analytics

MDPI

Article

Identification of Patterns in the Stock Market through Unsupervised Algorithms

Adrian Barradas *, Rosa-Maria Canton-Croda and Damian-Emilio Gibaja-Romero

Graduate School of Engineering, UPAEP-University, Puebla 72410, Mexico;
rosamaria.canton@upaep.mx (R.-M.C.-C.); damianemilio.gibaja@upaep.mx (D.-E.G.-R.)
* Correspondence: adrian.barradas@upaep.edu.mx

Abstract: Making predictions in the stock market is a challenging task. At the same time, several studies have focused on forecasting the future behavior of the market and classifying financial assets. A different approach is to classify correlated data to discover patterns and atypical behaviors in them. In this study, we propose applying unsupervised algorithms to process, model, and cluster related data from two different data sources, i.e., *Google News* and *Yahoo Finance*, to identify conditions in the stock market that might help to support the investment decision-making process. We applied principal component analysis (PCA) and a *k*-means clustering approach to group data according to their principal characteristics. We identified four conditions in the stock market, one comprising the least amount of data, characterized by high volatility. The main results show that, regularly, the stock market tends to have a steady performance. However, atypical conditions are conducive to higher volatility.

Keywords: k-means clustering; stock market; principal component analysis; economic indicators; unsupervised algorithms; financial news; SPY; S & P500

Citation: Barradas, A.; Canton-Croda, R.-M.; Gibaja-Romero, D.-E. Identification of Patterns in the Stock Market through Unsupervised Algorithms. *Analytics* **2023**, *2*, 592–603. https://doi.org/10.3390/analytics2030033

Academic Editor: Domenico Ursino

Received: 5 May 2023
Revised: 12 July 2023
Accepted: 20 July 2023
Published: 27 July 2023

1. Introduction

The proliferation of financial and economic news plays a vital role in determining asset prices since they promote the stock's market short- and long-term volatility [1,2]. Specifically, the empirical evidence points out a relationship between economic news and market fluctuations since high media coverage increases the trading volume in the stock market [3–5]. Hence, investors search for the right tools to understand the financial market behavior and make decisions that increase their benefits, given the wide variety of financial assets that they can trade in the stock market [6]. Although many studies estimate asset volatility by determining the economic indicators' impact on the stock market [7,8], forecasting asset prices is still challenging given the complex correlations between the variables that characterize stock market relationships [9]. Hence, classification techniques represent an alternative to predicting financial asset prices, given their capacity to analyze the stock market instead of solely integrating new data [10,11].

Over the past two decades, exchange-traded funds (ETFs) have become one of the most popular investment vehicles among retail and professional investors due to their low transaction costs and high liquidity [12,13]. The popularity of ETFs has caught the attention of academicians and decision makers, given the impact of financial news on asset prices through social networks, whose use has also increased in recent years [14]. This work aims to identify asset price changes by implementing unsupervised algorithms. Specifically, we consider financial news and economic indicators to identify stock market trends that help investors to make decisions. We focus on analyzing the *SPY* fund since it is a gauge of large-cap equities from the United States [15].

We provide a methodological approach based on classification techniques to discover hidden patterns, similarities, and differences in data gathered from the SPDR S & P 500

exchange-traded fund. Specifically, we propose a two-stage methodology that implements the principal component analysis in the first stage, where we identify the importance of each comprised variable. In the second stage, we apply the *k*-means clustering technique to classify market data. So, the previous framework deals with the issues related to forecasting techniques since it allows us to identify the behavior of asset prices without explicitly knowing what patterns to identify [16,17].

We organize this paper as follows. Section 2 presents a literature review concerning the literature closely related to our study. Section 3 describes the materials and methods used for the study's development in detail. Section 4 presents the results of processing and classifying data using the proposed unsupervised algorithms. Section 5 discusses the main results related to the patterns that the previous methodology identified. Finally, Section 6 summarizes the main findings and future works for this study.

2. Related Work

Artificial intelligence methods have become very popular for identifying relationships between asset prices and economic variables [18]. Supervised algorithms, such as artificial neural networks (ANNs) and support vector machines (SVMs), have succeeded in finding patterns and correlations between uncorrelated datasets in the stock market [19]. Moreover, unsupervised algorithms also identify hidden patterns in unlabelled data [18].

One of the advantages of applying unsupervised machine learning excels is their capacity to discover and organize data without being attached to assessing the potential solution [20]. For example, the *k*-means clustering technique groups data with similar characteristics into *k* sets [21,22], simplifying financial market analysis [23]. So, *k*-means clustering can be applied to segment active companies in the Tehran Stock Exchange (TSE) [23] or to obtain trading signals based on financial ratios [24]. Also, the performance of the previous algorithms can improve by combining them with other techniques. For example, the artificial fish swarm algorithm allows the *k*-means algorithm to classify 100 stocks in the Chinese stock exchange into poor and good performance [25].

While machine learning techniques aim to model the underlying distribution of data to discover patterns and information, it is typical for their combination with unsupervised statistical methods to reduce the dimensionality of data before applying machine learning. Concerning the previous objective, principal component analysis (PCA), fuzzy robust principal component analysis (FRPCA), and kernel-based principal component analysis (KPCA) are the most common techniques used to reduce data dimensionality. The PCA method excels over the previous ones because it provides a higher classification accuracy than the others [5]. Some studies have used the previously mentioned statistical techniques and the *k*-means method to classify companies according to their characteristics. For example, ref. [26] applied both methods to identify the most contributing companies among twenty economic sectors in Australia, and ref. [27] used a similar approach to classify companies listed on the Indonesia Stock Exchange in 2019 and 2021 to find the best-performing stocks before and during the COVID-19 pandemic. In [28,29], the authors concluded that such unsupervised algorithms could be used for stock trend forecasting by finding patterns in data and reducing the risk during the decision-making process. Some other studies used both techniques, applying them separately [30] or using them with a different purpose than identifying patterns in the stock market [31].

