Cluster Analysis of Firms based on their Financial Characteristics

Sudharshann Devanathan **Sanjan Prakash Kumar**

sd3770@nyu.edu spk363@nyu.edu

New York University

Abstract. The financial world is synonymous with meteoric rises and dramatic falls of stock prices, aggressive strategies of forming mergers and gaining acquisitions and shrewd investments that reap rich bounties. But what misses the eyes of the uninitiated is the significant amount of data analysis that goes on behind the scenes to influence such market-changing decisions. In our project, we attempt to deal with a small piece of the massive jigsaw puzzle that is financial data science. Our project attempts to analyze similarities and correlations between North American corporations and firms by performing clustering based on their financial data such as their stock prices, percentage breakdowns of their balance sheets and income statements. We also attempt to perform intra-cluster analysis and look at the trends between the stock prices of pairs of companies within a cluster and extract correlations between them.

1 Introduction

The aim of our project is twofold : (i) to create clusters of North American companies based on their annual financial performances; (ii) to perform correlation analysis of the stock prices of certain pairs of firms within a cluster.

We perform our clustering with the K-Means algorithm that puts companies that are similar to one another and very different when compared to certain other companies in a cluster of their own. This way, we maximize the intra-cluster similarity (minimize the intra-class variance) and minimize the inter-class similarity. For the sake of raw comparison of clusters, we also perform density-based spatial clustering of applications with noise (DBSCAN) which is more robust to outliers than K-Means.

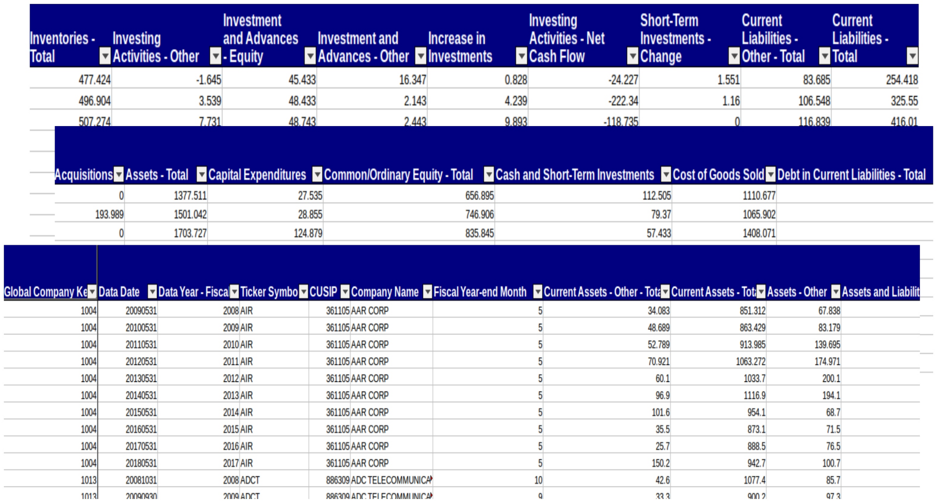
On obtaining our clusters, we examine each of them and generate pairs of companies. We try to extract correlations from the trends of their stock prices over the last two fiscal years. More specifically, we look at the Pearson's coefficients of such pairs.

The end game with the clustering analysis is to provide investors on the lookout with a set of lesser companies that exhibit a growth pattern similar to some of the giants in the market. If they get into the market early and choose to invest in such endeavors, they could potentially reap rich rewards. With the correlation analysis, we hope to produce some pairs of companies that exhibit a healthy positive correlation. This could prove to be of great use to advertisers, marketing managers and firms looking for lucrative acquisitions.

We use *Python* and its `scikit-learn` modules to perform the clustering. We use *ARIMA* to perform the time-series correlation analysis over the stock price trends. We largely stick to the CRISP-DM approach to a data science project that we learned in this course. We elaborate over each phase in the following sections.

