American Journal of Business and Society
Vol. 5, No. 4, 2020, pp. 207-216
http://www.aiscience.org/journal/ajbs

The Selection of Top Stocks Using a Statistical Approach

Carol Anne Hargreaves*, Hu Yiming, Jariah Bte Abdul Nassar

Department of Statistics and Applied Probability, Faculty of Science, National University of Singapore, Singapore

Abstract

Introduction: Investors continuously seek for a "strategic secret" to identify potentially profitable stock portfolios. Objective: The objective of this study is to systematically identify top performing stocks in different sectors of the Australian Stock Exchange, using three simple statistical techniques: Pearson Correlation, trend analysis and Principal Component Analysis (PCA). Methodology: Investors want to buy stocks that can give them good returns. We apply the Pearson Correlation technique to identify stocks whose prices are positively correlated with time, stocks which have a positive, upward trend in the most recent weeks. Secondly, investors also want to know why they should buy a particular stock and are interested in knowing the important factors only, while trying to keep their decision making as simple as possible. Principal Component Analysis is a statistical technique that reduces many predictor variables to a few factors. Results: Our results demonstrated that the Pearson Correlation analysis and the Principal Component Analysis, coupled with a simple short-term trading strategy was reliable in identifying winning stock portfolios in the Australian stock market. Our stock portfolios consistently reaped profits across consecutive time periods for different sectors. In addition, our stock portfolios in all three trading periods outperformed the Australian Stock Market Index. Our stock portfolios delivered return on investments at least 3.1 times higher than the stock market index over the three one-month trading periods. Conclusion: The objective of this study was to examine whether our stock portfolios would outperform the stock market index over three consecutive time periods. Our results demonstrated that our methodology was reliable and consistent across all 3 time periods, delivering significant profits from trading, further proving that our method was not only theoretically robust but also practically sound.

Keywords

Stock Analysis, Correlation Analysis, Principal Component Analysis, PCA, Top Stocks, Australian Stock Market, Trading Strategy, Stock Portfolio

Received: October 16, 2020 / Accepted: November 15, 2020 / Published online: December 11, 2020

1. Introduction

Identifying top performing stocks has been of interest for both investors and researchers for more than a decade because of the potential gain of significant profit. Many research papers have predicted the pricing of the stock index as well as stock performance across many European markets (e.g., UK, France, and Germany), are predictable [10]. Further, some researchers have combined the usage of both fundamental and technical variables for the prediction of profitable stocks using the support vector machine learning algorithm [20]. They systematically identified high returning healthcare stocks and traded them with the help of an automatic trading application without human error and sentiment interference and yielded 16.64% revenue at the end of a three-month trading period.

Further, other researchers investigated whether a healthcare sector stock portfolio will outperform the ASX All-Ordinaries Index (AORD) and Healthcare Sector Index (AXHJ) over the twenty days trading period using the logistic regression model,

* Corresponding author
E-mail address: carol.hargreaves@nus.edu.sg (C. A. Hargreaves)

a more statistical approach [12]. The healthcare stock portfolio returned 18.24%. In this paper, we would like to determine whether the selection of stocks in the Australian Stock Market using the correlation approach and Principal Component Analysis (PCA) will result in identifying top performing stocks that outperform the Australian stock market index.

We aim to maximise the potential return on investment by trading stocks in the Australian stock market on a short-term basis (one-month period). As stock markets are known to be volatile, the top stock(s) are likely to change over time [30]. Also, different sectors perform well at different time periods. Hence, in this paper, we develop a strategy to identify the top stock(s) for the top performing sector in each of the three consecutive time periods. A total of three stock portfolios were constructed and the performances of these portfolios were compared against the Australian stock market index.

Several researchers in the field of portfolio analysis and statistics had suggested and evaluated the use of correlation in investment portfolio analysis. Markowitz was the first to formally develop a mathematical model in the study and analysis of portfolios selection. He introduced the Modern Portfolio Theory based on mean-variance analysis for efficient diversification of investments, which involves the use of correlation coefficient between two stocks [22]. Moore introduced correlation analysis to estimate the characteristics of stock behaviour to study the successive movements in the price of common stocks [24].

In investment trading, stock price trends can provide arbitrage opportunities [31]. Trends can be defined as the direction in which the market is moving. Researchers are interested in identifying the future trend of the stock price by applying statistical techniques on past data [11]. A stock portfolio using the data mining approach was performed using the Australian Stock Market, where results demonstrated successfully, that data mining techniques can model the trend of stock prices which are nonlinear [12].

Some investors used economic variables such as interest rates and expected inflation to predict stock returns [6,7]. Others used fundamental variables such as return on assets, earnings yield, size and book to market equity in predicting future stock prices [8,15,19]. Principal Component Analysis (PCA), a dimensionality reduction technique, can be used to reduce the number of input variables into fewer factors, allowing investors to interpret the data in a more meaningful way. Pasini performed portfolio optimisation using PCA on three subgroups of stocks in the Down Jones Industrial (DJI) index [27]. Mani used PCA to reduce twenty-two fundamental stock variables to only four variables that were sufficient in identifying winning stocks in the Australian stock market [19], while Narayan showed that excess stock returns can be predicted using either institutional, macroeconomic factors, or a combination of both factors, for fifteen out of eighteen emerging markets studied using PCA [25].