In general, the implementation of classification techniques in the financial market focuses on using the intrinsic data of companies to cluster their assets. So, none of them considered financial news or economic indicators for the analysis. To the best of our knowledge, our study is the first that combines both techniques (PCA and *k*-means) to search for patterns in the stock market by considering financial news, transactional data of stock shares, and economic indicators.

3. Materials and Methods

In this work, we pretend to identify general patterns that characterize the pricing of assets within the stock market. Given the complexity of such a market, we propose a two-stage methodology based on the application of unsupervised algorithms. The first stage concerns *data collection* and *transformation*. Given its importance, we gathered data from the *SPY* ETF [32], which comprises selected stocks from five hundred issuers, all listed on U.S. stock exchanges. Also, it spans approximately twenty-four separate industry groups [32]. Later, the second stage focuses on data modeling through the unsupervised *k*-means clustering algorithm. Figure 1 illustrates the previous approach.

We coded and carried out computational routines employing Python, a programming language that provides data collection, visualization, preparation, and analysis tools through the Python library repository (PyPI: Python Package Index). Specifically, we used the free-to-use package `scikit-learn` to implement the PCA and *k*-means clustering algorithms. It is worth recalling that this study only reviews data behavior generally, which is why we did not consider the effect of seasonal variations over time. Our approach pretends to provide a classification analysis to avoid issues related to statistical significance. This last analysis remains an open question for future studies.

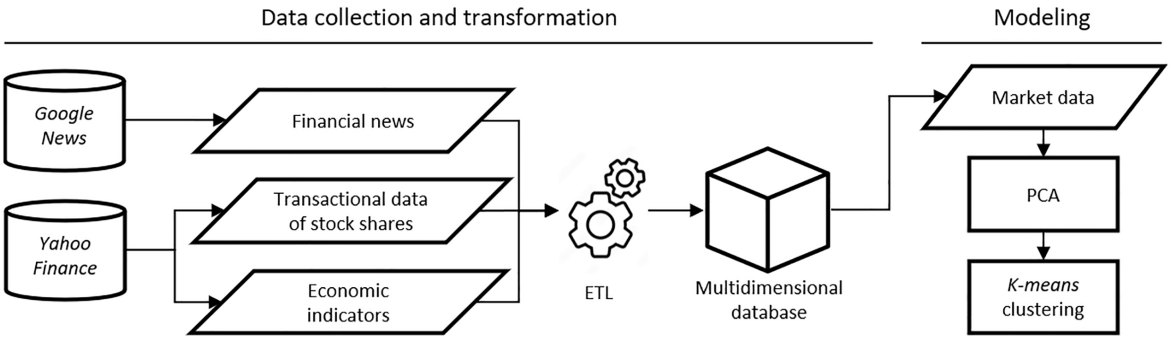


Figure 1. Proposed methodology. Source: compiled by authors.

3.1. Data Collection and Transformation

We gathered data from *Yahoo Finance* and *Google News* using an ETL (extract, transform, and load) process. After that, we generated three datasets concerning financial news (22,222 observations), transactional data of stock shares (2015 observations), and economic indicators (16,256 observations). The three follow a standard notation concerning data transformation, deletion of duplicate entries, replacement of null values, and the computation of new variables. Given that the financial news dataset comprises text data, we transformed them into numerical values by applying VADER (Valence Aware Dictionary and sEntiment Reasoner). This parser classifies a text string as positive or negative, and we obtained it from the Python library NLTK (Natural Language ToolKit Version: 3.6.5). VADER uses the English language's semantics and lexicon to provide scores ranging from -1 to 1. So, the parser assigns attributes that indicate whether the news is negative, positive, or neutral. The processing of financial news data through VADER does not allow for a transparent overview of the operation; nevertheless, VADER is commonly and extensively used in sentiment analysis research [33–35].

Later, we calculated two additional attributes. *Compound* reflects the weighted sum of the former attributes, and *compound sq* represents the popularity of the stock in the media, regardless of its sentiment [36,37]. Then, we related the three datasets through a multidimensional model that induces the creation of the market dataset, which comprises 2013 positions corresponding to daily data from 1 January 2014 to 31 December 2021. It is worth mentioning that data are only available for the days the U.S. stock exchange operates.

Respectively, Tables 1 and A1 describe the variables in this data set and show a sample of data.

Table 1. Description of variables in the market dataset.

Variable	Definition	Source
news	Weighted average of the sentiment score given to collected financial news related to the ETF (SPY).	Google News
news Sq	Squared value of variable *news*.	Google News
variation	Stock price variation with respect to the previous day.	Yahoo Finance
VIX	Volatility index.	Yahoo Finance
EEM	MSCI emerging markets ETF.	Yahoo Finance
TNX	Interest on 10-year treasury bonds.	Yahoo Finance
CLF	Crude oil future contracts.	Yahoo Finance
USD/MXN	Exchange rate between the U.S. dollar and the Mexican peso.	Yahoo Finance
USD/EUR	Exchange rate between the U.S. dollar and the euro.	Yahoo Finance
GFC	Gold future contracts.	Yahoo Finance

3.2. Modeling

The market dataset considers ten variables. So, we simplified the analysis by reducing the dimensionality of this dataset through principal component analysis (PCA). PCA is an unsupervised linear transformation algorithm that produces new features called principal components (PCs) by determining the maximum variance of the data [38]. After that, we applied the *k*-means clustering method, an unsupervised machine learning technique, to group data with similar features. In other words, it serves to identify conditions in the stock market that characterize assets.