2 Business & Data Understanding

We obtain the financial and operating characteristics obtained from the annual financial statements of companies. S&P's Compustat database collects a wealth of year-end financial statistics of over 8,000 North American corporations. The dataset is comprised of 80 financial variables such as investments, assets, liabilities, income statement variables and operational cash flow variables of the last 10 fiscal years. We gained access to the Compustat database through Wharton Research Data Services (WRDS) - a research platform and business intelligence tool developed at University of Pennsylvania. WRDS is being used by over 50,000 corporate, government and academic users and provides access to economical, financial, banking and healthcare-based databases. Below are a few snapshots of what the dataset looks like :



Inventories - Total	Investing Activities - Other	Investment and Advances - Equity	Investment and Advances - Other	Increase in Investments	Investing Activities - Net Cash Flow	Short-Term Investments - Change	Current Liabilities - Other - Total	Current Liabilities - Total
477.424	-1.645	45.433	16.347	0.828	-24.227	1.551	83.685	254.418
496.904	3.539	48.433	2.143	4.239	-222.34	1.16	106.548	325.55
507.274	7.731	48.743	2.443	9.893	-118.735	0	116.839	416.01

Acquisitions	Assets - Total	Capital Expenditures	Common/Ordinary Equity - Total	Cash and Short-Term Investments	Cost of Goods Sold	Debt in Current Liabilities - Total
0	1377.511	27.535	656.895	112.505	1110.677	
193.989	1501.042	28.855	746.906	79.37	1065.902	
0	1703.727	124.879	835.845	57.433	1408.071	

Global Company Ke	Data Date	Data Year - Fiscal	Ticker Symbo	CUSIP	Company Name	Fiscal Year-end Month	Current Assets - Other - Tota	Current Assets - Tota	Assets - Other	Assets and Liabilit
1004	20090531	2008	AIR	361105	AAR CORP	5	34.083	851.312	67.838	
1004	20100531	2009	AIR	361105	AAR CORP	5	48.689	863.429	83.179	
1004	20110531	2010	AIR	361105	AAR CORP	5	52.789	913.985	139.695	
1004	20120531	2011	AIR	361105	AAR CORP	5	70.921	1063.272	174.971	
1004	20130531	2012	AIR	361105	AAR CORP	5	60.1	1033.7	200.1	
1004	20140531	2013	AIR	361105	AAR CORP	5	96.9	1116.9	194.1	
1004	20150531	2014	AIR	361105	AAR CORP	5	101.6	954.1	68.7	
1004	20160531	2015	AIR	361105	AAR CORP	5	35.5	873.1	71.5	
1004	20170531	2016	AIR	361105	AAR CORP	5	25.7	888.5	76.5	
1004	20180531	2017	AIR	361105	AAR CORP	5	150.2	942.7	100.7	
1013	20081031	2008	ADCT	886309	ADC TELECOMMUNICA	10	42.6	1077.4	85.7	
1013	20090930	2009	ADCT	886309	ADC TELECOMMUNICA	9	33.3	900.2	97.3	

Fig. 1: Snapshots of our dataset

We obtain the stock price data of the last two fiscal years from Yahoo Finance.

3 Data Preparation

We realize that the raw data is not ready to be used to build our models. This brings us to the step of data preparation. Here, we pre-process our data, clean it and produce a more streamlined dataset that would be ready for use.

We deal with `NULL` values in the dataset by dropping the rows containing them. We could have used metrics like mean or median to replace these values, but we decide against it as we think that would introduce undesirable noise in the data. We combine pairs of attributes that just represent an upflow-downflow pair to reduce it to a single attribute of net flow. For example, we use `Incoming cash flow` and `Outgoing cash flow` to reduce it to just `Net cash flow`. Another such example - we use `Long term debt issuance` and `Long term debt reduction` to reduce it to just `Total long term debt`. Lastly, we normalize our numerical values using the min-max normalization to make sure that all values are bounded between 0 and 1.

4 Methodology

4.1 Clustering

The idea behind clustering is to find groups/clusters of items that are very similar to each other but are very different to items from other clusters. More specifically, in the case of K-Means clustering, our goal is to reduce the intra-cluster variance or within cluster sum of squares (WCSS) and maximize the inter-cluster variance between 2 items from different clusters or between cluster sum of squares (BCSS).

$$WCSS(K) = \sum_{j=1}^{K} \sum_{x_i \in (cluster) j} \|x_i - \bar{x_j}\|^2 \tag{1}$$

$$BCSS(K) = \sum_{x \in S_i} \|x - \mu_i\|^2 - WCSS(K) \tag{2}$$

Now, in order to decide upon an optimal value of K, we use the elbow method that performs a run of K-Means clustering for different values of K and observes that the WCSS(K) values fall steeply as K increases after which this value stays largely the same. The point at which the WCSS(K) values stop decreasing rapidly is referred to as the 'elbow' of the graph. We find $K = 20$ to be the most optimal value.

