and reducing overall market risk.

4.2.1 K-Means Algorithm

Creating the different clusters or what is the same the group of diversified assets is one of the key processes of the portfolio construction as it allows the portfolio to maximise

the market returns in bullish markets but at the same time it also allows to cover and be defensive in bearish markets, like the current situation, where uncertainty is present and therefore achieving a better performance when share prices are drastically dropping.

In order to create these mentioned cluster, it is essential to work with the correlation matrix, represented in a heat map in figure 14. Despite not being really representative and data is really difficult to interpret due to the large number of variable (stocks) it is clear that stocks are really correlated and therefore a deeper analysis is needed to pick the correct diversified groups.

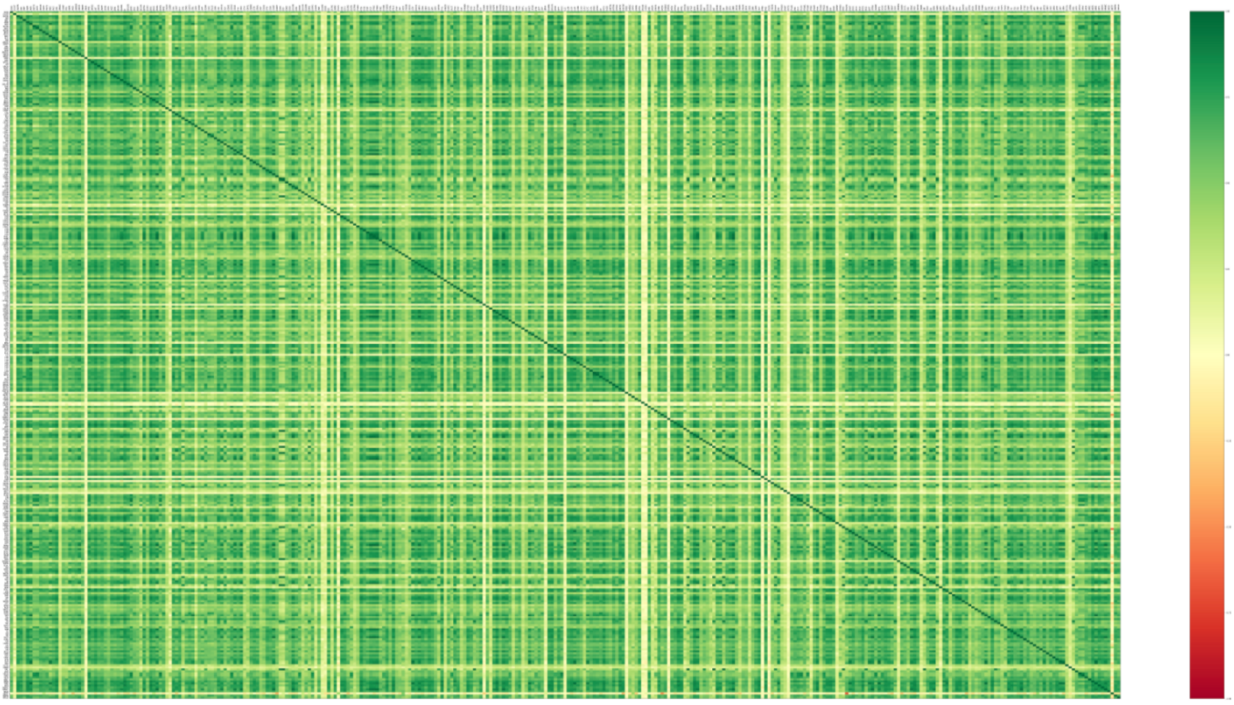


Figure 14 Heat Map of Filtered Stocks

Source: FactSet

The heat map is the starting point for the K-Means algorithm, which works in an iterative manner to assign to each stock one of the "K" groups based on their characteristics, which in this case are their correlations amongst each other. However, before getting into the algorithm and how it works there is a key step which is determining the value of K, or what it is the same, calculating the optimum number of clusters, to do so, it is essential to run the algorithm through the range of values of K see

and compare the features of the obtained groups. There is no exact formula to calculate the optimum number of clusters, however, a visual estimate using the elbow criterion is used in order to determine the optimum point. The elbow criteria stand for the shape of the graph which looks like an arm where the spike or elbow is the crucial point, figure 15 shows the optimum number of clusters for the portfolio.

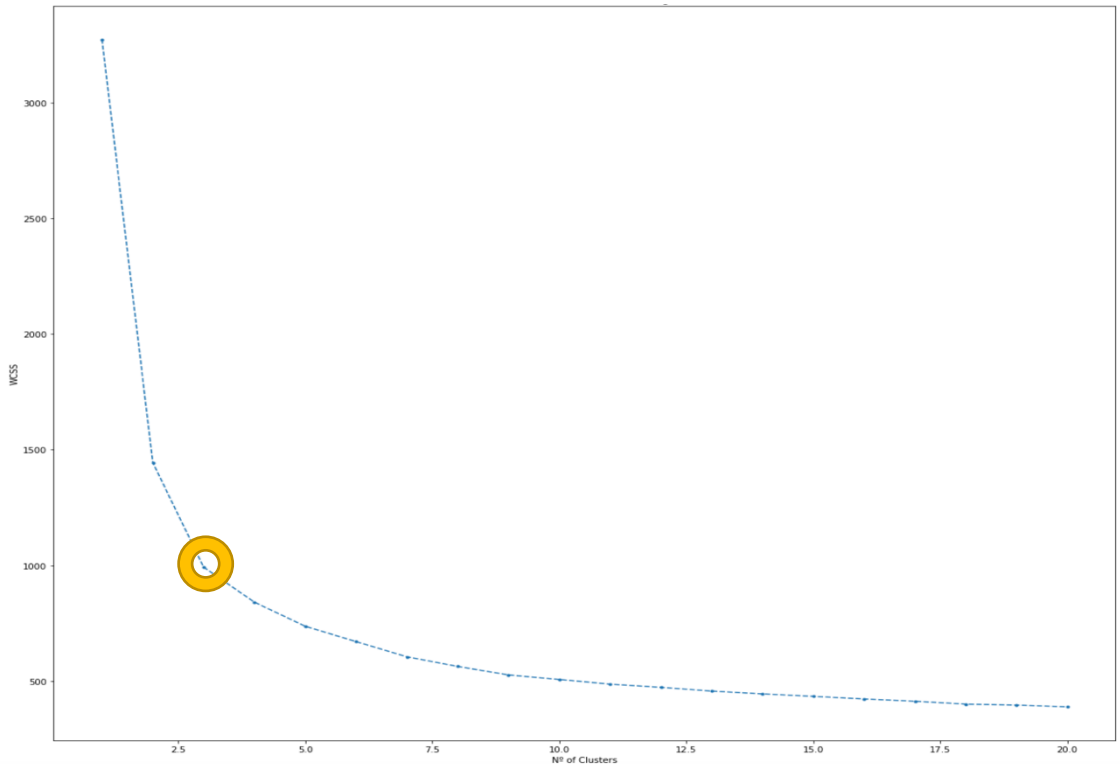


Figure 15 Optimum Number of Clusters

The orange circle in figure 15 shows the optimum number of clusters, 3, as previously mentioned, there is no exact numerical formula for it, instead the elbow of the arm shaped curve shows the optimum point.

Once the optimum number of clusters is identified, the stocks are group based on their common features and from running the algorithm the output will be:

⇒ ***Centroids:*** Each group or cluster has its own coordinates and a central point which will determine the assignment of it.

⇒ ***Labels:*** Each group will have a label which will allow the algorithm to assign a certain label (group number) to each stock.

The clusters will be defined in an organic manner where the centroids location will be adjusted as an iteration process till the algorithm converges to create the optimum clusters. It is essential to analyse the characteristic of the stocks in each group in order to draw conclusion from why each stock belongs to a certain group. The key part where machine learning helps to add efficiency to the process is the iterative process of the algorithm or in other words it constantly adjusts itself to create the final diversified groups. In order to run the algorithm, there are two key steps which must be developed:

⇒ ***Data Labelling:*** This is the process in which every stock of the pool of assets (in this case 343) are labelled towards the closest centroid which will be calculated by measuring the Euclidian distance, computed with the following formula:

$$\underset{c_i \in C}{\arg\min}\ dist(c_i,\ x)^2$$

⇒ ***Centroid Relocation:*** In this step the centroids of each cluster are recalculated, by calculating the distance of every assign stock in the previous step.

$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

This algorithm will iterate amid this two steps until one of the following criterion is met:

⇒ No changes amongst the cluster components.

⇒ The sum of the distance is minimised.