3.3. Principal Component Analysis

We applied principal component analysis to avoid the curse of dimensionality and to process data faster [39]. We used the free-to-use package `scikit-learn` for the analysis, which includes matrix decomposition algorithms in its *decomposition* module [40]. To determine the number of components more suitable for dimensionality reduction, we observed the *eigenvalues* of the corresponding covariance matrix and their proportion of explained variance. On the other hand, we also identified in which proportion each principal component represents each variable's variance. Consequently, together with the description of the variables in Table 2, we defined a conceptual meaning for each selected principal component.

Table 2. Eigenvalues from the PCA.

PC	Eigenvalue	Proportion	Cumulative Proportion
PC1	4.1470	0.379	0.379
PC2	2.6312	0.240	0.620
PC3	1.5831	0.144	0.765
PC4	1.0038	0.091	0.857
PC5	0.5113	0.046	0.904
PC6	0.4564	0.041	0.945
PC7	0.2762	0.025	0.971
PC8	0.1832	0.016	0.988
PC9	0.0786	0.007	0.995
PC10	0.0522	0.004	0.999

Source: compiled by authors.

3.4. K-Means Clustering

Once we reduced data dimensionality, we implemented the *k*-means clustering method. First, we determined an optimal number of clusters by applying the *elbow method* and *silhouette coefficient*. Then, we classified data according to the clusters selected. The selection of *k* may affect the performance of the clustering algorithm. Therefore, we chose a set of values for *k*. In that regard, it is essential for the number of values considered to be sufficiently large to reflect the specific characteristics of the dataset and to be significantly smaller than the number of objects in the dataset, which is the primary motivation for performing data clustering [21]. Thus, we analyzed the clustering of data in a range of *k* values according to the results obtained from applying the *elbow method* and *silhouette coefficient*. By applying this approach, we can compare the behavior of data when clustering them in different groups and therefore identify other characteristics in the stock market.

4. Results

Recalling that we followed a two-stage methodological approach, we first discuss the implications of applying PCA to simplify the stock market dataset. Later, we present the patterns that the clustering method identified.

4.1. Dimensionality Reduction

We significantly reduced the dimensionality of the dataset by applying PCA. In our case study, the cumulative proportion value concerning the five principal components indicates that these components adequately explain 90.4% of the variance (see Table 2). In this regard, we only used five components instead of ten to faithfully represent the whole data's performance, reducing the model's complexity and processing time.

Regarding the principal components' composition concerning the original variables, as shown in Table 3, all principal components are made up of the variables in the *market data* dataset, some of them unveiling more influence than others. For example, financial news (*News* and *News Sq*) shows a more significant impact over the third and fifth principal components. In contrast, the first, second, and fourth components are influenced to some extent by the variables equally.

Table 3. Variance explained per component.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Cumulative Proportion	**0.379**	**0.620**	**0.765**	**0.857**	**0.904**	**0.945**	**0.971**	**0.988**	**0.995**	**0.999**
news	0.350	0.018	−0.100	−0.260	0.890	0.005	−0.066	−0.026	−0.025	−0.037
news Sq	0.249	0.441	0.841	−0.154	−0.053	0.055	0.074	0.009	−0.009	−0.009
VIX	0.259	−0.213	0.137	0.620	0.074	0.585	−0.206	−0.302	−0.008	0.046
EEM	0.258	0.393	−0.339	−0.149	−0.206	0.032	0.190	−0.617	−0.426	0.031
TNX	−0.378	0.163	−0.088	−0.327	0.030	0.660	0.123	−0.025	0.196	−0.474
CLF	−0.152	0.516	−0.174	0.058	0.013	0.174	−0.624	0.356	−0.269	0.242
USD/MXN	0.372	−0.272	−0.079	−0.220	−0.180	0.377	0.324	0.526	−0.374	0.192
USD/EUR	0.089	−0.447	0.154	−0.515	−0.219	0.003	−0.609	−0.240	−0.127	−0.106
ESF	0.419	0.160	−0.242	−0.178	−0.222	0.102	−0.096	0.005	0.742	0.299
GFC	0.441	0.117	−0.148	0.216	−0.164	−0.186	−0.149	0.255	−0.010	−0.758

Source: compiled by the authors.

Since the principal components are linear combinations of the original variables in Table 1, they do not have a clear conceptual meaning. However, based on their composition and their higher weight variables, we set their meaning as follows:

- PC1: Gold and S & P500 future expectations.
- PC2: Oil and USD/EUR exchange rate future expectations.
- PC3: Media coverage.
- PC4: Volatility.
- PC5: Financial news sentiment.

4.2. Clusters Comparison

Figure 2a shows a graph that represents the implementation of the *elbow method*, while Figure 2b presents the *silhouette method*. Note that the optimal number of clusters lies between 2 and 4. In the case of the former, the elbow in the graph is visible for k equal to 2, while the graph for the *silhouette coefficient* shows the most significant value when k is equal to 4. In either case, given these results, we explored the k-means clustering model by considering both approaches, i.e., we applied and compared the results of the k-means technique by considering values of k equal to 2, 3, and 4. In other words, we classified and compared data by determining two, three, and four clusters.