However, we realize that using all 80 attributes of the original dataset will prove to be expensive. So we use a wrapper-based method for selecting the 30 most discriminative attributes. We also realize that the wrapper method produces such a subset of attributes by evaluating such subsets over the performance of a predictive model. For this, we first run K-Means clustering with $K = 20$ that makes use of all 80 attributes. We then label each of the companies in the dataset with the cluster to which it belongs. We use these ’pseudo-labels’ as the target variables for the predictive model within the wrapper method. We choose to use a logistic regression classifier with 20 labels whose performance is evaluated over all possible subsets of 30 attributes. The wrapper method finally returns the 30 most meaningful and variant attributes that we will use in further computations.

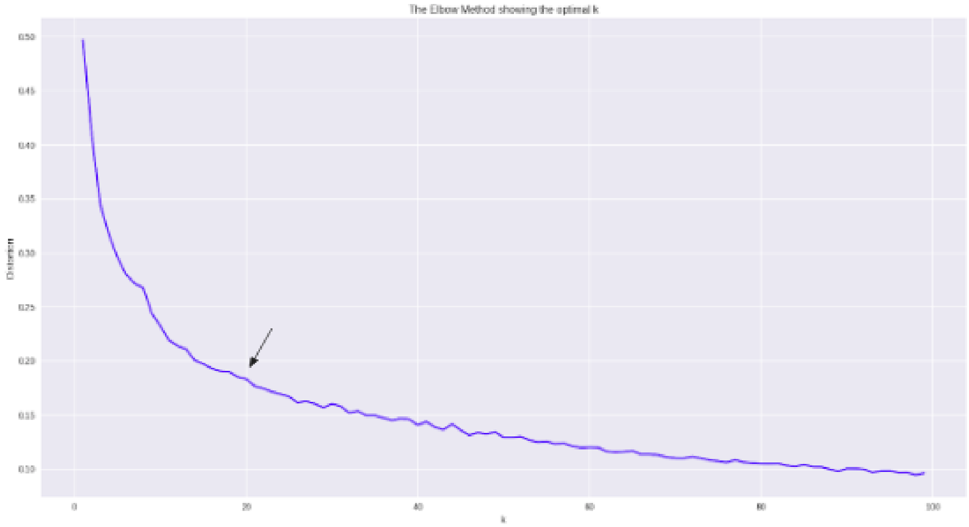


Fig. 2: Elbow method to find optimal K

Now that we have our most representative subset of attributes to work with, we perform a final run of K-Means clustering to obtain a our final 20 clusters.

A thing to note about the K-Means clustering algorithm is that it is sensitive to potential outliers in the data. So to counter this, we also perform DBSCAN which is a much robust clustering algorithm to outliers as it assigns data points to a cluster only if it lies within a certain specified distance `Distance` from a at least a certain number of other data points `min_samples`. Thus, we run DBSCAN with different configurations of `Distance` and `min_samples` and compare the clusters obtained with those obtained from K-Means clustering.

Next, we proceed to perform a correlation analysis of the stock price trends of companies that fall under the same cluster. For the sake of immediacy, we pick a small-sized cluster and hope to scale up our process for all other clusters in the future. In Fig. 6, we can see that corporations that belong to different domains and industries still manage to come together in the same cluster. This is because the clustering that we perform is solely based on the financial performances of these firms irrespective of their domains or specializations.

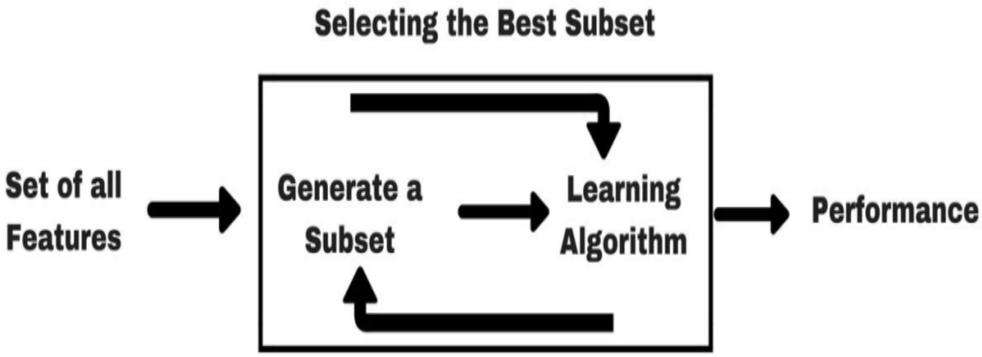


Fig. 3: The workings of a wrapper method.