⇒ The maximum number of iterations in done.

One of the drawbacks of the algorithm is that the result can be optimum locally which means that it would be extremely convenient to repeat the whole process a couple of times with random starting centroids in order to achieve the correct unbiased results.

Once the whole algorithm is executed through the data set this is how the different stocks are structured into the 3 different cluster calculated using the K-means algorithm.

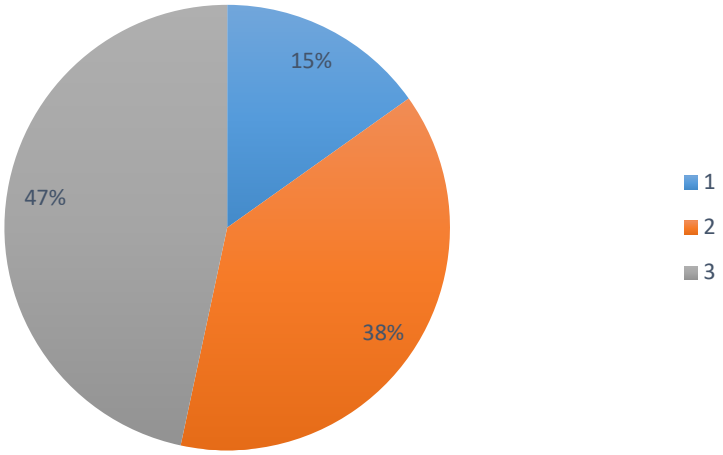


Figure 16 Clustering Distribution

This process of machine learning enables to divide the stocks into 3 differentiated group of stocks with common correlations hence making it much easier to diversify the portfolio and reduce the market risk. The clustering also proves that there is a big deal of stocks which are extremely similar (group 3 in figure 16) which count for 47% of the stocks in the whole portfolio, this is basically proving what the heat map in figure 14 was showing, there is a big portion of stocks which are highly correlated amongst each other.

4.3 Individual Stock Analysis

4.3.1 Fundamental Analysis

Once the three diversified clusters are defined, the correlation analysis between the price fluctuations is set aside, now the next step will be to look at the fundamentals of companies in order to assess whether the stocks are a convenient investment to beat the index.

Fundamental analysis is a technique used by analysts in the industry to calculate the intrinsic or real value of the studied asset, it can be used for both fixed income and equity assets however it is most commonly used for the equity products. One of its main assumptions is the denial of the efficient market hypothesis, which states that the share price reflects all available information and therefore are correctly priced. This assumption is also made in this study as the main goal is to generate alpha or what it is the same find mispriced stocks which trade at bargain prices in order to achieve a positive return in the future. Therefore, the assumption made by fundamental analysis is that in the long run prices will converge into the real intrinsic value, however the main issue which comes from this method is the fact that how long is the long run, however due to the fact that the portfolio created has a long term perspective, perfectly matches this requirement.

Fundamental analysis examines both economic and financial factors of a certain companies , by analysing ratios calculated from the firm's financial statements, which are : balance sheet, income statement and cash flow statement. This figures can be categorised into two main groups which are :

⇒ ***Quantitative:*** This are ratios calculated from the statements which are hard numbers there is no opinion of them they purely reflect the current reality of the

company, despite some adjustments could be made due to accounting issues carried out by each individual company. Examples of this type englobe profits, margins, liquidity ratios , ROE, Free Cash flow to Equity and so on.

⇒ ***Qualitative:*** Values which are related on based on an analysts value on an intangible asset this type of figures are dragged from the company's business model where the business shows its strategy, competitive advantage management team and corporate governance factors, which despite extremely important for the success of the company and the cash generation, but it is a complicated process to estimate.

As stated in the introduction, the goal of the thesis is to develop a new portfolio which beats the market index but doesn´t sum up to the management costs and therefore be able to compete with ETF funds which completely mirrors the index performance. The cost of analysing individual companies in depth and looking at every financial statement published must be eliminated. Nevertheless, in order to beat the index some company analysis needs to be carried out therefore, the study of each individual company will be done by analysing the quantitative fundamental factors, which will be discussed further on in the paper.

4.3.2 Fama & French 3 Factor Model

As previously mentioned, the key assumption undertaken in this paper is the fact markets aren´t efficient and there is still room for finding stock at bargaining prices which could benefit the investor from gains potential in the long run. Therefore, it is essential to select stocks which have the possibility to outperform the market, to do so, this paper will not deeply analyse companies, but it will carefully select fundamentals in order to find this mentioned bargain opportunities. To carry out this selection, the Fama and French model will be used to know which fundamentals to select.

The Fama & French Model was developed by Eugene Fama and Kenneth French in 1992 at the University of Chicago Booth School of business where both of them worked as researchers. The theory was an extension of the CAPM developed previously Jack L. Treynor, William Sharpe, John Lintner. Fama and French where pure defenders of the market inefficiency and wanted to prove that in the long run there was a real outperformance of value stocks and of small cap stocks (proven previously) therefore in order to reinforce their theory they included the following factors:

$$r = r_f + \beta_1(r_m - r_f) + \beta_2(SMB) + \beta_3(HML) + \varepsilon$$

Figure 17 Fama & French 3 Factor Model

⇒ ***Market Risk Premium:*** This factor is common in both models where the difference between the expected return of the market and the risk-free rate which for this paper due to the stocks are all in the US, the Risk-free rate used will be the US T bills yield. Which provides the reward each investor is receiving for the additional volatility it had to undertake for investing in the small cap universe.

⇒ ***Small Minus Big :*** This is the factor which wants to prove the size effect previously disclosed in this paper. It basically calculated the excess returns that small cap companies have on large cap companies, the coefficient of the equation will determine to which extent is this evident in the regression model, it is used to weigh the degree of success that small caps have over large caps.

⇒ ***High Minus Low :*** This factor analyses the value premium stocks have, measuring the spread between the companies which have a lower price to book

value, which are considered as value stocks, and a higher price to book value which are considered as growth stocks.

Once the theoretic concept of the 3-factor model is explained, it comes the time to analyse the results of the study carried out on the S&P600 index. The results of the OLS regression are shown in figure 18.

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.943
Model:	OLS	**Adj. R-squared:**	0.943
Method:	Least Squares	**F-statistic:**	1510.
Date:	Mon, 04 May 2020	**Prob (F-statistic):**	3.36e-169
Time:	19:38:07	**Log-Likelihood:**	812.77
No. Observations:	276	**AIC:**	-1618.
Df Residuals:	272	**BIC:**	-1603.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Mkt-RF	0.9768	0.018	54.002	0.000	0.941	1.012
Constant	-0.0007	0.001	-0.934	0.351	-0.002	0.001
Value	0.3689	0.026	14.238	0.000	0.318	0.420
Size	0.6961	0.024	28.597	0.000	0.648	0.744

Omnibus:	68.846	**Durbin-Watson:**	2.155
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	424.878
Skew:	0.826	**Prob(JB):**	5.48e-93
Kurtosis:	8.849	**Cond. No.**	36.6

Figure 18 Fama and French Model on the S&P 600

Source: Kenneth R. French

From the results dragged from the analysis it is essential to deeply observe certain observation which will create the basis of the fundamental analysis of the new portfolio. Firstly, in order to check the veracity of the analysis it is essential to analyse the R-squared value of 94.3%, this value tells how much of the performance of the stocks is explained by the 3 factors of the model. Which is a relevant figure as only 5.7% of the returns are not explained by the model. This creates a perfect path in order to select which fundamentals to choose in order to filter the 343 remaining stocks.

The first and highest coefficient is the market risk premium which is 0.9768 a figure really close to 1 which shows the importance of this factor. Recalling on what was previously stated, the market risk premium is a measure of the difference between the return of the market and the risk-free rate, this can be measured by analysing the stocks

beta. This figure is a measure of the unsystematic risk of a certain portfolio and its measures the relationship between the price fluctuations of the stock price and the market. Therefore, in order the beta of the stocks will be used as a filter factor for the final step in the stock selection process.

The next coefficient to analyse will be the value of high minus low, it is true that it is the lowest of the three but the fact that it is positive still reflects the fact that the value stocks outperform the growth stocks, figure 19 shows the evolution of the growth and value funds of the SP600.