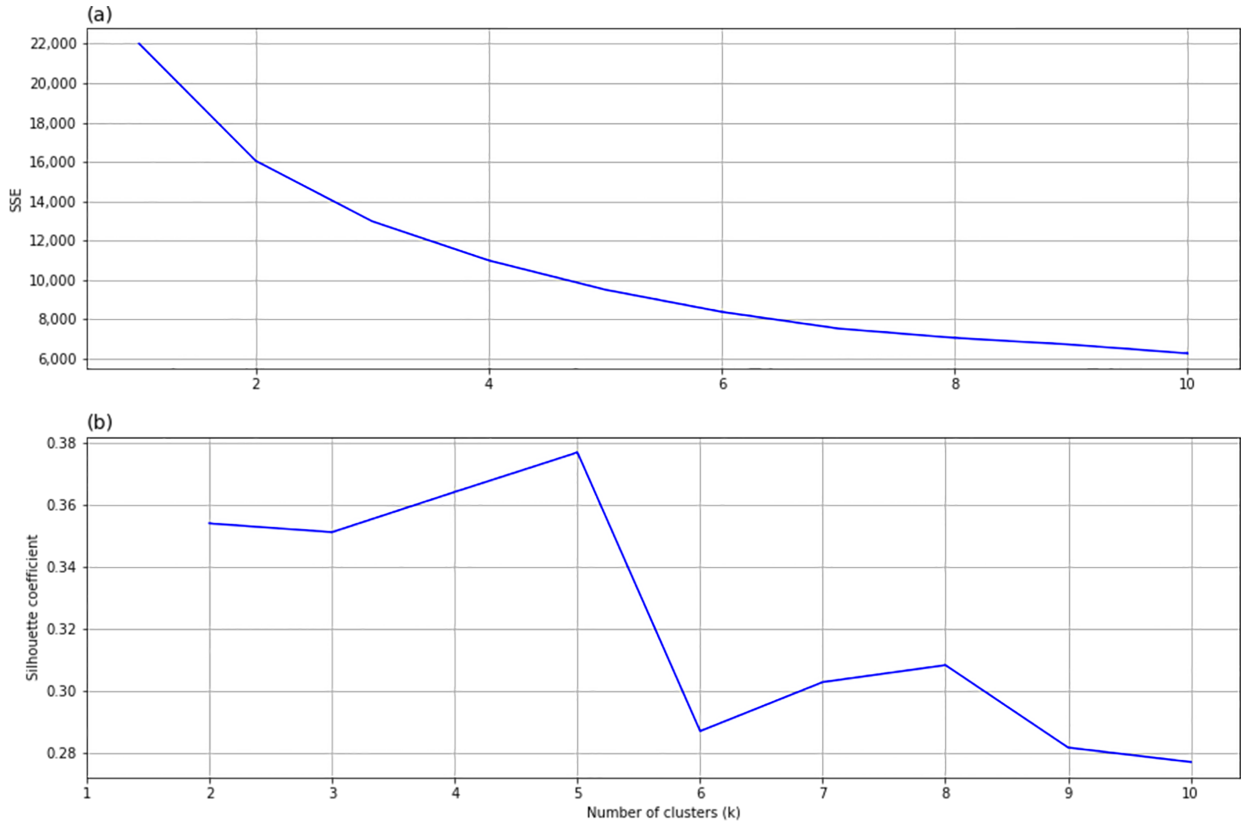
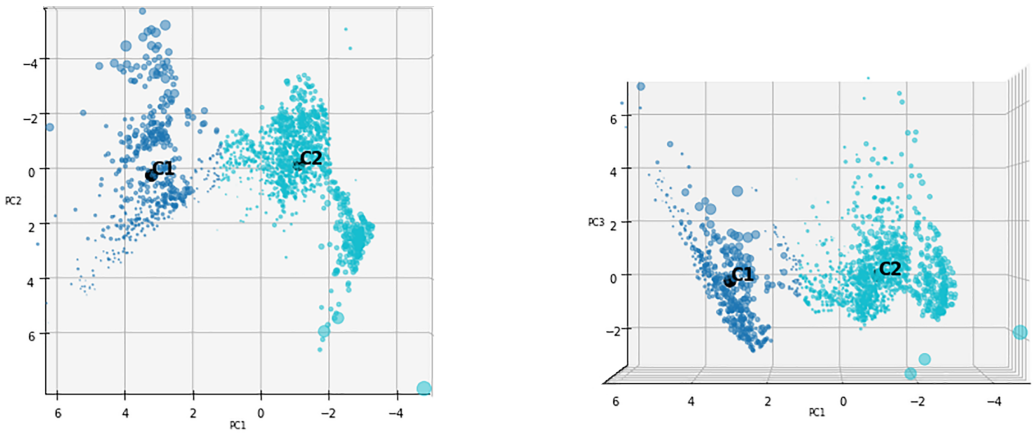


Figure 2. Determining the number of clusters. (**a**) Elbow method. (**b**) Silhouette method.