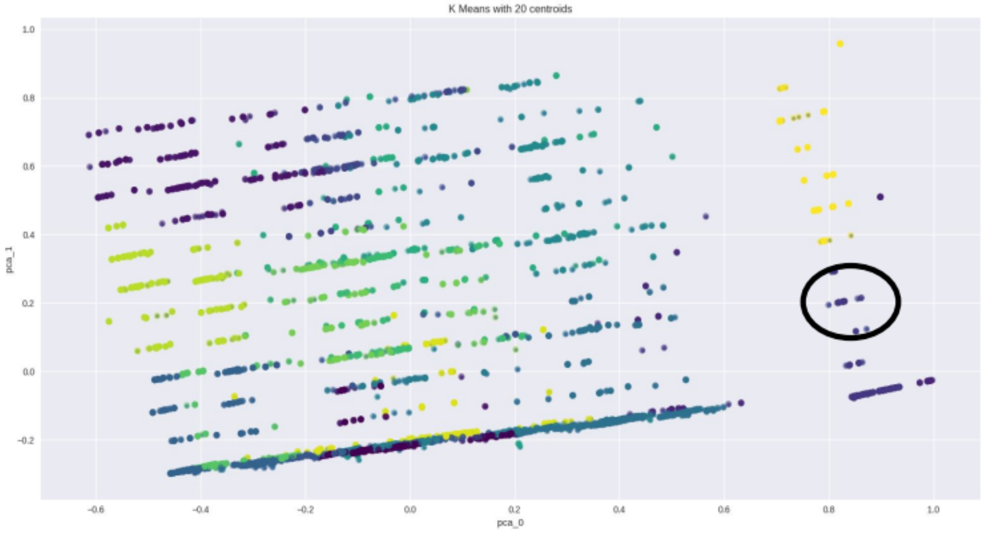


Fig. 5: K-Means clustering with $K = 20$ and over 30 attributes. We perform PCA over the dataset in order to obtain this 2D plot. The circled region is chosen for correlation analysis.

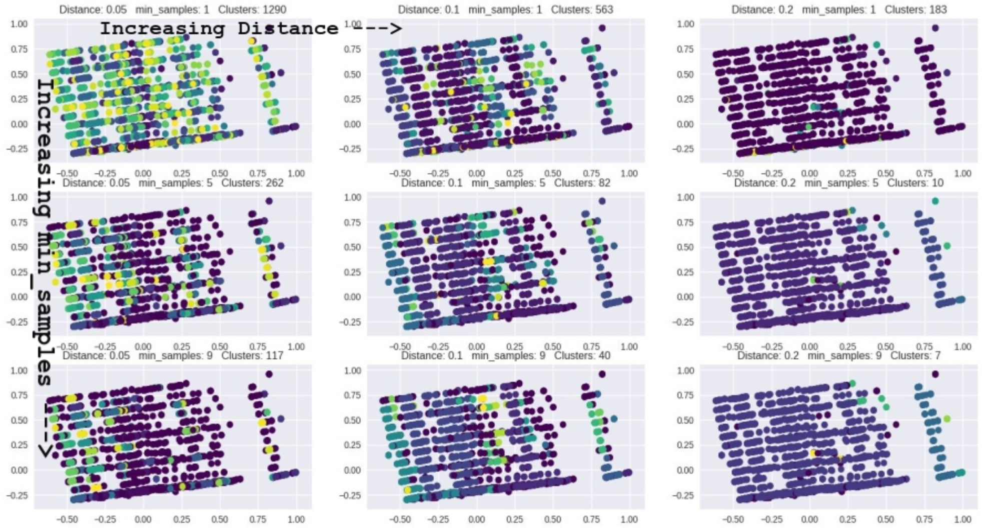


Fig. 4: Clusters obtained from various configurations of DBSCAN. We observe that the configuration of `Distance` = 0.2 and `min_samples` = 5 is the closest clustering we obtain to that generated from our K-Means run.