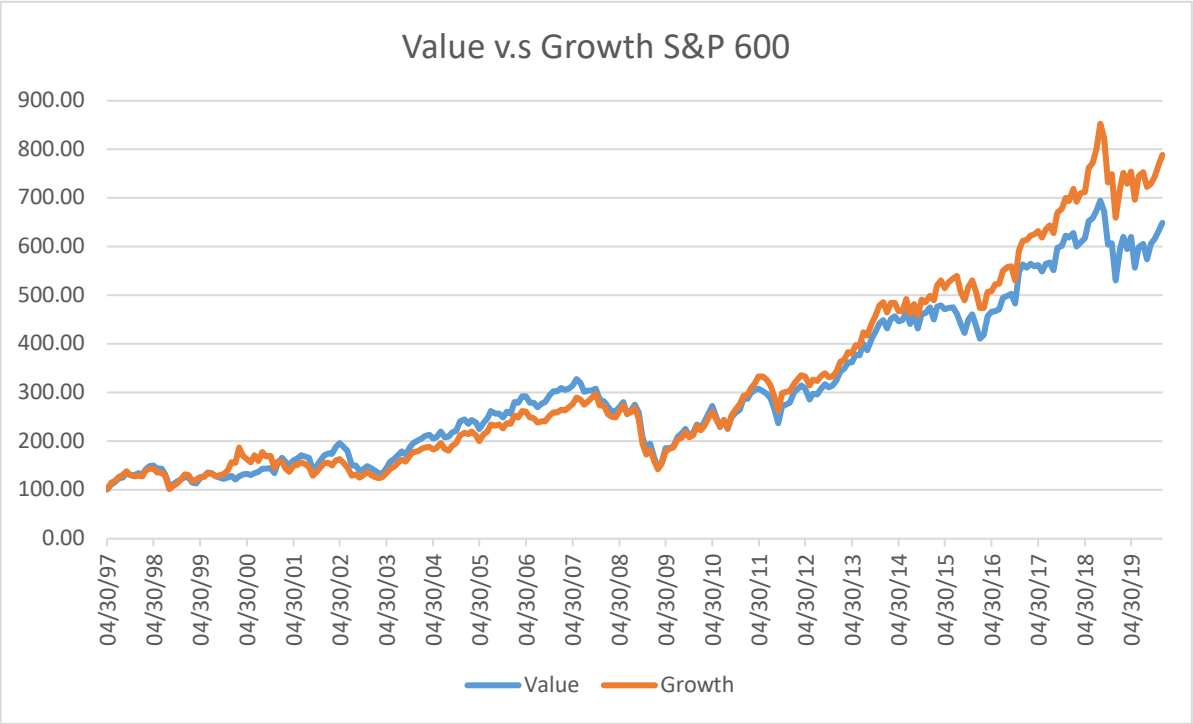


Figure 19 Value vs. Growth S&P 600 index.

Source: FactSet

Despite that in the las couple of years the growth stocks have been dominant over the value stocks traditionally before the great financial crisis and even after the dot com bubble at the beginning of the 20th century, value stocks leaded the charts. Therefore, according to our analysis value stocks should be overweighed in order to achieve a better than the market return.

(Merrill Lynch Bank of America, 2020)

On the other hand, value stocks, are companies that aren´t fully priced by the market, in other words, that the market prices doesn´t fully reflect the true intrinsic value and hence are trading at a discount. This bargain opportunity of being able to buy a stock which its value is not fully reflected by the market but also having the chance to profit also from the future growth of this company makes it a much more interesting investment. Specially if the investment horizon of the fund is taken into account as in the long run markets tend to reflect the true intrinsic value of the company, having this in mind markets aren´t always generous and their greediness and some of their irrationality can make a good companies share price trade at low prices for many years. This last assumption of markets not reflecting the true value of a company clearly matches with the inefficient market assumption this paper is undertaking.

Once the concept is clear, the key step is identifying this kind of stock , in order to add them to the possible stock options in the portfolio. The factors for identifying these stocks are ambiguous and many times subjected to an opinion, however for the sake of this paper the following will be used in order to identify the high-quality value stocks:

⇒ **Low Price to Book Value:** This is a ratio which is calculated as shown in figure 20, shows the relationship between the current stock price and the net value of the company (equity) to its market capitalization. In the cases where the market is overpricing a certain stock or investors are willing to pay a higher price for this type of stocks will be known as growth stocks. Nevertheless, those companies which have a lower values for this multiple, or in other words where the true value of the stock is not reflected by the market will be considered as value stocks.

$$P/B\ ratio = \frac{Stock\ price}{Shareholders'\ equity\ per\ share}$$

Figure 20 P/BV formula

⇒ ***Low Earnings Per Share (EPS):*** This ratio allows investors to identify the market value of the stock compared to its ability to generate earnings, in other words it represents how expensive or cheap is it for an investor to pay for each dollar generated by the company. If an investor has to pay a high amount (high EPS) for each dollar the company is generated , the company might be overvalued. On the other hand, if the company has a low EPS, it shows that each euro is invested in that company is worth more (earnings-wise) worthy hence cheaper. This is one of the main attributes which value investors seek amongst which Warren Buffet excels clearly. The fact that the EPS is low also shows that the appetite of other investors for that stock at that moment is low, hence when the market acknowledges the quality of the earnings the price will reflect the improvement. Again, it is essential to mention the fact that in order for the market to converge to the intrinsic value many years may pass and still the stock could underperform, therefore again the long-term investment horizon of the portfolio must be kept in mind. Figure 21 shows how the EPS is calculated.

$$EPS = \frac{Total\ income - Preferred\ dividends}{Outstanding\ shares}$$

Figure 21 EPS Formula

⇒ ***Low Debt to Equity Ratio (D/E) :*** This ratio is used by analysts and investors to examine the way a company finances its assets, in a way it is a way to measure how efficiently the company manages its leverage and if its growth comes from a reinvestment of equity and retained earnings or whether the growth comes from pure debt which despite being cheap nowadays with the low interest rates it is risky as the expansion of a company is always full of many uncertainties.

Nevertheless, it doesn´t measure the degree of growth in a company but it tells the investors how efficiently managed the funds inside a company are. Therefore as the purpose of this value tilted portfolio, it will be essential to select stocks with lower debt to equity which not only will reduce the possibility of underperforming the index, in the long run, but also it ensures to the investors that the companies invested in have a higher quality and are less exposed to possible bear markets and situations where debts cannot be met will cause severe damage to highly leveraged companies with high debt to equity ratios. Figure 22 shows the formula to calculate the debt to equity ratio.

$$Debt - To - Equity\ Ratio = \frac{Total\ Liabilities}{Shareholders\ Equity}$$

Figure 22 Debt to Equity formula

⇒ ***Free Cash Flow to Equity:*** It is a calculation of how much cash the company is able to generate after paying the operating expenses and the capital expenditures known as CAPEX. Therefore, it is a sign of the company's ability to generate liquidity and be able to generate sufficient cash to survive. One of the most known expressions by financial analyst is that "cash is king", as it ensures the company's success in the nearby future. In addition, generating cash is a metric which shows how much cash is going to be available for shareholders either in the way of a dividend or in a stock buyback. This last to, will generate value for the company and hence increase the price and returns for shareholders. Cashflow maximisation, will be one of the requirements for the assets in the portfolio ensuring the sustainability of the company and the increase in share price in the long run.

⇒ ***Return on Equity (ROE) :*** This is a metric used to evaluate the performance of a company financial-wise, calculated by dividing the net income of the company over the average shareholder´s equity, shown in figure 23. It is a way analysts can evaluate and measure numerical the efficiency to the generate profits of the

management of the company´s assets. As all the ratios there is not a good or bad figure it depends in every particular sector, however having a higher than the average ROE for a company shows that the management of the assets inside the company is being effective, hence , seeking the maximum number will be key for the stock selection process.

$$\text{Return on Equity} = \frac{\text{Net Income}}{\text{Average Shareholders' Equity}}$$

Figure 23 ROE Formula

The final coefficient to analyse from the Fama & French model will be the size factor, which directly refers to the theory that small cap stocks outperform large cap stocks in the long run. Therefore, for the portfolio construction and the selecting of the most efficient and profitable stocks, the small caps stocks will be chosen over the largest cap as according to the Fama & French analysis in the long run it will bring shareholders higher profits.

4.4 Stock Ranking Process

The Fama & French model allowed to pick up the most relevant fundamentals which have to be analysed from the constituent stocks of the S&P 600 index, which where the following:

- ⇒ Beta
- ⇒ Price to Book Value
- ⇒ EPS
- ⇒ Free Cash Flow to Equity
- ⇒ ROE

⇒ Market Capitalisation

Once this fundamentals are identified, one of the most important if not the most crucial part of the portfolio construction process comes into place.

