# CS701 Final Project Report

Sarah Breckinridge
Middlebury College
sbreckinridge@middlebury.edu

Adisa Majors
Middlebury College
amajors@middlebury.edu

May 16, 2018

**Abstract**

As computer science majors graduating and beginning jobs in the finance industry we aimed to pursue a project that would utilize our coding skills in order to automate commonly used financial analysis techniques. Often, analysts at investment banks are tasked with the role of comparing a given company's value relative to its peers. Such an analysis is used when attempting to achieve a company's holistic price valuation when such a firm is interested in selling itself, merging with another institution or acquiring an asset for instance. The process of selecting comparable companies (i.e. a peer group) is often the most challenging part as it takes time and research especially for firms in niche or unknown industries.

Our project therefore aims to automate the process of conducting a comparable companies analysis by utilizing a clustering algorithm. We implemented the front end functionality in java using eclipse, and did our back end clustering in R, based on particular metrics including size, market capitalization and industry. We therefore constructed a program that takes in a company of interest to a user or investor and outputs competitor companies and their corresponding financial metrics. The output displays the data in a list and graphical form in order for the user to visibly see how their company of interest compares to some of its primary competitors in its general industry group. From this information, an investor could then conclude whether such a public company is a potential good investment since it may be under or over valued compared to its peers.

## 1 Introduction

Our project seeks to automate the widely used comparable company analysis performed by Investment Banking analysts worldwide to value a company in relation to its peers or competitors. The analysis itself is currently done by hand and can often be a time consuming and subjective process. As two soon to be graduates pursuing jobs at different financial institutions, we saw the potential to automate this process. With majors in both economics and computer science, we also found this project to be a suitable opportunity to utilize and expand our knowledge in both fields. Further, within computer science, we both have

been particularly interested in artificial intelligence. While approaching how to perform such an analysis computationally, we found the opportunity to utilize machine learning and clustering techniques. We aimed to automate a comparable companies analysis by using a clustering algorithm in order to determine worthy comparable companies of a company of interest from either the SP 500 or the NASDAQ index. Combining both Java and R, we then present the corresponding data of the comparable companies to a user who can then be the judge of the company's relative value.

Our project therefore contributes to the ongoing study of clustering algorithms and their applications. More specifically, our project is very much tied to the finance industry and could have positive implications on increasing time efficiency and work product quality.

The next section of this paper will discuss the problem statement that drove our project. We then follow by discussing related work to our project that helped to guide our progress and educated us about the uses of clustering. The fourth section discusses the methods used to create the Analyzer. The Results section then provides an overview of our final product which is then talked about in depth in the following concluding Discussion section.

## 2 Problem Statement

Our project aims to automate the process of creating a comparable company analysis in order to better understand whether a company is valued accurately. Investment Banks are hired as advisers to companies when they make substantial and significant financial changes. These include going public, selling bonds, selling their company itself, or acquiring another entity. In all of such actions, an Investment Bank advisory team must reach conclusions about the value of the company in question in order to correctly manage the respective deal. There are many techniques used to value a company, however, the most basic process is conducting a comparable company analysis. This analysis essentially compares the current stock price of the given company to that of its competitors. In addition, since stock price is not always the most effective indicator of a true company value, one must also examine ratios from which projected prices can be derived. With these different indicators of value, one then generates a "football-field" graph which we will explain later on in the paper. However, the most time consuming part of this process is selecting the competitors themselves. Overall, one must select the peer group by hand based on extensive industry research and a developed understanding of the different companies' businesses. Because this process is time consuming, subjective and conducted so often in the financial industry, we decided to create a program that will select the peer group and produce the full analysis.

In order to first gain access to the data necessary for the analysis, we extracted data from Bloomberg for two stock indexes. Our data is therefore from one particular instance in time, so we had to make the assumption that this data was a good representation despite the fact that the stock market could

have been experiencing unusual fluctuations or shocks on that particular day. As we will explain later, we would strive in future work to have a real time connection with Bloomberg in order to always display current data.