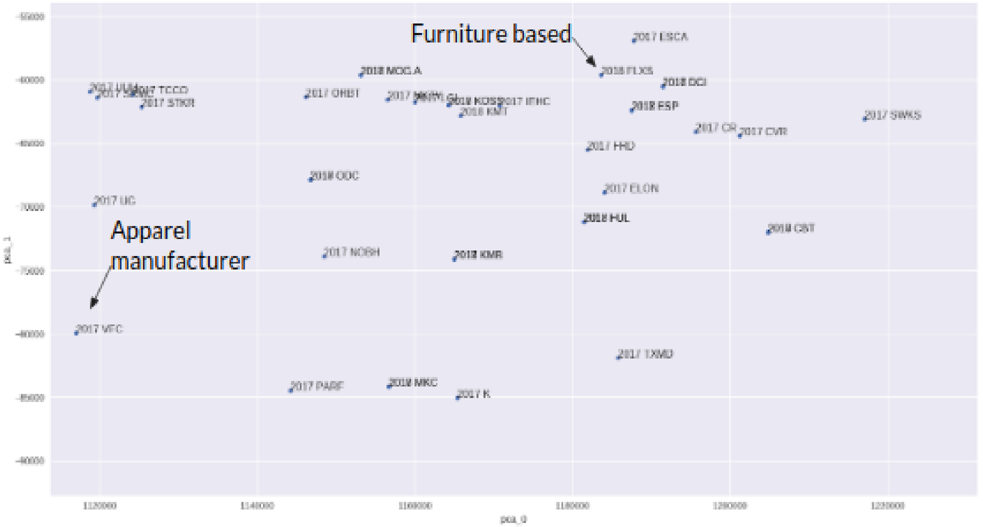


Fig. 6: A zoomed up view of the circled region in Fig. 5. We observe an interesting diversity in the domains of companies that belong to the same cluster.

One specific instance of a pair of companies that we choose to look at is VF Corporation, which is an apparel and footwear company, and Flexsteel Industries, Inc., a furniture company. It is plain to see that these 2 companies come from very diverse niches of the market. If we are able to show a strong correlation between how the stocks of both of them grow, this could prove to be valuable insight to an investor.

4.2 Correlation analysis

Now, we look at the trends of the stock prices of VF Corporation (VFC) and Flexsteel Industries, Inc. (FLXS) over the last two fiscal years. We first proceed by examining the linear relationship between the two trends. We do so by making use of the Pearson product-moment correlation coefficient ρ_{xy}.

$$\rho_{xy} = \frac{\mathcal{E}[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \tag{3}$$

ρ_{xy} will be a value between -1 and 1 with -1 indicating a perfect negative correlation (stock trends move in opposite directions) and 1 indicating a perfect positive correlation (stock trends move in the same direction).

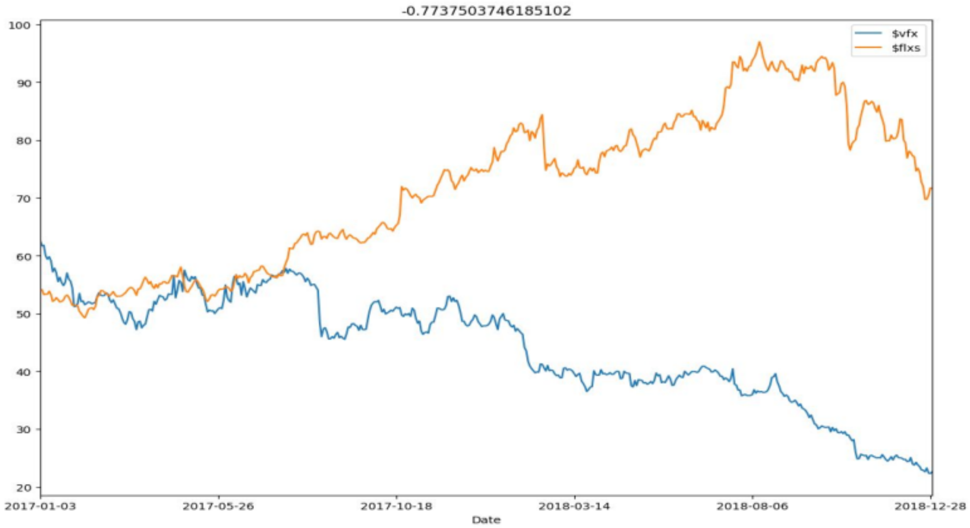


Fig. 7: The Pearson correlation coefficient between the stock trend of VFC and FLXS is -0.77, which is also visible in the trends as they largely grow apart.

However, the Pearson correlation is only reflective of a linear relationship between the two variables. In order to perform a more comprehensive time-series analysis, we compute the cross-correlation between these two trends to find by how much one lags the other.