In this investigation Joe Greenblatt's famous *Magic Formula* will be rethought and reconstructed for manging a small cap stock portfolio. This theory was developed in Greenblatt´s publication of the "The Little Book That Beats the Market" published in 1980, where he created what is known as the magic formula in order to make investors with no financial knowledge able to have a portfolio of stocks which was able to meet the market. The formula , however , was oriented towards the large cap stocks of the S&P500 index, where the first requirement was to set a minimum market capitalization for the stocks to analyse. The next step of the formula was to eliminate any stock which was in the financial and utility sector, he argued that the stocks in this sector where too correlated to the market and therefore they will usually perform as the market did, making it impossible to beat the market. In addition, the formula also excluded American Depositary Receipts which where foreign companies trading in the USA. Once the filters had been set , the formula ranked the stock by two main criteria, which where the earning´s yield of the company , calculated by computing the EBIT to enterprise value multiple. The next calculation required was the companies return on capital invested (similar ratio to the ROE) which was calculated by dividing the EBIT by the net fixed assets plus the working capital. Once this two ratios where computed stocks where ranked highest to lowest, further on, the investor should invest in the top companies in the ranking . Despite its simplicity it has been proven to be an extremely successful way to invest and to beat the market, in fact, Greenblatt claims that his portfolio has generated returns of over 30% annualised returns. In fact, he was the fund manager at Gotham Asset Management a portfolio which despite had a difficult start and stock price didn´t converge to the intrinsic value till 5 years passed , has demonstrated outstanding performances in its history achieving a return of 32% annualised, which actually is outstanding.

Despite proven successful, in this paper, the goal is to beat the small cap index , therefore a rethinking of this formula must be done in order to address a completely different sector which has proven to be more profitable in the long run than the large cap. The steps for the stock ranking or small cap magic formula are the following:

1. Separate the stocks by their clusters
2. Rank the stocks from lower to higher market cap. Then assign each stock a number which will be their ordinal position in the rank, for example the 3rd smallest company will be assign the number 3
3. Rank the stocks from highest to lowest Beta and repeat the same scoring process.
4. Rank the stocks from lowest to highest price to book value and repeat the same scoring process.
5. Rank the stocks from lowest to highest earnings per share and repeat the same scoring process.
6. Rank the stocks from highest to lowest free cash flow to equity and repeat the same scoring process.
7. Rank the stocks from highest to lowest return on equity and repeat the same scoring process.
8. Once each stock has been awarded a grade, sum up all the scores and select the stocks with the lowest overall added score

This formula must be repeated separately for each individual cluster of stocks in order to achieve the diversification of the portfolio. Now, the next step in the portfolio construction process will be the selection of the appropriate number of stocks. Diversification is a tool used by portfolio managers to hedge against the risk of downturns of certain companies, by balancing the loses of a certain stock with the gains achieved by another successful stock inside the portfolio. However, selecting the correct number of stocks to diversify the total portfolio risk is essential, figure 24 shows the effect on the variations of the standard deviation of the whole portfolio as stocks are added to it. Analysing it carefully the obvious conclusion to drag would be the greater

the number of stocks the greater the diversification thus , the lower the portfolio risk. However there is a point where the marginal contribution of each stock towards reducing the overall portfolio volatility starts to diminish, this means that there is a point at which adding an extra stock will not contribute as adding the stock previously creating inefficiencies as the greater the number of stocks the greater the costs for trading and holding them. Taking a look back at figure 24, the optimum point where the portfolio risk will be reduced and have an appropriate number of stocks in the portfolio will be around 30, This number is perfect for the new constructed portfolio as it will mean that the portfolio will have the top 10 rated stocks of each cluster.

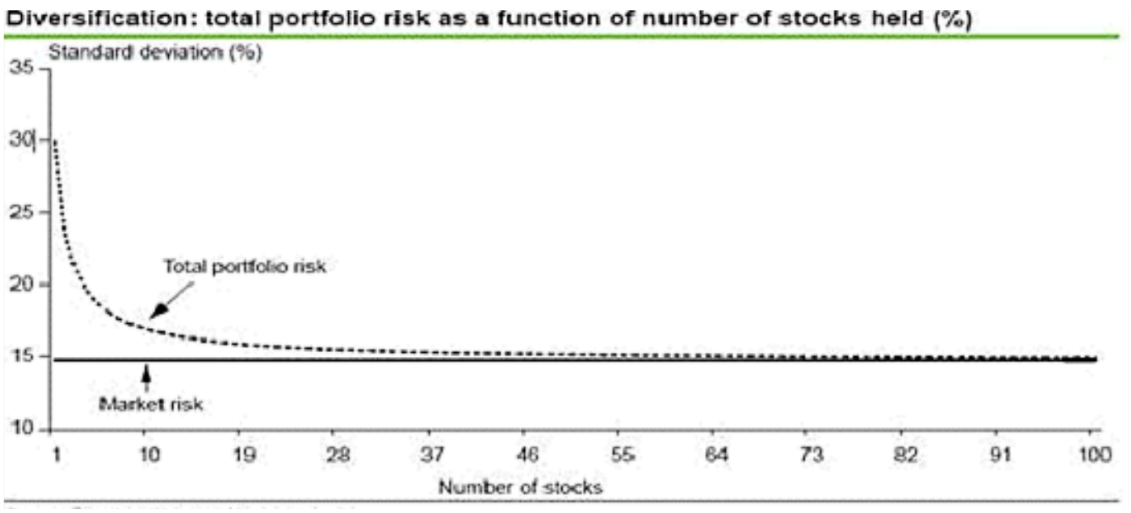


Figure 24 Diversifying Portfolio Risk

Source: Kleinwort Hambros Capital Markets

4.4 Stock Weighting Process

Once the 30 stocks are chosen from the formula applied to every clusters, it is essential to know the importance each stock will have in the portfolio which will beat the market, to do so in this study Harry Markowitz´s efficient frontier theory which is based on setting the optimal portfolios that maximise the future returns for a certain level of volatility or risk undertaken by the specific portfolio. This is a graphical tool which

plots the expected returns of the portfolios in the y axis against the volatility of the portfolio. The expected returns of all the possible returns are exclusively dependant on the weight each asset has on the set of portfolios whereas the volatility of the portfolio will depend only in the covariances of the composing assets in every single portfolio. Therefore, investors will seek to invest in portfolios which lie on the efficient frontier where the risk undertaken is maximised by achieving the highest excepted return possible.

However, the main purpose of the paper is to find a portfolio which beats the market therefore, in order to achieve this goal, it is essential to analyse the maximum Sharpe ratio portfolio. The efficient frontier gives investors the opportunity to easily locate the portfolio by finding the portfolio lying in the efficient frontier which is tangent to the Capital Market Line (CML), this line is derived from the Capital Asset Pricing Model (CAPM) that depicts the additional returns that the portfolio obtains above the risk free rate for every level of risk. The portfolios in this line are optimised, or in other words maximise Sharpe Ratio therefore the tangent point will coincide with the portfolio with the highest possible shape ratio for the 30 chosen stocks.

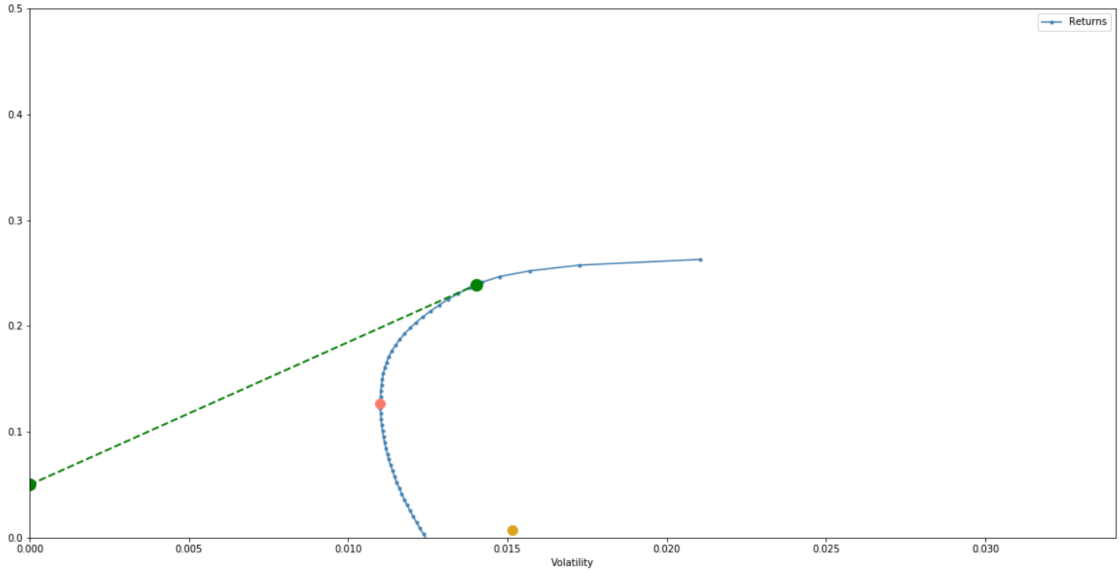


Figure 25 Efficient Frontier

Figure 25, shows 100 possible portfolios in the efficient frontier from the 30 stocks selected by the formula used, The yellow dot shows the performance of the equally weighted portfolio, as you can see equally weighting the stocks doesn´t add value to a portfolio, or what is the same there is barely any alpha generated by the portfolio manager. The pink portfolio which actually sits in the efficient frontier is known as the global minimum variance portfolio, this is known as the combination which offers investors to hold a portfolio with the minimum risk , or what is the same the least volatility and still being an efficient portfolio. Due to the main goal of the new portfolio is beating the market, the best possible portfolio lying in the efficient frontier will be the prementioned maximum Sharpe ratio portfolio, as seen in figure 25 is tangent to the CML line. This is the portfolio which will be studied in order to beat the market.