Overall, we initially had plans to simply output comparable companies for any stock in the SP 500 as our prototype. However, for our final deliverable, we anticipated completing our comparable analysis in full, generating a nicely formatted graph, applying our code to a larger index, joining together the different components used and creating a simple web app. From the halfway point we also set the goals to automate our labeling of industries and improve how we select comparable companies from within clusters. Overall, we accomplished these tasks, and looking to the future we can imagine enlarging the scale of the project. Since this could be a revolutionary tool for the finance industry, we imagine applying our program to different data sets, establishing a real time connection with Bloomberg data and also improving upon our user experience.

# 3 Related Work

Investment Banking teams advise companies on their financial transactions and are consistently tasked with determining the true value of such companies. One common way to observe the performance of a public company and its value, is to look at its stock price as well as other valuation ratios and metrics. In order to astutely consider whether such numbers are "good" or whether the stock may be a wise investment, investment bankers compare metrics to those of comparable companies. This technique is commonly called a comparable analysis and is performed by investment bankers along with other more complicated quantitative modeling to determine the value of a company. While the analysis itself is very straightforward and requires one to simply compare numbers, one of the more challenging tasks is to actually find the companies to use in one's comparison.

Machine Learning has been used to forecast stock market movements and predict future stock prices. This has been done by looking at news, financial statements and press releases of different companies. [2] [5] However, to our knowledge, machine learning has yet to be used in order to perform a full comparable analysis commonly used by investors and bankers in the finance industry. We hope to perform this task by using machine learning techniques namely a clustering algorithm. A substantial amount of research has been performed in regards to clustering and classification algorithms as well. There are many different clustering techniques most of which are based on finding a partition that optimizes a particular objective function. [3] We also familiarized ourselves with the clustering process in depth. Reading Karypis, et al., we developed a further understanding of hierarchical clustering and the process' significant advantages compared to those of partitioning techniques. [1] Further, we understand from the basis of the Karypis et al. article that euclidean distance when being used by a hierarchical clustering algorithm does have its faults in regards to handling and identifying clusters for complex data sets. This served as a consideration when we determine the effectiveness our industry labels' linearity and overall

clustering technique.

Clustering has been used in regards to Finance as well. Nanda clustered data from the Indian Stock Market in order to create investment portfolio groups that include a diverse group of stocks to best minimize risk. [4] In many ways Nanda's work is the opposite of what we hope to achieve in finding firms that are similar to one another. Further, Sim et al. creates a 3D subspace clustering algorithm to find undervalued stocks. [6] Sim et al. therefore has the same goal as us in regards to finding stock market anomalies, however they go about it in a different way. We, on the other hand, plan to find value in the stock market through a formalized method commonly produced in financial firms.

Further, Many investment sites that exhibit stock data such as NASDAQ.com lists the comparable companies to a given stock. However, the comparable company lists produced by these sites often lack practicality and precision. While we do not know the algorithms that these sites use, we hope to replicate and improve upon the work done on these sites while familiarizing ourselves with clustering techniques and financial analysis.

# 4 Methods

Our program goes through numerous specific steps in order to generate the full comparable company analysis for a user.