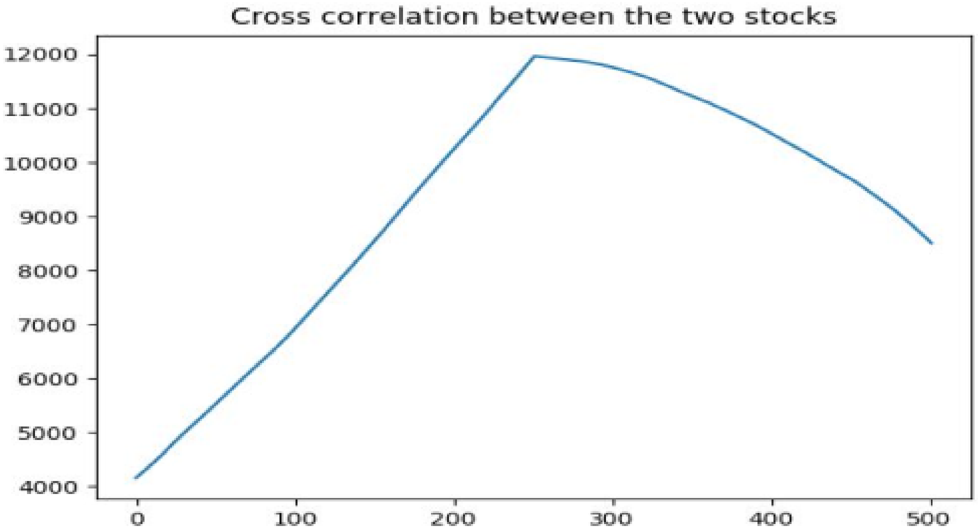


Fig. 8: The cross-correlation between the stock trend of VFC and FLXS. We observe it is at its highest at around the 250^{th} day which implies that the trend of FLXS lags that of VFC by around 250 days.

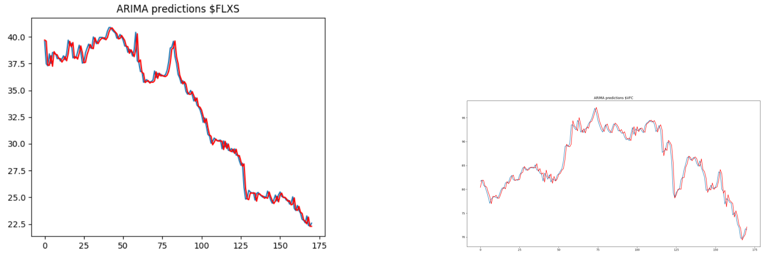
Now, we proceed to attempt forecasting future values of the trends by using *ARIMA*'s auto-regressive models. *ARIMA* stands for Auto Regressive Integrated Moving Average - a class of models that uses its own past values, i.e., its lags and lagged forecast errors as predictors to forecast future values of a time-series.

It takes in 3 parameters : p, d and q. p is the auto-regressive term which refers to the number of lags of a series to be used as predictors. d is the minimum number of differencing needed to be performed in order to make the time-series stationary. Here, we make a rough assumption that our stock data is stationary and do not perform any differencing, thus making $d = 0$. q is the moving average term that refers to the number of lagged forecast errors that should serve as predictors in the *ARIMA* model. Due to our assumption of a stationary time-series, q is also set to be 0. With these assumptions in mind, our model would then be an auto-regressive model only where the forecasted value will only depend on its lags.

Thus, our model will look something like :

$$X_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_p X_{t-p} \tag{4}$$

where X_t is the forecasted value at time t, X_{t-i} is its i^{th} lag from time t, β_i is the coefficient of the i^{th} lag and α is the intercept term. The *ARIMA* model estimates α and the β_is to produce a forecast.



(a) Flexsteel Industries, Inc.

(b) VF Corporation

Fig. 9: Stock price forecasting using *ARIMA* models. (Red line - forecast; Blue line - expected)

5 Conclusion and Future Work

From our findings, we see that the example of VFC and FLXS did not exhibit an appreciable correlation in their stock trends. We intend to keep looking for pairs of seemingly unrelated companies that show a strong correlation. Further, we also intend to roll forward the time-series stock trend of a company by however much it lags the other's trend and then repeat our correlation analysis. We are also yet to fully exercise the power of *ARIMA* models as this is the first time we use them. Thus, as a future resolution for this project, we plan to integrate moving average models into our forecasting model by taking lagged forecast errors into account as possible predictors, which would also mean doing away with our naive assumption of working with a stationary time-series.

6 Acknowledgements

We would like to take the time to thank Prof Bari for giving us the opportunity to dabble in a data science project involving such a large volume of data. We would also like to thank him for making Big Data Science a fun course and we certainly did take away a lot from it and are inspired to continue with this project and even undertake more data science projects in the future. In addition, we wish to thank NYU Stern as it was due to their affiliation that we were able to access the Compustat database made available on WRDS.