5. Portfolio Performance Analysis

In order to assess the usefulness and effectivity of the investing guidelines developed in the paper, it is essential to take an insight on the performance of the portfolio and deeply analyse its behaviour in every market situation and always keeping clear that the benchmark is the market index.

The first thing to analyse will be the weighting of each stock in the maximum Sharpe ratio portfolio, which is shown in figure 26. It is evident that the portfolio is overweighed towards 3 mains assets and the rest of the portfolio uses the remaining stocks as a diversification tool for hedge in case there are main drops int the market, however the strong fundamental figures and the powerful past performance is why the three stocks are dominant in the portfolio, this is evident in figure 27 where as it is evident, the returns of the stocks where more or less concentrated but there are some which excel and produce outstanding results and it is why they have such an important stake.

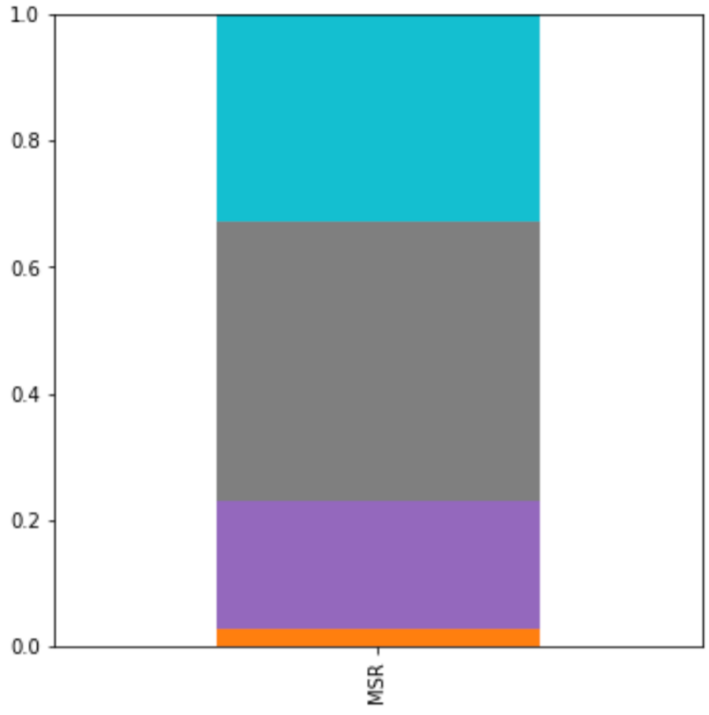


Figure 26 MSR Stock Weights

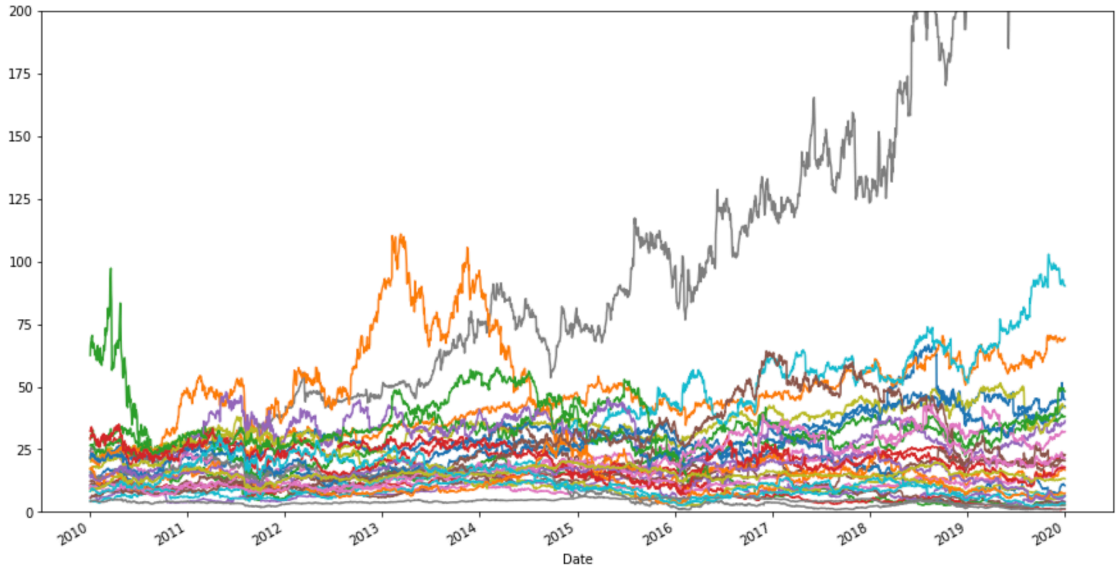


Figure 27 MSR Stock Performance

Source: FactSet

Before getting into analysing the returns, it is essential to highlight a second factor which is extremely important about this portfolio and a key factor in order to achieve the main goal of beating the market index., which is the volatility of the portfolio. In order to reduce the volatility of the portfolio it is important to acknowledge the fact that volatility is exclusively dependant on the correlations of the stocks within the portfolio, therefore figure 28, shows the correlation amid the components of the new portfolio. It is true that heatmaps are extremely difficult to interpret as seen previously in this paper, nevertheless, they key aspect to highlight is the fact that there are few stocks which performances are highly related to each other, or in other words, most of the cost's performances are not correlated . This proves the efficiency of the clustering developed by using the K-mean algorithm , in addition, it also shows that the volatility of the whole portfolio will be reduced drastically.

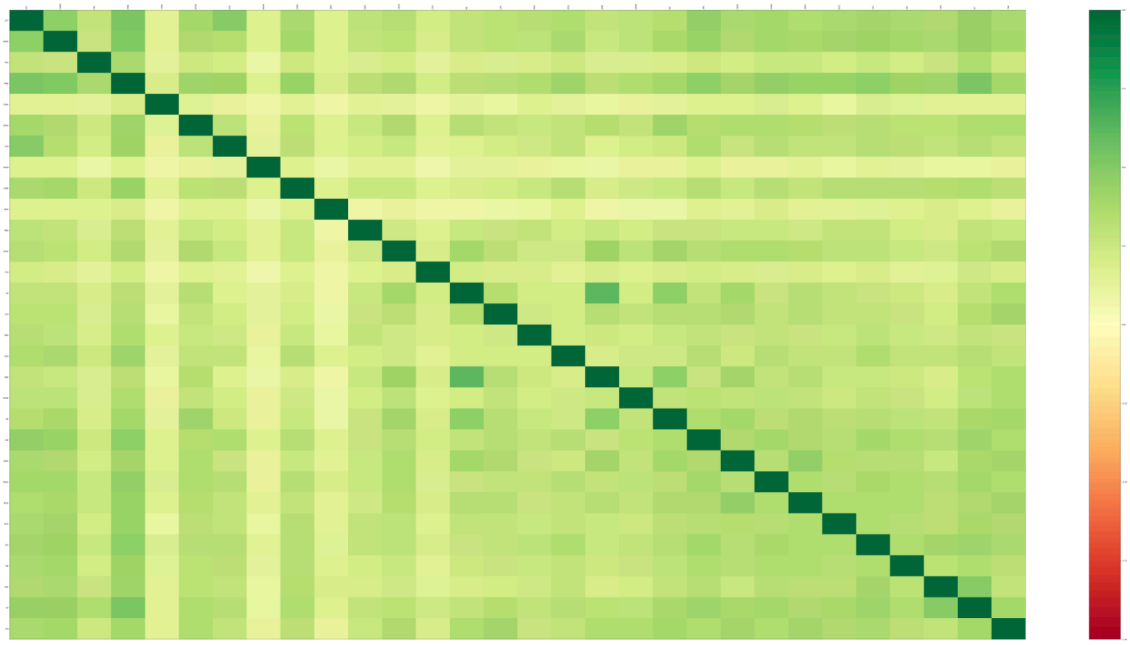


Figure 28 Heat Map New Portfolio

Source: FactSet

The next and most essential test to check the well-functioning of the portfolio construction methodology is analysing the returns and figure 29 is self-explanatory. This graph represents two hypothetical cases, the orange line represents what would have happened to an investor if $1,000 would have been invested in the S&P 600 SC index, on the other hand the blue line represents what would have happened if all of the money had been invested on the maximum Sharpe ratio portfolio. Graphically , the gap is evident that the maximum Sharpe ratio portfolio created using the steps and the guidelines provided by this paper works not only to beat the market but also to achieve outstanding results. In numbers, it can be even more evident, the investor in the index would´ve obtained $3,006.65 for investing its money in the index approximately a 300% of returns in 10 years, which actually it is quite an impressive figure. However, the investor who chose the maximum Sharpe ratio would´ve obtained a total absolute return of $12,286.35 , returns over 1,220% which is approximately 28.5% returns annually, this are extremely outstanding figures.