In this study, we first apply the Pearson correlation technique to select the top performing stock sector with an upward trend. Next, we use a combination of PCA and correlation analysis to identify top performing stock(s) in the selected sectors. Up to three top stocks were selected to form an investment portfolio with an equal weightage. With the stock portfolios identified, a short-term trading over a one-month period was conducted. The return for each portfolio was compared against the performance of the Australian stock market index (AORD). We propose that the stock portfolio will outperform the Australian stock market index (AORD). The methodology was applied to three consecutive time periods to rule out possible random results and demonstrate its consistency and reliability.

This paper is structured into 5 sections. While Section 1 is the introduction, Section 2 describes the data extraction and cleaning process, Section 3, a brief overview of the methodology used, Section 4 the analysis results, after which Section 5 presents the conclusion.

2. Data Extraction and Cleaning

We used the open-source R programming software (version 3.5.1) for our study. We collected the Australian stock technical data from by using the *"quantmod"* package in R [32]. We retrieve the daily adjusted close prices by sector, for all stocks, and the market index for three time periods. The three time periods are shown in table 1 below.

Table 1. Time Periods Used for Analysis.

Period	Dates
1	1 February 2019 to 31 March 2019
2	1 April 2019 to 31 May 2019
3	1 June 2019 to 31 July 2019

Each of the three time periods were further split into a one-month training period followed by a one-month trading period as shown in table 2 below. Relatively short training and trading periods were chosen as trends in stock price tend to be short-lived [3].

Table 2. Training and Trading Periods Used for Analysis.

Period	Training Period	Trading Period
1	1 February 2019 to 28 February 2019	1 March 2019 to 31 March 2019
2	1 April 2019 to 30 April 2019	1 May 2019 to 31 May 2019
3	1 June 2019 to 30 June 2019	1 July 2019 to 31 July 2019

We scraped the Australian stock fundamental data from Yahoo Finance using the webscraper.io for the three training periods stated in table 2 [21, 32]. The data was formatted in a structured format for further analysis.

3. Methodology

In our study, we used a three-step process for portfolio construction followed by the simple trading strategy as shown in figure 1 below.

Expected stock market returns varied overtime in a predictable way. Stocks that had statistically significant and high correlation values against time in the current period had potential upward price trend in the immediate period that followed [9]. We constructed a two-layered filter to identify the stocks that were more likely to perform well in the period immediately after the training period, otherwise known was the trading period.

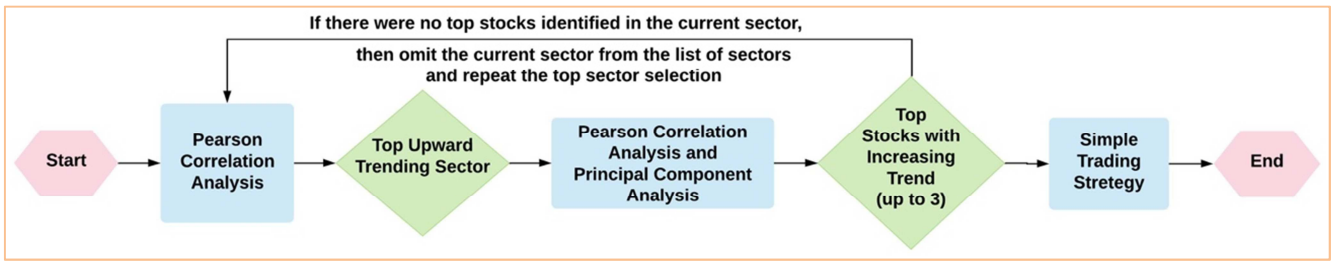

Figure 1. Methodology Flowchart.

3.1. Selection of Top Upward Trending Sector: Pearson Correlation Analysis

In the first stage of filtering, technical data were used. We selected the top sector whose sector index prices had a strong and significant positive correlation with respect to time over the one- month training period.

The correlation was calculated using the Pearson correlation coefficient formula [26].

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \quad (1)$$

The Pearson correlation coefficient indicates the direction and strength of a linear association between two variables. The values range between -1 and 1 [5]. If the correlation coefficient is near -1 or 1, then there is strong linear relationship between the two variables. However, if the correlation coefficient is near zero, then there is a weak linear relationship between the variables. A positive (negative) correlation coefficient indicates a positive (negative) linear relationship – as one variable increases (increases), the other variable increases (decreases) and vice versa [5]. A zero-correlation coefficient indicates that the two variables do not exhibit a linear relationship between the movement of the two variables.

The null hypothesis under the Pearson correlation statistic is that there is no significant linear relationship between the two variables [26]. The probability value (p-value) of the correlation coefficient is an indicator of whether the null hypothesis is rejected. In general, a smaller p-value suggests that there is a stronger evidence in favour of the alternative hypothesis. In our study, we reject the null hypothesis when the corresponding p-value is less than 5% and conclude that there is sufficient evidence to show that the two variables have a linear relationship at 5% level of significance [18].

In figure 2 below, we compared the scatter plot of two market sector indices across the last three years, namely, the healthcare sector (^AXHJ) and the industrial sector (^AXUJ). We can observe that there is indeed a linear trend for ^AXHJ, which has a statistically significant, high and positive correlation coefficient of 0.9399 (p-value of < 2.2E-16). Comparatively, no linear trend can be observed for ^AXUJ, which has a statistically significant, low and negative correlation coefficient of -0.3269 (p-value of 2.6E-20).