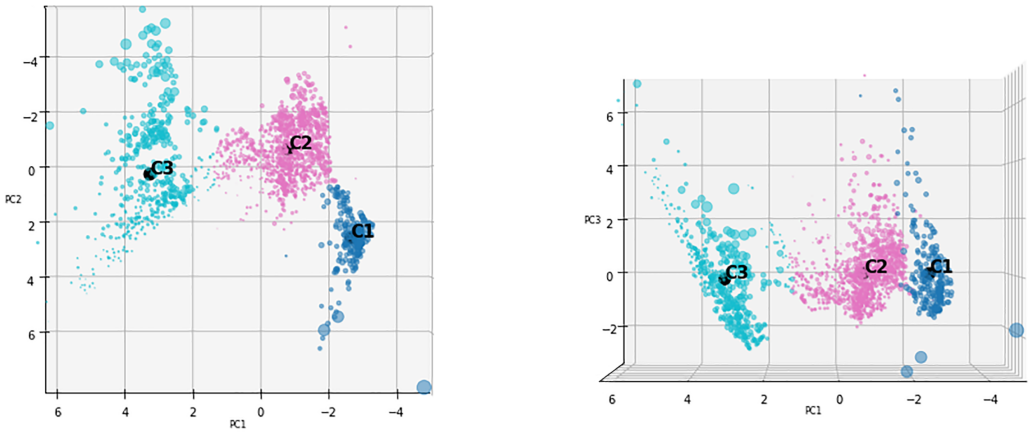
Graphically, we first provide a three-dimensional representation of the dispersion of the data in Figure 3. Later, Figure 4 shows the data classification into groups. At the same time, Figure 5 illustrates the centroids through the scatter plots. Note that the previous representation allows us to show the four principal components: the first three (*PC1, PC2*, and *PC3*) are associated with the coordinates of the axes X, Y, and Z, respectively, and the last major component (*PC4*) is appreciable through the size of each data point. At first sight, we can clearly distinguish between negative and positive *PC1* values. Although concentrated more on the negative side of *PC2*, data show a slight dispersion trend toward positive values. Concerning *PC3*, a tendency toward positive values is also recognizable. Notably, the impact of *PC4* is more significant at the extremes of the rest of the components, whereas the data point size has a more significant variation.



(**a**) *PC1–PC2* plane view (**b**) *PC1–PC3* plane view

Figure 3. Scatter plot of data grouped into two clusters: $k = 2$. Centroids are displayed in black.

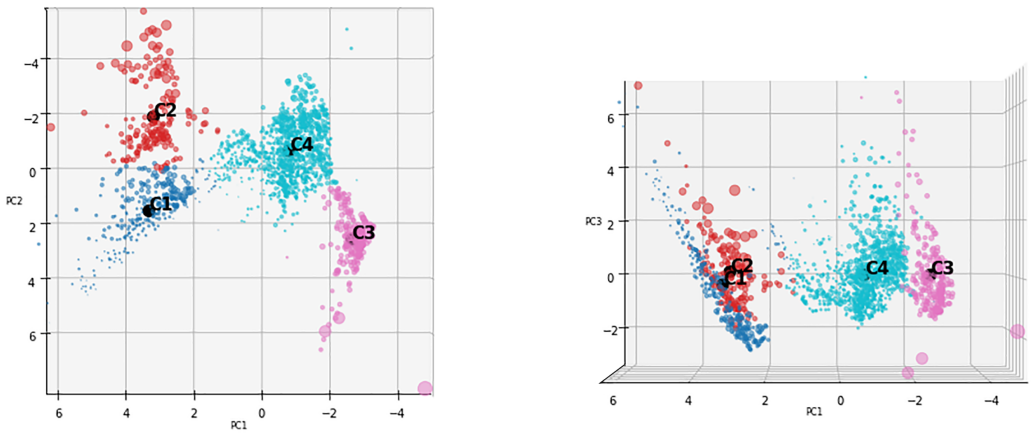
As expected, we obtained new clusters with different characteristics as the value of k increases. Also, the centroids' locations change and data appear to be further clustered. In the case of $k = 3$—see Figure 4—data associated with negative *PC1* values are split into two clusters, one (C1) for those whose values of *PC4* and *PC2* are high but with low values for *PC1*; and the other concerns the rest (C2). It is worth emphasizing that cluster (C3) remains unchanged on the right side.



(**a**) *PC1–PC2* plane view (**b**) *PC1–PC3* plane view

Figure 4. Scatter plot of data grouped into three clusters: $k = 3$. Centroids are displayed in black.

Notice that the *elbow* and *silhouette* methods present a better performance when $k = 4$. The positive values of *PC1* (in the data's right side) are divided into two groups by computing the corresponding clusters. The split of the cluster is similar to that for the left side cluster; see Figure 4. In other words, we obtained a new cluster for those data with high *PC4* values and negative values for *PC2*. In this case, clusters have no significant differences when observing *PC3*. We recalculated their centroids correspondingly; we observe them in Figure 5.



(**a**) *PC1–PC2* plane view (**b**) PC1–PC3 plane view

Figure 5. Scatter plotof data grouped into four clusters: $k = 4$. Centroids are displayed in black.

Finally, Figure 6 shows clusters in an axonometric projection. In this one, the size of each data point shows the percentage of daily variation; that is, the bigger the data point, the higher its daily variation. We can see that C2 contains points whose sizes are larger than the rest. C3 contains data points with high daily variation values but only at positive values for *PC2*.

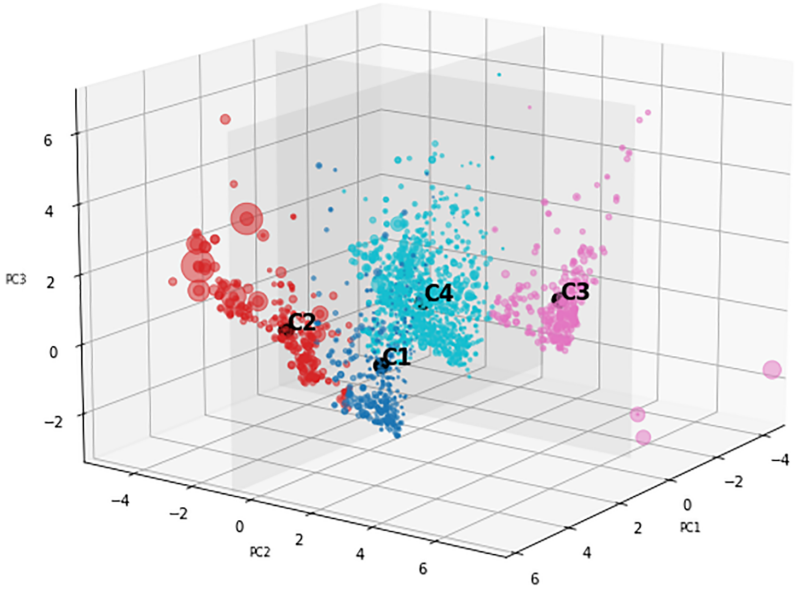


Figure 6. Three-dimensional scatter plot of data grouped into four clusters: $k = 4$. The size of each data point is represented by the daily variation.