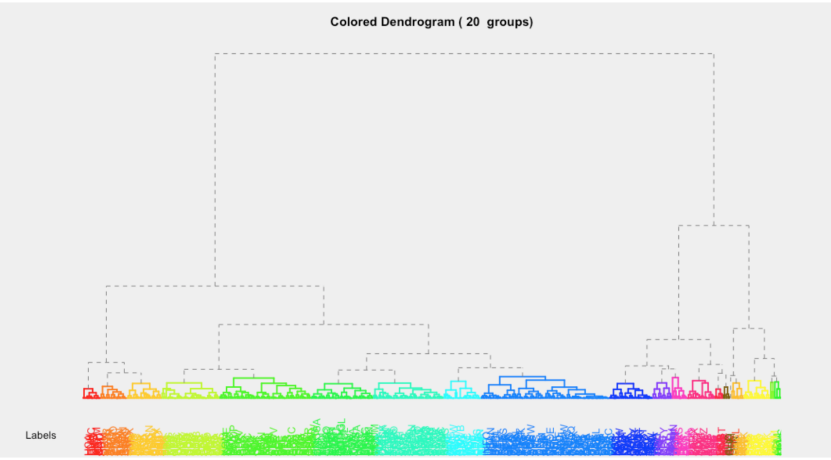
First, in order to have a comprehensive set of company data, we extracted a list of companies, their corresponding metrics and applicable data from the Bloomberg terminal. We chose to use Bloomberg to retrieve our data because it is the most reliable source we had access to and is commonly used at financial institutions as the data source of choice. The terminal has access to real time immediate and accurate stock data. More specifically, we took data from companies in two indexes, the SP 500, which contains the roughly 500 largest companies in the stock market, and the NASDAQ, which includes over three thousand commonly traded equities listed on the stock exchange. In java, we then created an object for each company and stored the respective metrics in each object for later use.

We then used R in RStudio to cluster our data. We decided to cluster based on three particular metrics. The first and most important characteristic of a company is its general industry group or sub-industry group. Our extracted data included industry labels strings for each company however in order to perform hierarchical clustering using euclidean distance, we needed to assign numeric values to each industry. While initially we labeled the industries numerically by hand, we then wrote an excel script to automate this process and easily cluster companies from any index or data set imported. The excel script includes a lengthy if statement that accounts for every potential industry group in Bloomberg and assigns a value from 1 to 5000. Industries that are rather similar to each other are close together on this 2-dimensional number line. For instance, companies in the food industry are particularly close to beverage companies. Furthermore, since these companies exist in the broader consumer and

retail industry, they are relatively closer on the 1 to 5000 number scale to consumer brands and retail chains than they are to industrial or energy companies for example.

We also cluster by the log of two other metrics, market capitalization and employee count. We used the log of these values in order to normalize the data and remove potential clustering problems from data outliers. Prior to logging this data, we did find that our euclidean distance hierarchical technique struggled to generate accurate clusters. This made sense given the warnings from Karypis et al. [1] However, the log normalized our data and conformed it to prevent these types of problems. Market capitalization more specifically is a company's stock price multiplied by the number of shares outstanding. This measurement helps to gauge the size of a given company and its generic worth that the stock market subscribes. The number of employees, our third measurement, serves to represent the capacity and size of the company. Overall, when finding comparable companies we want the companies to have similar metrics in order to create the best peer group. Competitors should be in the same general industry group and should also have relatively the same size and capacity.

Figure 1: Clustering Example with 20 Clusters




Using specific packages in R including the cluster, hclust and flashClust libraries, we then cluster the companies using a hierarchical technique. The companies are clustered starting as one full and funneling down to each grouping containing one company each. This clustering method creates a tree as seen in Figure 1. We then were able to specify how many clusters we wanted. An example of this process can be seen in Figure 1, where each color at the base of the tree diagram represents a different single cluster out of the twenty specified to generate. For a smaller index like the SP 500 we chose 25 clusters which was specific enough to account for niche industries but also just enough so that each cluster has at least 5 companies within it. For the NASDAQ index we