In our study, we check the correlation coefficient of all the market sector indices in each time period. As shown in figure 3 below, we only considered those sectors with a positive correlation coefficient between the stock price and time of at least 0.7, at a 5% significance level. The correlation coefficient is calculated using equation (1), where the sector index prices were defined by x and the total number of days in the training period, 1, 2, 3, …, n, defined the y variable. Days when no trading occurred were omitted for conciseness. The y values, which denoted time, increased linearly from 1 to n. Stock market sector indices with a statistically significant, high and positive correlation implied that the stocks in the sector had an overall strong positive linear association between price and time. Thus, we assumed these market sector indices had a linear trend in the time period understudy.

The top sector with the highest statistically significant

correlation above 0.7 was chosen in this step. Only the stocks in the chosen top sector were selected for further analysis in the next section.

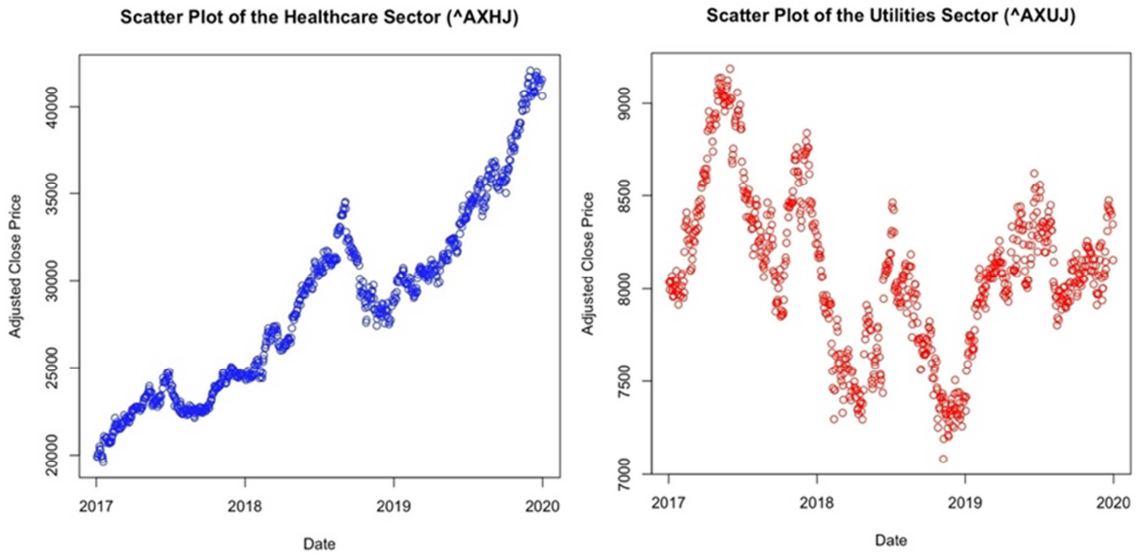

Figure 2. Scatter Plots of ^AXHJ. AX and ^AXUJ. AX.

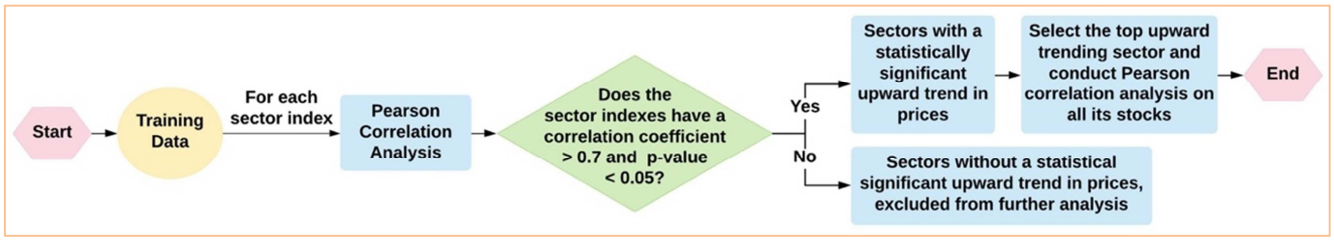

Figure 3. Top Upward Trending Sector Selection Process.

3.2. Selection of the Top Three Upward Trending Stocks

Using the filtered portion of stocks from section 3.1, we next selected a maximum of three top stocks as shown in figure 4 below.

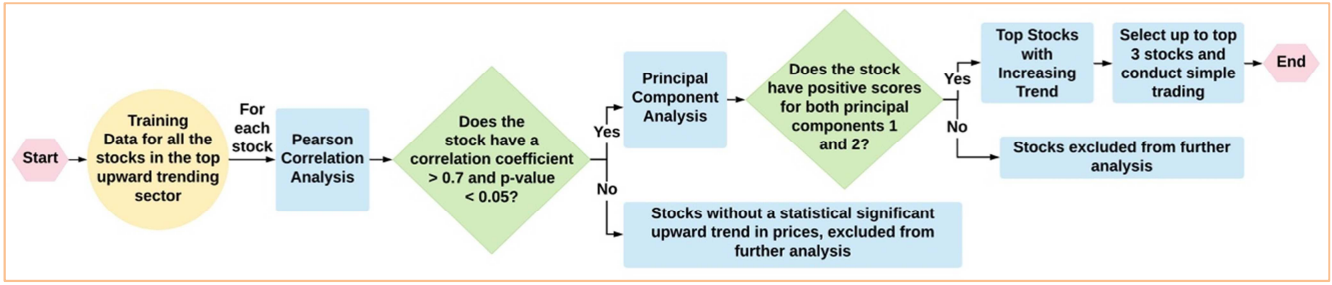

Figure 4. Top Upward Trending Stocks Selection Process.