Table 4 summarizes the statistical properties of each cluster, i.e., mean values and variance, together with the maximal and minimal daily variation in the *SPY* ETF. Regarding those metrics, it is worth noting the behavior of cluster C2. It contains the lowest proportion of data (9.04%) but also presents the highest value for variance (5.11), identifiable in their minimum and maximum daily variation values, which range between −10.21% and 8.37%. Thus, we can conclude that the main characteristic of this cluster is high volatility. Since we can consider that last one as a synonym of risk [41,42], it might help investors to identify

challenging market conditions and to make better investment decisions. In this case, we can suggest avoiding those assets belonging to C2.

Table 4. Statistical metrics by cluster.

	Data		Daily Variation [%]			
	Qty.	**%**	**Max.**	**Min.**	**Variance**	**Average**
C1	310	15.40%	2.20	−2.32	0.582	0.04650
C2	182	9.04%	8.37	−10.21	5.114	0.03212
C3	237	11.77%	1.79	−2.14	0.400	−0.00062
C4	1284	63.79%	4.64	−3.96	0.631	−0.01493

Source: compiled by the authors.

The majority of data (63%) concentrate on cluster C4. This cluster presents low expectations in future gold and S & P 500 prices (*PC1*), oil, and USD/EUR exchange rate (*PC2*). A low volatility indicates that, according to data, the daily variations in the stock market tend toward regular and steady values. Hence, we can consider non-risky assets in such a cluster, which provides decision makers confidence to invest in them. Concerning clusters C1 and C3, whose centroids are on the positive side of axis Y (*PC2*), their data show homologous behavior when observing the statistical metrics in Table 4. Minimum and maximum values of daily variation in the ETF's price, as well as their corresponding variance, are similar to each other. The main difference between them is that they have opposite values for *PC1*; hence, C1 and C3 can be interpreted as optimistic and pessimistic clusters, respectively, since the component *PC1* refers to the future expectations concerning the value of gold and the S & P 500 index.

It is worth noting that *PC3* data (associated with media coverage) show approximately the same distribution, regardless of the cluster. In all cases, cluster centroids are close to zero in *PC3*. In Figure 6, it is noticeable that, concerning *PC2*, *PC3* presents a linear correlation, which is positive or negative depending on each cluster. This latter point, together with the other findings in the data, allows us to assume that stock market movements, financial news, and economic data are intrinsically correlated, and that the market behaves according to the current economic conditions, while, most of the time, it has a steady behavior. Lastly, it is worth noting that the fifth principal component (*PC5*), while not plotted because it exceeds the capabilities of graphical dimensional representation, is considered in the *k*-means clustering model.

We can summarize the previous discussion by identifying each cluster with the feature that characterizes it. So, we have that:

- C1: Optimistic expectations.
- C2: High volatility.
- C3: Pessimistic expectations.
- C4: Usual steady behavior.

5. Discussion

According to the PCA results, five components represent 90.4% of data variance, which reduces the dataset size from ten to five variables. Also, this simplifies the model's complexity and processing time. In general terms, the selected principal components comprise proportional parts of the ten variables in the original dataset; however, some have greater importance than others. Even though the principal components do not have a clear definition because of their composition, we defined a conceptual meaning for each according to their significant weight variables. In this regard, clusters represent specific economic and stock market conditions.

By applying two methodologies (*elbow method* and *silhouette method*) for determining the optimal number of clusters (*k*) to classify the new dataset obtained from the PCA, we

found that a value for k equal to four is adequate for clustering the data. However, different market conditions are identified depending on the number of clusters.

Clustering data in two by considering the components' conceptual meaning results in the separation of data mainly by positive and negative expectations of gold and S & P500 price. Grouping data into four clusters splits the two former clusters into two that are, in this case, based on positive and negative expectations of oil and USD/EUR exchange rate prices, but also media coverage and volatility. Having four clusters give us a more detailed insight into the data. Therefore, it is possible to distinguish between different market states. We identified four conditions, one concentrating about 9% of data and showing particular features related to volatility; data clustered into this group could be considered atypical since the variance of the daily variation is about 5.11%.

Moreover, like previous studies [3–5], data behave differently when media coverage (*PC3*) is higher. Under this condition, clusters C1 and C3 exhibit higher volatility (*PC4*) values than the others. Remarkably, the percentage of daily variation for data in C1 is similar to that in C3, which may point out that the risk of investing in the stock market is the same in either case. On the other hand, most data (63.79%) lie in cluster C4, characterized by a steady behavior according to its features. In words, it indicates that the stock market operates, most of the time, under similar conditions.

6. Conclusions

Understanding the stock market's behavior provides valuable insights for decision making when trading financial assets. Specifically, relating data from different sources and clustering them can help investors to seize better investment opportunities. In this paper, we use unsupervised machine learning to find patterns and describe specific trading conditions in the market concerning other variables. We could categorize the market's behavior into four groups, each representing a market state with specific features that help to define optimal timing for investing. Our main contribution relies on recognizing that, although financial news affects the behavior of the stock market, the percentage of daily variation is not significantly affected. Also, we observe that high volatility is less frequent and can be considered an atypical behavior. From this finding, we conclude that the stock market behaves steadily most of the time; nevertheless, there are short periods when investors should make decisions carefully to prevent or reduce losses during high-risk times. In that context, distinguishing between high- and low-volatility periods provides advantages to obtaining higher profits.