Figure 29 MSR vs. Index

Despite the yields obtained from the portfolio are more than outstanding it is essential to analyse the statistical figures in order to be able to know whether the formula has worked properly as it would be meaningless to achieve higher returns just by increasing the risk of the portfolio, as in bear markets , like the COVID-19 market crash which recently happened, having a way too risky portfolio could mean a destruction on all of the value generated in the previous years. Figure 30 shows the same statistical results as previously done with SCs , LCs and the benchmark index.

Skewness	Kurtosis	Cornish-Fisher VaR (5%)	Max Drawdown	Sharpe Ratio
0.154438	8.264174	0.019837	-0.230005	0.360275

Figure 30 Statistical Analysis MSR

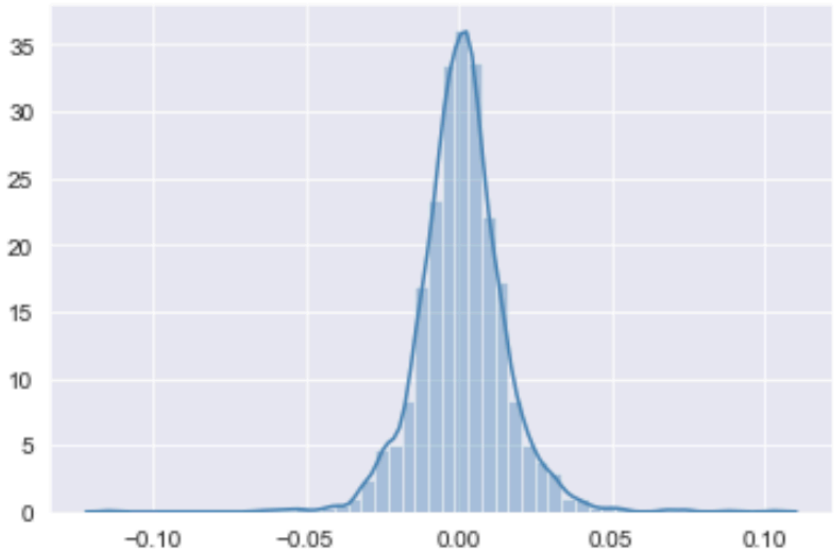


Figure 31 Portfolio Returns Dispersion

The first thing to comment on regarding the statistical results is the skewness, recalling from the first section in the paper the skewness shows the dispersion of the returns and how the tails of the results are shaped, a negative skew, like in the index means that results tend to be negative more times than the times they are positive. Therefore , despite very little, but the fact that kurtosis is positive shows that results tend to be positive more times than negative showing a clear improvement in the portfolio management of the maximum Sharpe ratio portfolio.

The next statistical figure to highlight is the kurtosis, stated how differed from normality where the returns of the portfolio, despite having a leptokurtic kurtosis (3<), it is much lower. This means that the returns are less dispersed than the returns of the portfolio. This might be insignificant, but it really shows that the investor who choses the new created portfolio will not only achieve better results, but the results will be consistent and therefore int will have less fluctuations than investing on the index.

Another improvement done by the stock picking formula has been the decrease in the maximum drawdown. The drawdown from the index was close to 60% which is a substantial drop in the value of the portfolio, nevertheless, thanks to the formula, the maximum drawdown the portfolio would´ve experienced in the last 10 years would´ve been of 23%, more than 50% less of the maximum fall. However, it is essential to

comment on the obvious factor is that the great market crash of the global financial crisis happened in 2008 and the hypothetical portfolio starts on 2010.

Finally, it is essential to analyse, which is considered as one of the key aspects in portfolio management which is the Sharpe ratio. It is true that the yields obtained are substantially greater than average, nevertheless, are they due to a proper management of the stocks or due to the high risk undertaken by the portfolio? Sharpe ratio is the indicator of this aspect, the result was 0.36, this is quite a low figure compared to other portfolio managers in the industry, nevertheless, it is essential to state that the index had a -0.53 Sharpe ratio. This is an extremely substantial increase in the way risk is managed in the portfolio, as it assures that investors are correctly rewarded for each unit of risk they undertake in the portfolio.

6. Conclusions

After cautiously detailing every step in the portfolio construction and after having analysed the investment universe of the SCs companies, several conclusions can be dragged from this paper.

First of all , after comparing the results of the index with the performance of the portfolio created by the guidelines in this paper, it is evident that the methodology works, and it is able to beat the market. Not only by looking at the spread between the returns of the portfolio and the index which notably surpasses it, but also by taking a look at the Sharpe ratio, which tells us that not only the fund is a better investment opportunity but also because it is more efficient hedging the risk of the portfolio. The improvement in the Sharpe ratio is extremely notable and reinforces the fact that investors are properly rewarded by the risk-taking investing in SCs which have quite a volatile nature. Therefore, it is simple and easy to state that the developed portfolio construction methodology is able to beat its benchmarking index.

Despite there is clear and simple evidence to support the formula created there are also certain issues which could be addressed. The first and most important is the opportunity to expand the fundamental analysis on the companies, the use of the Fama & French 3 factor model is extremely useful to analyse where the investment opportunities can arise and know which fundamentals can be used to rank the companies, nevertheless, there have been improvements to the model and there is 5 factor model which can be used to be able to identify more fundamentals from the company and add more value to the portfolio selection and stock ranking process. Beyond this, another factor which could be debated is the fact that in order to achieve higher returns, analysts should spend more time looking at intangibles and looking at the business model of the company in order to estimate real potential growth of the company and thus of the stock price, however, it is essential to recall one of the main objectives of the paper which was to use data analytics such as the k-mean algorithm to help portfolio managers beat the market and create a fund which can manage low management fees whilst it still beats the ETF which mirrors the particular index.

Finally, the last issue to comment is the efficient frontier model which allows to create the weights of the assets in the portfolio. The theory, inside Markowitz's modern portfolio theories is world-wide known and has been proven successful, nevertheless, it basis its analysis on past data and it is key to highlight that despite past performance is always a good indicator, successful past performance doesn´t necessarily mean that in the future that trend is going to continue. Having this stated, the use of fundamentals in the previous selection process tries to compensate the bias the efficient frontier model has on looking exclusively at the past performances. Again, making reference to the outstanding performance of the portfolio in the last years, it is essential to state the fact that the selection methodology actually works.

Bibliography

BGF Asset Management. (2020). *Asset class returns.* BGF Asset Management.

Bierig, R. F. (2000). *THE EVOLUTION OF THE IDEA OF "VALUE INVESTING": FROM BENJAMIN GRAHAM TO WARREN BUFFETT.* Duke University.

BoA Merryl Lynch. (2019). *Research Equity Portfolios.* BoA Merryl Lynch.

Cheng, P. L. (2009). *Efficient Portfolio Selections beyond the Markowitz Frontier.* Cambridge University.

Dwyer, A. (2014). *An Application of Portfolio Theory and the Efficient Frontier Concept to the Risk-Return Decisions of Beothuk Hunter-Fisher-Gatherers.* Mount Royal University.

ETFGI. (2016).

Fisher, P. A. (1960). *Common Stocks and Uncommon Profits.*

Flood, C. (2019). Popularity of passive investing changes rules of the game. *Financial Times*.

French, E. F. (1992). *The Cross-Section of Expected Stock Returns.* The Journal of Finance.