3.2.1. Pearson Correlation Analysis on Individual Stocks

In this stage of filtering, we selected the stocks whose prices had a strong and significant positive correlation with respect to time based on the technical data.

The correlation coefficients of the stocks in the training period were used to identify if they had upward price trend in the training period. A similar method to that was described in section 3.1 was applied, where, only the stocks with a positive correlation coefficient between the stock prices and time, of at least 0.8 at a 5% level of significance was considered. The correlation coefficient was calculated using equation (1) by substituting the stock prices for the x variable and values 1 to n for the y variable. Stocks with a statistically significant, high and positive correlation implied that it had a strong positive linear association between the stock price and time. Stocks filtered for an upward price

trend in this stage were then further analysed in the next section.

3.2.2. Principal Component Analysis

Australian stock fundamental data have a large number of variables. The use of Principal Component Analysis (PCA) allowed us to interpret it more meaningfully. PCA is a dimension reduction technique that reduces a large number of input variables into fewer number of factors. The resulting factors are linear combination of the input variables and explain most of the variation in the data set. PCA is commonly used to explore financial time series, compute financial risks, and statistical arbitrage [1, 14, 29]. It can be applied to different fields such as feature extraction, statistical variables analysis and visualisation of high dimensional data [13, 16, 19].

PCA was performed using the IBM SPSS Statistics 26 standard software. All thirty-four fundamental variables were used. We examined the scree plots and eigenvalues to estimate the number of principal components that could be extracted. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy test was used to verify the sampling adequacy of the data for the PCA. KMO statistics varies between 0 and 1. High KMO values indicate that a PCA will be valuable. A KMO value of above 0.5 is adequate. On the other hand, the Bartlett's test of sphericity checks the redundancy between variables to see if they can be summarised with a few factors [17]. It tests the null hypothesis that the correlation matrix is an identity matrix. At a 5% significance level, if the p-value of Bartlett's test is less than 0.05, then there is sufficient evidence to reject the null hypothesis, which indicates that the variables are related and PCA should be conducted. Thus, if the KMO value is greater than 0.5 and the p-value of Bartlett's test is less than 0.05, then we may conclude that the sample is adequate [19].

There is a general rule of thumb that the extracted principal components should contribute at least 70% to the total variance. The number of factors to be extracted is approximately the total number of eigenvalues that were greater than one. Varimax rotation method were used to make the final principal components more interpretable. To ensure that the extracted principal components are meaningful with high loadings and explain the maximum variability, variables with low loadings of less than 0.45 were removed after each iteration and the PCA is repeated several times until all of the following are satisfied [19].

1) The variables loading on the principal components have eigenvalues greater than 1.
2) The principal components extracted explains at least 70% of the total variance.
3) The variables loaded in the rotated component matrix is at least 0.7.

We specifically examined the first two principal components; we call them principal component 1 (PC1) and 2 (PC2). The reliability and consistency of the principal components were verified by checking that the Cronbach's alpha coefficient is at least 0.7 [4].

The filtered portion of stocks from the previous section (section 3.2.1) were ranked in descending order of their correlation coefficients. Among those stocks, we were only interested in those stocks that were high on both PC1 and PC2. We check through the list of ordered stocks and selected the first three top stocks with positive scores for both PC1 and PC2. Only these stocks were traded in the next and final stage.

If none of the stocks were chosen, then the current top sector is omitted from the list of sectors and we repeat the steps in section 3.1 and 3.2 on the remaining sectors to identify the top stocks in the next top sector.

3.3. Trading Strategy

The top stocks that were identified in the top sector based on the one-month training data formed a stock portfolio with equal weightage for each time-period. The stock portfolio was traded for the next one month, which is known as the trading period. We assumed a constant transaction cost of $25 and traded with an initial investment of $10,000 for each of the top stocks.

On the first day of the trading period, we purchased the maximum number of shares possible, for each of the top stocks, with the initial investment less transaction cost, as shown in equation (2).

$$S_0 = \left\lfloor \frac{I_0 - C}{P_0} \right\rfloor \tag{2}$$

where S_0 is the number of shares of the stock purchased on the first day of the trading period,

P_0 is the stock price on the first day of the trading period,

I_0 is the amount of initial investment for the stock,

C is the buying transaction cost,

and $\lfloor\ \rfloor$ is the floor function, rounding off to the nearest integer

Since $I_0 = 10{,}000$ and $C = 25$, equation (2) can be further simplified as follows.

$$S_0 = \left\lfloor \frac{9{,}975}{P_0} \right\rfloor \tag{3}$$

On the last day of the trading period, all the shares of stocks were sold at the market price. The total amount received in

exchange for the stocks is known as the cash value, calculated by equation (4) below.