From a technical point of view, our results show that the k-means clustering and PCA help to explore and analyze data since they provide a better understanding of the stock market through data characterization. This work helps to broaden the literature by providing a framework for detecting typical and atypical conditions in the stock market. While we analyzed daily data in this study, a different approach would be to examine data in a lower granularity, e.g., weekly or monthly. Considering the seasonality in the stock market might provide a broad overview of its behavior under specific periods. In future studies, we will pretend to apply different criteria for identifying the optimal number of clusters since such a classification technique only provides exclusive results. Regarding the clustering algorithm, a "mixture modeling" technique could be applied to identify the degree of data belonging to different clusters since it assigns data partially to one or more groups; this could help to find particular unrecognized patterns and insight regarding the behavior of the stock market.

Author Contributions: Methodology, A.B.; supervision, D.-E.G.-R. and R.-M.C.-C.; writing—review and editing, A.B. and D.-E.G.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Yahoo Finance and Google News and are available at https://finance.yahoo.com, accessed on 14 January 2023 and https://news.google.com, accessed on 14 January 2023 with the permission of Yahoo Finance and Google News.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Example of Market dataset.

Date	*News*	*News Sq*	*Variation*	*VIX*	*EEM*	*TNX*	*CLF*	*USD/MXN*	*USD/EUR*	*ESF*	*GCF*
3 January 2014	−0.134	0.018	−0.000	13.76	40.12	2.995	93.96	13.09	0.731	1825.5	1238.4
6 January 2014	−1.586	2.51	−0.0028	13.55	39.74	2.961	93.43	13.06	0.735	1820.75	1237.8
7 January 2014	−0.595	0.35	−0.0071	12.92	39.91	2.937	93.67	13.07	0.733	1830.75	1229.4
29 December 2021	4.332	18.76	0.0012	16.95	48.56	1.543	76.56	20.66	0.883	4784.5	1805.1
30 December 2021	1.6	2.60	−0.0027	17.33	49.09	1.515	76.99	20.55	0.880	4772.25	1812.7
31 December 2021	3.029	9.14	−0.0025	17.22	48.85	1.512	75.21	20.45	0.883	4758.5	1827.5

Source: compiled by the authors.

References

1. Peress, J. The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes. *J. Financ.* **2014**, *69*, 2007–2043.
2. Rangel, J.G. Macroeconomic News, Announcements, and Stock Market Jump Intensity Dynamics. *J. Bank. Financ.* **2011**, *35*, 1263–1276. [CrossRef]
3. Alanyali, M.; Moat, H.S.; Preis, T. Quantifying the Relationship Between Financial News and the Stock Market. *Sci. Rep.* **2013**, *3*, 3578. [CrossRef]
4. Goonatilake, R.; Herath, S. The Volatility of the Stock Market and News. *Int. Res. J. Financ. Econ.* **2007**, *3*, 53–65.
5. Zhong, X.; Enke, D. Forecasting Daily Stock Market Return Using Dimensionality Reduction. *Expert Syst. Appl.* **2017**, *67*, 126–139. [CrossRef]
6. Chen, T.L.; Chen, F.Y. An Intelligent Pattern Recognition Model for Supporting Investment Decisions in Stock Market. *Inf. Sci.* **2016**, *346–347*, 261–274. [CrossRef]
7. Grouard, M.H.; Lévy, S.; Lubochinsky, C. La volatilité boursière: Des constats empiriques aux difficultés d'interprétation. *Banq. Fr.* **2003**, 61–79.
8. Atkins, A.; Niranjan, M.; Gerding, E. Financial News Predicts Stock Market Volatility Better than Close Price. *J. Financ. Data Sci.* **2018**, *4*, 120–137. [CrossRef]
9. Kumar, G.; Jain, S.; Singh, U.P. Stock Market Forecasting Using Computational Intelligence: A Survey. *Arch. Comput. Methods Eng.* **2021**, *28*, 1069–1101. [CrossRef]
10. Mystakidis, A.; Tjortjis, C. Big Data Mining for Smart Cities: Predicting Traffic Congestion Using Classification. In Proceedings of the 2020 11th International Conference on Information, Intelligence, Systems and Applications IISA, Piraeus, Greece, 15–17 July 2020; pp. 1–8. [CrossRef]
11. Francis, B.K.; Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *J. Med. Syst.* **2019**, *43*, 162. [CrossRef]
12. Ben-David, I.; Franzoni, F.; Moussawi, R. Exchange-Traded Funds. *Annu. Rev. Financ. Econ.* **2017**, *9*, 169–189. [CrossRef]
13. Poterba, J.M.; Shoven, J.B. Exchange-Traded Funds: A New Investment Option for Taxable Investors. *Am. Econ. Rev.* **2002**, *92*, 422–427. [CrossRef]
14. Shah, D.; Isah, H.; Zulkernine, F. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *Int. J. Financ. Stud.* **2019**, *7*, 26. [CrossRef]
15. S&P Dow Jones Indices. S&P 500. 2023. Available online: https://www.spglobal.com/spdji/en/indices/equity/sp-500/ (accessed on 29 April 2023).
16. Malik, A.; Tuckfield, B. *Applied Unsupervised Learning with R*; Packt Publishing Ltd.: Birmingham, UK, 2019.
17. Huang, H.; Ding, C.; Luo, D.; Li, T. Simultaneous Tensor Subspace Selection and Clustering: The Equivalence of High Order Svd and k-Means Clustering. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, New York, NY, USA, 24–27 August 2008; pp. 327–335. [CrossRef]
18. Kumbure, M.M.; Lohrmann, C.; Luukka, P.; Porras, J. Machine Learning Techniques and Data for Stock Market Forecasting: A Literature Review. *Expert Syst. Appl.* **2022**, *197*, 116659. [CrossRef]
19. Vargas, M.R.; de Lima, B.S.L.P.; Evsukoff, A.G. Deep Learning for Stock Market Prediction from Financial News Articles. In Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Annecy, France, 26–28 June 2017; pp. 60–65. [CrossRef]