French, E. F. (1993). *Common risk factors in the returns on stocks and bonds**. Journal of Financial Economics.

Goumas, A. (2010). *Value Investing and The Magic Formula - a method for successful stock investments.*

Graham, B. (1949). *The Intelligent Investor.* New York.

Greenblatt, J. (1996). *The Little Book that Beats the Market.* New York.

Greenblatt, J. (1997). *You Can Be a Stock Market Genius.* New York.

Guthrie, J. (2020). The fallacy behind the rise of passive fund management. *Financial Times*.

Henderson, R. (2019). US small-caps rebound as equity rally broade. *Financial Times*.

Hilpisch, Y. (2017). *Python for Finance.* O´Reilly.

Hulbert, M. (n.d.). *The Big Reason Behind Small-Caps' Struggle.* New York: Wall Street Journal.

Iyiola Omisore, M. Y. (2012). *The modern portfolio theory as an investment decision tool.* Journal of Accounting and Taxation.

Kawa, L. (2019). Small-Cap Stocks Match a Record Streak of Big Advances. *Bloomberg*.

Kenneth Fisher, P. L. (1994). *The Warren Buffett Way.*

Kusrini, K. (2018). *Grouping of Retail Items by Using K-Means Clustering.* Procedia Computer Science.

Langley, K. (n.d.). *Small-Caps Outperform in Strong Period for Stocks.* Wall Street Journal.

Mangram, M. (2013). *A Simplified Perspective of the Markowitz Portfolio Theory.* Colorado: Colorado Technical University.

Markowitz, H. M. (1976). *Markowitz Revisited.* Financial Analysts Journal.

Merrill Lynch Bank of America. (2020, May). *https://www.merrilledge.com/article/growth-vs-value-investing-two-approaches-to-stocks#:~:text=value%3A%20two%20approaches%20to%20stock%20investing&text=Growth%20and%20value%20are%20two,and%20stock%20mutual%20fund%20investing.&text=Growth%20investors%20se.*

Mills, D. Q. (2001). *Who's to Blame for the Bubble?* Boston: Harvard Business School.

Momeni, M. (2015). *Clustering Stock Market Companies via K-Means Algorithm.* Kuwait Business School.

Nanda, S. (2016). *Clustering Indian stock market data for portfolio management.* Science Direct.

Ning, V. (2018). *International Small Cap Investing.* New York: S&P Global.

Otani, A. (2020). *America's Smallest Stocks Are Staging a Comeback.*

Pai, G. A. (2018). *Python for Portfolio Optimization.*

Phelim Boyle, L. G. (2011). *Keynes Meets Markowitz: The Trade-Off Between Familiarity and Diversification.* Management Science.

Rajablu, M. (2015). *Value investing: review of Warren Buffett's investment philosophy and practice.* Columbia Business School.

S&P Dow Jones Index . (2020). *S&P 600 Factsheet.* New York: S&P Global .

S&P Dow Jones Index. (2020). *S&P 600 Factsheet.* New York: S&P Global.

S&P Dow Jones Indices. (2020). *S&P 600 Index Factsheet.* New York: S&P Global.

Snow, D. (2019). *Machine Learning in Asset Management.* NYU.

William N. Goetzmann, A. K. (2008). *Equity Portfolio Diversification.* Review of Finance.