$$V_1 = (S_0 \text{ x } P_1) - C \quad (4)$$

where V_1 is the cash value of the stock on the last day of the trading period,

C is the selling transaction cost of $25,

and P_1 is the stock price on the last day of the trading period

Net return [28] can be found by subtracting initial value of investments from final value of investment via the equation below.

$$Net\ Return = V_1 - (S_0 \text{ x } P_0) \quad (5)$$

The return on investment (ROI) can be calculated by the following equation.

$$ROI = Net\ Return\ /(S_0 \text{ x } P_0) \text{ x } 100\% \quad (6)$$

3.4. Confidence Interval for the Mean Stock Price During the Training Period

Using the training data, for each of the top stocks identified in section 3.2, a 99% confidence interval for the mean stock price was calculated using equation (7) below. We expect the mean stock price of the top stocks to move towards the upper bound of the 99% confidence interval during the trading period.

$$\bar{P} \pm Z_{0.005} \frac{s}{\sqrt{n}} \quad (7)$$

where $\bar{P}$ is the mean stock price,

$Z_{0.005}$ is the Z value for 99% confidence interval, equals 2.5758,

s is the standard deviation of the stock price,

and n is the number of days in the training period.

4. Results and Findings

Full analysis for the different sectors for three consecutive time periods was performed. We use the Materials sector for period 1, to illustrate our methodology and results in detail.

Correlation analysis was first conducted on all the sector indices to identify the top sector, next correlation analysis and PCA were applied to all the stocks in the top sector to identify the top three stocks.

4.1. Selection of Top Upward Trending Sector-Pearson Correlation Analysis

We applied the Pearson correlation analysis as specified in section 3.1 on all ten sector indices in the Australian stock market. The ten sectors are namely Consumer Discretionary (^AXDJ), Consumer Staple (^AXSJ), Energy (^AXEJ), Finance, (^AXFJ), Health Care (^AXHJ), Industrials (^AXNJ), Information Technology (^AXIJ), Materials (^AXMJ), Communication Services (^AXTJ) and Utilities (^AXUJ) [2].

Table 3. Correlation Coefficient and Probability Value for All Sectors in Period 1.

Sector Index	Correlation Coefficient	Probability Value
^AXMJ	0.9302	2.91E-09
^AXNJ	0.8983	7.62E-08
^AXEJ	0.8624	1.01E-06
^AXIJ	0.8601	1.16E-06
^AXDJ	0.7714	6.83E-05
^AXFJ	0.7033	5.41E-04
^AXTJ	0.6788	1.00E-03
^AXUJ	0.2772	2.37E-01
^AXHJ	-0.4453	4.91E-02
^AXSJ	-0.6854	8.52E-04

The sorted correlation coefficients of all ten sector indices in period 1 are shown in table 3 above. In this case, the Materials sector was selected as the top sector as it was the top on the list with the highest correlation coefficient value of 0.9302 and was statistically significant with a p-value of 2.91E-9.

All the stocks in Materials sector were selected for further analysis in the next section.

4.2. Selection of the Top Three Upward Trending Stocks

4.2.1. Pearson Correlation Analysis on Individual Stocks

We again applied the Pearson correlation analysis as specified in section 3.2.1 on all forty-six stocks in the Materials sector. The sorted correlation coefficients of all the stocks in the Materials sector in period 1 is shown in table 4 below. In the training period, there were eighteen well performing stocks with statistically significant, high and positive correlation coefficients. These stocks can be seen in table 4 below.

Table 4. Pearson Correlation Coefficient and Probability Value for All Stocks in the Materials sector in Period 1.

Stock	Correlation Coefficient	Probability Value	Stock	Correlation Coefficient	Probability Value
CSR. AX	0.9620	1.38E-11	NST. AX	0.6675	1.30E-03
AWC. AX	0.9545	6.79E-11	CII. AX	0.6518	1.85E-03
BKW. AX	0.9437	4.41E-10	BLD. AX	0.6142	3.97E-03
S32. AX	0.9383	9.85E-10	RRL. AX	0.5318	1.58E-02
AMC. AX	0.9352	1.51E-09	ORE. AX	0.5212	1.85E-02

Stock	Correlation Coefficient	Probability Value	Stock	Correlation Coefficient	Probability Value
ZIM. AX	0.9322	2.25E-09	BSL. AX	0.4823	3.13E-02
BHP. AX	0.9255	5.12E-09	NCM. AX	0.4672	3.78E-02
IGO. AX	0.9218	7.81E-09	ANO. AX	0.3860	9.28E-02
RIO. AX	0.9187	1.09E-08	SFR. AX	0.3541	1.26E-01
JHX. AX	0.9012	5.93E-08	ORA. AX	0.2697	2.50E-01
SGM. AX	0.8959	9.29E-08	OGC. AX	0.0918	7.00E-01
MWY. AX	0.8891	1.60E-07	ORI. AX	0.0224	9.25E-01
OZL. AX	0.8677	7.20E-07	MIN. AX	0.0078	9.74E-01
CYL. AX	0.8602	1.15E-06	FBU. AX	-0.1510	5.25E-01
AQG. AX	0.8482	2.31E-06	IPL. AX	-0.1566	5.10E-01
CIA. AX	0.8352	4.60E-06	WSA. AX	-0.3917	8.76E-02
FMG. AX	0.8010	2.20E-05	SBM. AX	-0.6369	2.53E-03
TBR. AX	0.7475	1.52E-04	KPT. AX	-0.8414	3.35E-06
LYC. AX	0.7438	1.70E-04	RND. AX	-0.8658	8.13E-07
ABC. AX	0.7367	2.12E-04	EVN. AX	-0.8893	1.57E-07
GXY. AX	0.7226	3.20E-04	PGH. AX	-0.8935	1.13E-07
ILU. AX	0.7000	5.89E-04	SAR. AX	-0.9245	5.71E-09
AGG. AX	0.6905	7.52E-04	NUF. AX	-0.9644	7.59E-12