20. Sathya, R.; Abraham, A. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *Int. J. Adv. Res. Artif. Intell. IJARA* **2013**, *2*, 34–73. [CrossRef]
21. Pham, D.T.; Dimov, S.S.; Nguyen, C.D. Selection of K in K-means Clustering. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2005**, *219*, 103–119. [CrossRef]
22. Sharma, M.; Jyoti, Y. A Review of K-mean Algorithm. *Int. J. Eng. Trends Technol. IJETT* **2013**, *4*, 2972–2976.
23. Momeni, M.; Mohseni, M.; Soofi, M. Clustering Stock Market Companies via K-Means Algorithm. *Kuwait Chapter Arab. J. Bus. Manag. Rev.* **2015**, *4*, 1–10. [CrossRef]
24. Ghorbani, A.; Yahyazadehfar, M.; Nabavi Chashmi, S.A. Stock Trading Signal Prediction Using a Combination of K-Means Clustering and Colored Petri Nets (Case Study: Tehran Stock Exchange). *J. Adv. Comput. Res.* **2020**, *11*, 1–17.
25. Fang, Z.; Chiao, C. Research on Prediction and Recommendation of Financial Stocks Based on K-means Clustering Algorithm Optimization. *J. Comput. Methods Sci. Eng.* **2021**, *21*, 1081–1089. [CrossRef]
26. Wijesinghe, G.; Rathnayaka, R. ARIMA and ANN Approach for Forecasting Daily Stock Price Fluctuations of Industries in Colombo Stock Exchange, Sri Lanka. In Proceedings of the 2020 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2–4 December 2020; pp. 1–7. [CrossRef]
27. Mulyaningsih, S.; Heikal, J. K-Means Clustering Using Principal Component Analysis (PCA) Indonesia Multi-Finance Industry Performance Before and During Covid-19. *APMBA Asia Pac. Manag. Bus. Appl.* **2022**, *11*, 131–142.
28. Powell, N.; Foo, S.Y.; Weatherspoon, M. Supervised and Unsupervised Methods for Stock Trend Forecasting. In Proceedings of the 2008 40th Southeastern Symposium on System Theory (SSST), New Orleans, LA, USA, 16–18 March 2008; pp. 203–205. [CrossRef]
29. Jeng, A.M. Using K-Means and PCA in Construction of a Stock Portfolio. 2016. Available online: https://www.diva-portal.org/smash/get/diva2:1079232/FULLTEXT01.pdf (accessed on 13 June 2023).
30. Hargreaves, C.A. An Automated Stock Investment System Using Machine Learning Techniques: An Application in Australia. *Int. J. Math. Comput. Sci.* **2019**, *13*, 199–202.
31. Liu, B.; Qiu, H.; Shen, Y. Establishment and Implementation of Securities Company Customer Classification Model Based on Clustering Analysis and PCA. In Proceedings of the 2012 International Conference on Control Engineering and Communication Technology, Shenyang, China, 7–9 December 2012; pp. 325–329. [CrossRef]
32. State Street Global Advisors. SPY: SPDR S&P 500 ETF Trust. Available online: https://www.ssga.com/us/en/intermediary/etfs/funds/spdr-sp-500-etf-trust-spy (accessed on 17 April 2023).
33. Elbagir, S.; Yang, J. Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. In Proceedings of the International MultiConference of Engineers and Computer Scientists 2019, Hong Kong, China, 13–15 March 2019.
34. Agarwal, A. Sentiment Analysis of Financial News. In Proceedings of the 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, 25–26 September 2020; pp. 312–315. [CrossRef]
35. Heiden, A.; Parpinelli, R.S. Applying LSTM for Stock Price Prediction with Sentiment Analysis. In Proceedings of the Anais Do 15. Congresso Brasileiro de Inteligência Computacional. SBIC, 2021, Joinville, Santa Catarina, Brazil, 3–6 October 2021; pp. 1–8. [CrossRef]
36. Hutto, C.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proc. Int. AAAI Conf. Web Soc. Media* **2014**, *8*, 216–225. [CrossRef]
37. Bonta, V.; Kumaresh, N.; Janardhan, N. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian J. Comput. Sci. Technol.* **2019**, *8*, 1–6. [CrossRef]
38. Ghojogh, B.; Samad, M.N.; Mashhadi, S.A.; Kapoor, T.; Ali, W.; Karray, F.; Crowley, M. Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review. *arXiv* **2019**, arXiv:1905.02845. [CrossRef].
39. Anowar, F.; Sadaoui, S.; Selim, B. Conceptual and Empirical Comparison of Dimensionality Reduction Algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput. Sci. Rev.* **2021**, *40*, 100378. [CrossRef]
40. Scikit-Learn. API Reference. Available online: https://scikit-learn/stable/modules/classes.html (accessed on 29 March 2023).
41. Huang, D.; Schlag, C.; Shaliastovich, I.; Thimme, J. Volatility-of-Volatility Risk. *J. Financ. Quant. Anal.* **2019**, *54*, 2423–2452. [CrossRef]
42. Bhowmik, R.; Wang, S. Stock Market Volatility and Return Analysis: A Systematic Literature Review. *Entropy* **2020**, *22*, 522. [CrossRef]