Appendix

```python
import pandas as pd
import numpy as np
import math
from scipy.stats import norm

def annualize_rets(r, periods_per_year):
    """
    Annualizes a set of returns
    We should infer the periods per year
    but that is currently left as an exercise
    to the reader :-)
    """
    compounded_growth = (1+r).prod()
    n_periods = r.shape[0]
    return compounded_growth**(periods_per_year/n_periods)-1


def annualize_vol(r, periods_per_year):
    """
    Annualizes the vol of a set of returns
    We should infer the periods per year
    but that is currently left as an exercise
    to the reader :-)
    """
    return r.std()*(periods_per_year**0.5)
```

```
def sharpe_ratio(r, riskfree_rate, periods_per_year):
    """
    Computes the annualized sharpe ratio of a set of returns
    """
    # convert the annual riskfree rate to per period
    rf_per_period = (1+riskfree_rate)**(1/periods_per_year)-
    excess_ret = r - rf_per_period
    ann_ex_ret = annualize_rets(excess_ret, periods_per_year
    ann_vol = annualize_vol(r, periods_per_year)
    return ann_ex_ret/ann_vol
def drawdown(return_series: pd.Series):
    """Takes a time series of asset returns.
       returns a DataFrame with columns for
       the wealth index,
       the previous peaks, and
       the percentage drawdown
    """
    wealth_index = 1000*(1+return_series).cumprod()
    previous_peaks = wealth_index.cummax()
    drawdowns = (wealth_index - previous_peaks)/previous_peal
    return pd.DataFrame({"Wealth": wealth_index,
                         "Previous Peak": previous_peaks,
                         "Drawdown": drawdowns})
```

```
def sharpe_ratio(r, riskfree_rate, periods_per_year):
    """
    Computes the annualized sharpe ratio of a set of returns
    """
    # convert the annual riskfree rate to per period
    rf_per_period = (1+riskfree_rate)**(1/periods_per_year)-1
    excess_ret = r - rf_per_period
    ann_ex_ret = annualize_rets(excess_ret, periods_per_year)
    ann_vol = annualize_vol(r, periods_per_year)
    return ann_ex_ret/ann_vol
def drawdown(return_series: pd.Series):
    """Takes a time series of asset returns.
       returns a DataFrame with columns for
       the wealth index,
       the previous peaks, and
       the percentage drawdown
    """
    wealth_index = 1000*(1+return_series).cumprod()
    previous_peaks = wealth_index.cummax()
    drawdowns = (wealth_index - previous_peaks)/previous_peaks
    return pd.DataFrame({"Wealth": wealth_index,
                         "Previous Peak": previous_peaks,
                         "Drawdown": drawdowns})
```

```
def kurtosis(r):
    """
    Alternative to scipy.stats.kurtosis()
    Computes the kurtosis of the supplied Series or DataFrame
    Returns a float or a Series
    """
    demeaned_r = r - r.mean()
    # use the population standard deviation, so set dof=0
    sigma_r = r.std(ddof=0)
    exp = (demeaned_r**4).mean()
    return exp/sigma_r**4

def skewness(r):
    """
    Alternative to scipy.stats.skew()
    Computes the skewness of the supplied Series or DataFrame
    Returns a float or a Series
    """
    demeaned_r = r - r.mean()
    # use the population standard deviation, so set dof=0
    sigma_r = r.std(ddof=0)
    exp = (demeaned_r**3).mean()
    return exp/sigma_r**3
```

```python
def var_historic(r, level=5):
    """
    Returns the historic Value at Risk at a specified level
    i.e. returns the number such that "level" percent of the returns
    fall below that number, and the (100-level) percent are above
    """
    if isinstance(r, pd.DataFrame):
        return r.aggregate(var_historic, level=level)
    elif isinstance(r, pd.Series):
        return -np.percentile(r, level)
    else:
        raise TypeError("Expected r to be a Series or DataFrame")

def var_gaussian(r, level=5, modified=False):
    """
    Returns the Parametric Gauusian VaR of a Series or DataFrame
    If "modified" is True, then the modified VaR is returned,
    using the Cornish-Fisher modification
    """
    # compute the Z score assuming it was Gaussian
    z = norm.ppf(level/100)
    if modified:
        # modify the Z score based on observed skewness and kurtosis
        s = skewness(r)
        k = kurtosis(r)
        z = (z +
                (z**2 - 1)*s/6 +
                (z**3 -3*z)*(k-3)/24 -
                (2*z**3 - 5*z)*(s**2)/36
            )
    return -(r.mean() + z*r.std(ddof=0))


def summary_stats(r, riskfree_rate=0.03):
    """
    Return a DataFrame that contains aggregated summary stats for the returns in the
columns of r
    """
    ann_r = r.aggregate(annualize_rets, periods_per_year=12)
    ann_vol = r.aggregate(annualize_vol, periods_per_year=12)
    ann_sr = r.aggregate(sharpe_ratio, riskfree_rate=riskfree_rate, periods_per_year=12)
    dd = r.aggregate(lambda r: drawdown(r).Drawdown.min())
    skew = r.aggregate(skewness)
    kurt = r.aggregate(kurtosis)
    cf_var5 = r.aggregate(var_gaussian, modified=True)
    hist_cvar5 = r.aggregate(cvar_historic)
    return pd.DataFrame({
        "Annualized Return": ann_r,
        "Annualized Vol": ann_vol,
        "Skewness": skew,
        "Kurtosis": kurt,
        "Cornish-Fisher VaR (5%)": cf_var5,
        "Max Drawdown": dd,
        "Sharpe Ratio": ann_sr,
    })
```

```
def summary_rets(r, riskfree_rate=0.03):
    """
    Return a DataFrame that contains aggregated summary stats for the returns in the
columns of r
    """
    ann_r = r.aggregate(annualize_rets, periods_per_year=12)
    skew = r.aggregate(skewness)
    kurt = r.aggregate(kurtosis)
    avg = r.mean()
    median= r.median()
    return pd.DataFrame({
        "Annualized Return": ann_r,
        "Skewness": skew,
        "Kurtosis": kurt,
        "Mean": avg,
        "Median" : median

    })


def summary_vol(r, periods_per_year):

    volatility = r.std()*(periods_per_year**0.5)

    return pd.DataFrame({
        "Annualized Volatility" : volatility
    })
```

```
def optimal_weights(n_points, er, cov):
    target_rs = np.linspace(er.min(),er.max(),n_points)
    weights = [minimize_vol(target_return,er,cov) for target_return in target_rs]
    return weights

def minimize_vol(target_return,er,cov):

    n = er.shape[0]
    init_guess =np.repeat(1/n,n)
    bounds = ((0.0,1.0),)*n
    return_is_target =  {
        "type" : "eq",
        "args" : (er,),
        "fun"  : lambda weights, er:target_return - portfolio_return(weights,er)
    }

    weights_sum_to_1 = {
        "type":"eq",
        "fun" : lambda weights:np.sum(weights)-1
    }

    results = minimize(portfolio_vol, init_guess,
                       args=(cov,), method="SLSQP",
                       options={"disp":False},
                       constraints= (return_is_target,weights_sum_to_1),
                       bounds=bounds
                      )

    return results.x
```

```python
def msr(riskfree_rate, er, cov):
    n = er.shape[0]
    init_guess = np.repeat(1/n,n)
    bounds = ((0.0,1.0),)*n
    weights_sum_to_1 = {
        'type':'eq',
        'fun' : lambda weights: np.sum(weights)-1
    }
    def neg_sharpe_ratio(weights,riskfree_rate,er,cov):
        r=portfolio_return(weights,er)
        vol=portfolio_vol(weights,cov)
        return  -(r - riskfree_rate)/vol

    results = minimize(neg_sharpe_ratio, init_guess,
                       args=(riskfree_rate,er,cov,), method="SLSQP",
                       options={"disp":False},
                       constraints= (weights_sum_to_1),
                       bounds = bounds
                      )

    return results.x

def gmv(cov):
    n = cov.shape[0]

    return msr(0,np.repeat(1,n),cov)
```

```
def plot_ef(n_points, er, cov,show_cml=False,style=".-",riskfree_rate=0, show_ew=Fa
    weights = optimal_weights(n_points, er, cov)
    rets = [portfolio_return(w,er) for w in weights]
    vols = [portfolio_vol(w, cov) for w in weights]
    ef = pd.DataFrame ({
        "Returns": rets,
        "Volatility": vols,
    })

    ax = ef.plot.line(x="Volatility", y="Returns",style=style, color="steelblue",fi
    if show_ew:
        n = er.shape[0]
        w_ew =np.repeat(1/n,n)
        r_ew= portfolio_return(w_ew, er)
        vol_ew =portfolio_vol(w_ew,cov)
        ax.plot([vol_ew],[r_ew], color="goldenrod", marker="o", markersize=10)
    if show_gmv:
        w_gmv=gmv(cov)
        r_gmv= portfolio_return(w_gmv, er)
        vol_gmv =portfolio_vol(w_gmv,cov)
        ax.plot([vol_gmv],[r_gmv], color="salmon", marker="o", markersize=10)
    if show_cml:
        ax.set_xlim(left = 0)
        plt.ylim(0,200)
        w_msr=msr(riskfree_rate, er, cov)
        r_msr=portfolio_return(w_msr,er)
        vol_msr= portfolio_vol(w_msr,cov)
        cml_x = [0,vol_msr]
        cml_y = [riskfree_rate, r_msr]
        ax.plot(cml_x, cml_y,color="green",marker="o",linestyle="dashed",markersize
        plt.ylim(0,0.5)

    return ax
```

```
def plot_ef(n_points, er, cov,show_cml=False,style=".-",riskfree_rate=0, show_ew=False,show_gmv=False):
    weights = optimal_weights(n_points, er, cov)
    rets = [portfolio_return(w,er) for w in weights]
    vols = [portfolio_vol(w, cov) for w in weights]
    ef = pd.DataFrame ({
        "Returns": rets,
        "Volatility": vols,
    })

    ax = ef.plot.line(x="Volatility", y="Returns",style=style, color="steelblue",figsize=(20,10), legend=True)
    if show_ew:
        n = er.shape[0]
        w_ew =np.repeat(1/n,n)
        r_ew= portfolio_return(w_ew, er)
        vol_ew =portfolio_vol(w_ew,cov)
        ax.plot([vol_ew],[r_ew], color="goldenrod", marker="o", markersize=10)
    if show_gmv:
        w_gmv=gmv(cov)
        r_gmv= portfolio_return(w_gmv, er)
        vol_gmv =portfolio_vol(w_gmv,cov)
        ax.plot([vol_gmv],[r_gmv], color="salmon", marker="o", markersize=10)
    if show_cml:
        ax.set_xlim(left = 0)
        plt.ylim(0,200)
        w_msr=msr(riskfree_rate, er, cov)
        r_msr=portfolio_return(w_msr,er)
        vol_msr= portfolio_vol(w_msr,cov)
        cml_x = [0,vol_msr]
        cml_y = [riskfree_rate, r_msr]
        ax.plot(cml_x, cml_y,color="green",marker="o",linestyle="dashed",markersize=12,linewidth=2)
        plt.ylim(0,0.5)

    return ax
```

```
price_sml= pd.read_excel("SML Monthly Data.xlsx", parse_dates=True,index_col=0) #load the performance of the inde
price_sml.index = pd.to_datetime(price_sml.index,format = "%Y%m%d" )# change to date format
price_sml = price_sml["Price"] #Name the column Pruce
price_sml= price_sml[:"2018"] # Filter data til the end of 2018
price_sml
```

```
def compound(r):

    return np.expm1(np.log1p(r).sum())

# function to load the data of the fama and french directly using the monthly performance
def get_fff_returns():
    rets = pd.read_csv("data/F-F_Research_Data_Factors_m.csv",
                       header=0, index_col=0, na_values=-99.99)/100
    rets.index = pd.to_datetime(rets.index, format="%Y%m").to_period('M')
    return rets
```

```
rets_m = rets.resample('M').apply(compound).to_period('M') #change dayli to monthly data
rets_m = pd.DataFrame(rets_m)
rets_m
```

```
fff = get_fff_returns()
fff.head()
fff= fff["1996":"2019"]
fff
#filter the Fama and French DF
```

```
rets_excess = rets_m - fff.loc["1996":"2018", ['RF']].values
mkt_excess = fff.loc["1996":"2018",['Mkt-RF']]
exp_var = mkt_excess.copy()
exp_var["Constant"] = 1
lm = sm.OLS(rets_excess, exp_var).fit()
```

```
lm = sm.OLS(rets_excess, exp_var).fit()
lm.summary() #now the model is exaplained by 94.3%
```

```
kmeans_pca = KMeans(n_clusters = 3, init="k-means++", random_state=42)
# using 3 clusters using elbow criterion
```

```
kmeans_pca.fit(scores_pca)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=3, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=42, tol=0.0001, verbose=0)
```

```
df_segm_pca_kmeans = pd.concat([df.reset_index(drop = True), pd.DataFrame(scores_pca)], axis = 1)
df_segm_pca_kmeans.columns.values[-3:]=["C1","C2","C3"]
df_segm_pca_kmeans["Segment K-means PCA"] = kmeans_pca.labels_
#add the column of groups at the end of the PCA with the new components 1-96
```

```
df_segm_pca_kmeans
```

```
df_segm_pca_kmeans["Segment"]=df_segm_pca_kmeans["Segment K-means PCA"].map({0:"1st",
                                                                             1:"2nd",
                                                                             2:"3rd",
                                                                            })
# prepare each segment and map it
```

```
x_axis = df_segm_pca_kmeans["C1"]
y_axis = df_segm_pca_kmeans["C2"]
plt.figure(figsize=(10,8))
sns.scatterplot(x_axis,y_axis, hue = df_segm_pca_kmeans["Segment"], palette = "Spectral")
plt.title("Cluster by PCA")
plt.show()

#plot the 3 different groups
```