4.2.2. Principal Component Analysis

We applied the PCA as specified in section 3.2.2. The KMO value was 0.863, greater than 0.5 and the Bartlett's test is statistically significant with p-value of 0.000. The values could be observed in table 5 below.

Table 5. KMO and Bartlett's Test.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.863
Bartlett's Test of Sphericity	Approx. Chi-Square	11,615.655
	df	171
	Sig.	0.000

The variables loadings on the first two principal components were 8.750 and 7.326 respectively, both greater than 1. PC1 and PC2 explained 46.05% and 38.56% of the variation respectively, which sums to 84.61% of the total variation. These values can be seen in table 6 below.

Table 6. Principal Component Analysis Total Variance Explained.

Total Variance Explained			
Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	8.750	46.050	46.050
2	7.326	38.560	84.611
3	1.974	10.389	95.000
Extraction Method: Principal Component Analysis			

The variables loaded in the rotated component matrix are shown in table 7. The Cronbach's alpha coefficient were 0.821 and 0.929 for PC1 and PC2 respectively, both greater than 0.7. This confirms the reliability of PC1 and PC2.

Table 7. Rotated Component Matrix.

Rotated Component Matrix [a]			
	Component		
	1	2	3
Levered Free Cash Flow	.959		
Total Debt	.948		
Gross Profit	.944		
Total Cash	.928		
Enterprise Value	.923		
Market Cap (intraday)	.918		
Operating Cash Flow	.911		
EBITDA	.906		
Revenue	.809		
Book Value Per Share		.886	
52 Week High		.880	
Diluted Earnings Per Share		.878	
200 Day Moving Average		.876	
50 Day Moving Average		.869	
52 Week Low		.868	
Revenue Per Share		.858	
Total Cash Per Share		.843	
Enterprise Value Revenue			.993
Price Sales			.993
Extraction Method: Principal Component Analysis Rotation Method: Varimax with Kaiser Normalization.			
a. Rotation converged in 4 iterations.			

Among the eighteen top performing stocks from section 4.2.1, there were only two tops stocks, AMC. AX and RIO. AX, with positive scores on both PC1 and PC2 as shown in table 8 below.

Table 8. AMC. AX and RIO. AX scores on PC1 and PC2.

Stock	PC1	PC2
AMC. AX	0.45	1.13
RIO. AX	3.85	8.91

In this case, only AMC. AX and RIO. AX were selected as the top stocks in period 1, which were traded in the next and final step.

4.3. Trading Strategy Results

Stocks AMC. AX and RIO. AX were traded as discussed in section 3.3. Assuming that I_0 is \$10,000 for each stock and C is \$25 per transaction. As shown in table 9, the variables S_0, V_1, $Net\ Return$ and ROI for the top three stocks were calculated based on equation (3) and (4) since P_0 and P_1 were known.

Table 9. Simple Trading Outcome for the Top Stocks in the Materials sector in Period 1.

Stock	P_0 ($)	P_1 ($)	S_0 (Shares)	V_1 ($)	*Net Retun* ($)	*ROI* (%)
AMC. AX	13.82	14.52	721	10,443.92	479.70	4.81
RIO. AX	87.16	93.17	114	10,596.38	660.14	6.64
Portfolio	-	-	-	21,040.30	1,139.84	5.73

From table 9, we see that the final cash value for our stock portfolio in period 1 was $21,040.30, which amounted to a net return of $1,139.84 and a ROI of 5.73% in the short span of a one month trading period. In the same trading period, the stock market index had a ROI of -0.19%, calculated based on the simple trading method.

By comparing the return on investment of our stock portfolio to the market index in period 1, we observed that our methodology was performing well as our stock portfolio's ROI was positive and more than five times that of the Australian stock market index.

4.4. Confidence Interval for the Mean Stock Price During the Training Period

The 99% confidence interval for the stocks AMC. AX, RIO. AX and WOW. AX were calculated using equation (7) as discussed in section 3.4.

Table 10. The 99% Confidence Interval for All Top Stocks Across Three Time Periods.

Training Period	Stock	$\bar{P}$ ($)	s	99% Confidence Interval	
				Lower Bound ($)	*Upper Bound* ($)
1	AMC. AX	13.50	0.3925	13.27	13.72
	RIO. AX	84.76	2.3590	83.41	86.12
2	AMC. AX	14.75	0.3033	14.57	14.93
3	WOW. AX	31.51	0.7399	31.08	31.95

From table 10, we see that the upper bound for the 99% confidence interval for AMC. AX RIO. AX, AMS. AX and WOW. AX were 13.72, 86.12, 14.93 and 31,95 respectively.

Table 11. Summary Statistics for All Top Stocks Across Three Time Periods.

Trading Period	Stock	Buy Price P_0 ($)	Selling Price P_1 ($)	Average Price $\bar{P}$ ($)	Standard Deviations
1	AMC. AX	13.82	14.52	14.11	0.2046
	RIO. AX	87.16	93.17	88.48	1.8600
2	AMC. AX	15.39	15.71	15.55	0.1647
3	WOW. AX	32.12	34.65	33.47	0.6816

We observed that both the trading period mean stock price ($\bar{P}$, 14.11) and the selling price (P_1, 14.52) for AMC. AX in table 11 was higher than its upper bound of the 99% confidence interval, in table 10 (13.72). Based on table 10 and 11, the same conclusion can be made for all the other top stocks. This shows that the stock price for all the top stocks across all 3 time periods had increase on average by the end of the trading period.

4.5. Trading Strategy Result Summary

Table 12 below shows the portfolio's cash value (V_1), net return and ROI for the Materials sector in period 1 and 2, and Consumer Staple sector in period 3.

Table 12. Summary of Stock Portfolios for Three Sectors Across Three Time Periods.

Period /Sector	Stock	V_1 ($)	*Net Retun* ($)	*ROI* (%)	Market Index ROI (%)
Period 1/Material	AMC. AX	10,443.92	479.70	4.81	− 0.19
	RIO. AX	10,596.38	660.14	6.64	
	Portfolio	21,040.30	1,139.84	5.73	
Period 2/Material	AMC. AX	10,161.56	188.84	1.89	0.39
	Portfolio	10,161.56	188.84	1.89	
Period 3/Consumer Staple	WOW. AX	10,716.50	759.30	7.63	2.46
	Portfolio	10,716.50	759.30	7.63	

(see Appendix Table 13 and Table 14 for a detailed calculation of period 2 and 3)

The three stock portfolios generated from the correlation analysis performed very well, with portfolios in all three time periods outperforming the Australian Stock Market Index. Based on table 12 above, we see that the ROI of our portfolio is higher than that of the stock market index in all three time periods, with the ROI in all three periods having at least more than three times higher ROI than the stock market index. Despite the market index performing fairly well in period 3,

achieving a ROI of 2.46%, the Consumer Staple portfolio still outperformed the market index by thrice as much, hitting a ROI of 7.63% within the short span of a one month trading period.

5. Conclusion

Most studies on stock portfolio selection were conducted through the use of advanced and sophisticated techniques. In this study, we focused on using one of the simplest and most fundamental statistical techniques – Pearson correlation analysis coupled with Principal Component Analysis, to identify winning stocks in the Australian stock market. The objective of this study was to examine whether the stock portfolios would outperform the stock market index over three consecutive time periods. In section 4.5, our three stock portfolios based on stocks selected from different sectors across three non-overlapping time periods demonstrated that our methodology was reliable and consistent across all 3 time periods, delivering significant profits from trading, further proving that our method was not only theoretically robust but also practically sound.

In conclusion, correlation analysis can be used to identify winning stocks trading on a short-term basis. Overall, our stock portfolios delivered return on investments at least 3.1 times higher than the stock market index over the three-month period.

Further research can be performed to test whether portfolio selection through the use of correlation analysis and principal component analysis can be applied to other stock markets. Further, adjustment can be made to the length of time period and/or weightage of the stocks in the portfolio to increase the profitability of the portfolios.

Appendix

Table 13. Simple Trading Outcome for the Top Stocks in the Materials sector in Period 2.

Stock	P_0 ($)	P_1 ($)	S_0 (Shares)	V_1 ($)	*Net Retun* ($)	*ROI* (%)
AMC. AX	15.39	15.72	648	10,161.56	188.84	1.89
Portfolio	-	-	-	10,161.56	188.84	1.89

Table 14. Simple Trading Outcome for the Top Stocks in the Consumer Staple sector in Period 3.

Stock	P_0 ($)	P_1 ($)	S_0 (Shares)	V_1 ($)	*Net Retun* ($)	*ROI* (%)
WOW. AX	32.12	34.65	310	10,716.50	759.30	7.63
Portfolio	-	-	-	10,716.50	759.30	7.63

References

[1] Alexander C. (2009) *Market Risk Analysis, Value At Risk Models*, Vol. 4 ed., John Wiley & Sons, San Francisco.

[2] Australian Securities Exchange-Sector Index Overviews. [online] https://www.asx.com.au/products/sector-indices.htm (Accessed 18 May 2020).

[3] Coronel-Brizio, HF, Hernandez-Montoya, AR, Olivares Sánchez, HR, and Scalas, E. (2012) 'Analysis of short term price trends in daily stock-market index data', *arXiv,* [online] https://arxiv.org/pdf/1211.3060.pdf (Accessed 16 May 2020).

[4] Cronbach, LJ (1951) 'Coefficient alpha and the internal structure of tests', *Psychometrika,* Vol. 16, No. 3, pp. 297-334.

[5] Egghe, L and Leydesdorff, L. (2009) 'The relation between Pearson's correlation coefficient and Salton's cosine measure', *Journal of the American Society for Information Science and Technology*. Vol. 60, Issue 5, pp. 1027-1036.

[6] Fama, EF and French, KR. (1988) 'Permanent and temporary components of stock prices', *Journal of Political Economy,* Vol. 96, No. 2, pp. 246-273.

[7] Fama, EF and French, KR. (1988) 'Dividend yields and expected stock returns', *Journal of Financial Economics,* Vol. 22, No. 1, pp. 3-25.

[8] Fama, EF and French, KR. (1992) 'The cross-section of expected stock returns', *The Journal of Finance,* Vol. 47, No. 2, pp. 427-465.

[9] Ferson, WE, and Harvey, CR. (1991) 'The Variation of Economic Risk Premiums', *Journal of Political Economy*, Vol. 99, No. 2, pp. 385-415.

[10] Ferson, WR and Harvey, CR. (1993) 'The Risk and Predictability of International Equity Returns', *Review of Financial Studies*, Vol. 6, Issue 3, pp. 527-566.

[11] Gajanan, LA. (2008) *Financial Forecasting: Comparison of ARIMA, FFNN, SVR Models*, Unpublished MTech thesis, Indian Institute of Technology, Bombay, India.

[12] Hargreaves, CA, Dixit, P and Solanki, A. (2013) 'Stock Portfolio Selection using Data Mining Approach', *IOSR Journal of Engineering*, Vol. 3, Issue 11, pp. 42-48.

[13] Hotelling H. (1933) 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology,* Vol. 24, No. 6, pp. 417.

[14] Ince H and Trafalis TB. (2007) 'Kernel principal component analysis and support vector machines for stock price prediction', *IIE Transactions*, Vol. 39, No. 6, pp. 629-637.

[15] Jaffe J, Keim DB and Westerfield R. (1989) 'Earnings yields, market values, and stock returns', *The Journal of Finance,* Vol. 44, No. 1, pp. 135-148.

[16] Jolliffe I. (2011) 'Principal component analysis', *In: International Encyclopedia of Statistical Science, Springer*, pp. 1094-1096.

[17] Kaiser, HF. (1974) 'An index of factorial simplicity', *Psychometrika*, Vol. 39, No. 1, pp. 31-36.

[18] Krzywinski, M, and Altman, N. (2013) 'Significance, P values and t-tests', *Nature Methods*, Vol. 10, No. 11, pp. 1041-1042.

[19] Mani, CK and Hargreaves, CA. (2015) 'The Selection of Winning Stocks Using Principal Component Analysis', *American Journal of Marketing Research,* Vol. 1, No. 3, pp. 183-188.

[20] Mani, CK and Hargreaves, CA. (2016) 'Stock Trading using Analytics', *American Journal of Marketing Research*, Vol. 2, No. 2, pp. 27-37.

[21] Mani, CK. (2015) 'Web scraping in a simple way' [online] 5 April. http://chandrikakadirvelmani.blogspot.sg/2015/04/web-scraping-in-simple-way.html (Accessed 18 May 2020).

[22] Markowitz, HM. (1959) *Portfolio Selection: Efficient Diversification of Investments*, Yale University Press, New York.

[23] Mbeledogu, NN, Odoh, M and Umeh, MN. (2012) 'Stock feature extraction using Principle Component Analysis', *International Conference on Computer Technology and Science*, IACSIT Press, Singapore DOI: 10.7763/IPCSIT.2012.V47.44.

[24] Moore, AB. (1964) 'Some Characteristics of Changes in Common Stock Prices', in Paul H. Cootner (Eds.), *The Random Character of Stock Market Prices*, MIT Press, Cambridge, MA, pp. 139-161.

[25] Narayan PK, Narayan S and Thuraisamy KS. (2014) 'Can institutions and macroeconomic factors predict stock returns in emerging markets?', *Emerging Markets Review,* Vol. 19, pp. 77-95.

[26] Obilor, EI and Amadi, EC. (2018) *Test for Significance of Pearson's Correlation Coefficient (r).* https://www.researchgate.net/publication/323522779_Test_for_Significance_of_Pearson's_Correlation_Coefficient (Accessed 18 May 2020).

[27] Pasini G. (2017) 'Principal component analysis for stock portfolio management', *International Journal of Pure and Applied Mathematics,* Vol. 115, No. 1, pp. 153-167.

[28] Schlarbaum, G, Lewellen, W and Lease, R. (1978) 'Realized Returns on Common Stock Investments: The Experience of Individual Investors', *The Journal of Business,* Vol. 51, No. 2, pp. 299-325.

[29] Shukla R and Trzcinka C. (1990) 'Sequential tests of the arbitrage pricing theory: a comparison of principal components and maximum likelihood factors', *The Journal of Finance*, Vol. 45, No. 5, pp. 1541-1564.

[30] Washer, KM, Jorgensen, R and Johnson, RR. (2016) 'The increasing volatility of the stock market?', *The Journal of Wealth Management*, Vol. 19, Issue 1, pp. 71-82.

[31] Yadav, PK and Pope, PF. (1994) 'Stock index futures mispricing: profit opportunities or risk premia?', *Journal of Banking & Finance*, Vol. 18, Issue 5, pp. 921-953.

[32] Yahoo Finance – stock market live, quotes, business & finance news. (n. d.). (online) https://au.finance.yahoo.com/ (Accessed 16 May 2020).