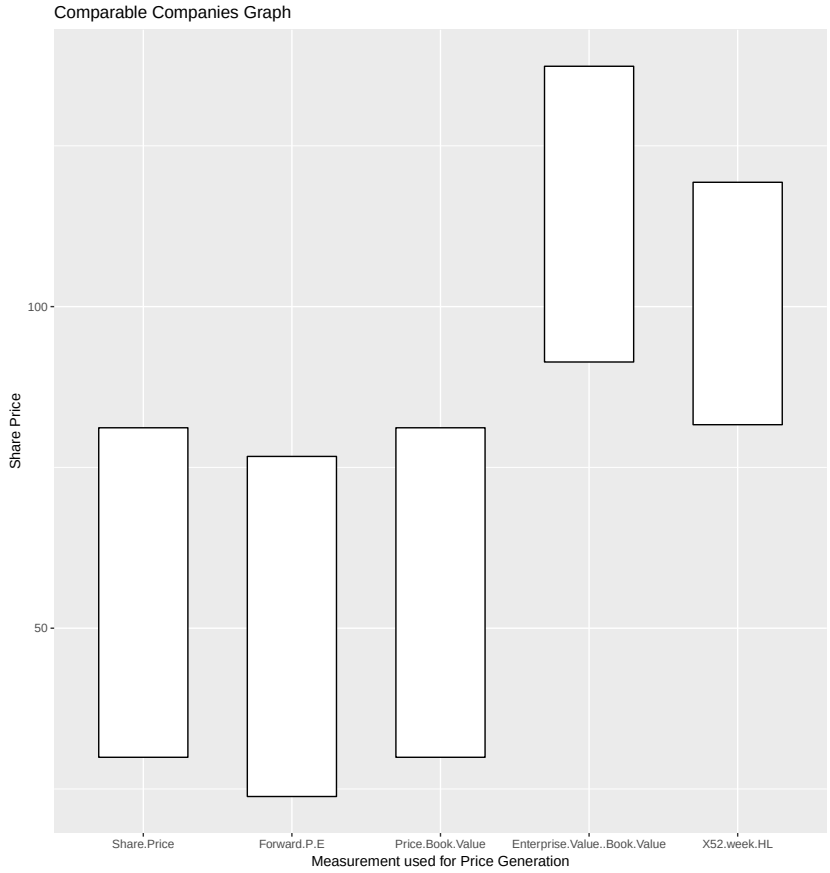
generate 40 clusters given the larger nature of the index and inclusion of many more industry types. Again we made sure that each cluster had at least five companies within it.

Once the clusters were established we then exported them back to eclipse to work with using Java. The specific cluster a company is in is recorded and stored as part of the company object. Therefore, the process of extracting company data, labeling industry groups and then clustering the data concludes our back end process which we therefore only have to complete once.

Our program then addresses the user, requesting first which of the two indexes the user is interested in and then requesting the specific company the user wants to perform a comparable analysis on. Once we take in the user's request we then find the respective company's object cluster. From within that cluster we then determine our five peers. While we initially chose these five randomly within the cluster we developed a more accurate process in our final stages of the project. We first choose two companies at random while three are chosen based on closest euclidean distance to the company of interest based solely on market capitalization. We believe that this approached granted the most accurate competitor while also providing variety which mimics common analyses in the financial industry.

Once these companies are selected from the target company's cluster, we then extract the metrics of interest from each of the five comparable company's objects. More specifically we extract prices that we have derived from the following ratios: Forward Price/Earnings, LTM EV/EBITDA, Forward Enterprise Value/EBITDA, Price/Book Value and Enterprise Value/Book Value. All of these ratios are commonly used in the financial industry for this analysis and can vary based on the industry of focus. When we first received the data from Bloomberg which included the ratios mentioned, we were able to derive prices. For instance, with a forward price/earnings ratio, we could multiply the ratio by the particular company's earnings per share to then be left with what the numerator(price) equals per share. We performed the same process with each metric to get at the derived price from each metric. However, in some cases there is missing data. This is often because the companies are obscure, have negative ratios, or used different valuation metrics depending on its industry sector. For instance, financial institutions often use Price/Book Value to value themselves instead of Enterprise Value/EBITDA. Our program accounts for this by labeling the two groups of companies, those with only the Price/Book Value, and those with Enterprise Value/EBITDA. If the company has both, we grouped it with those that only had Enterprise Value/EBITDA, because this ratio is more commonly used. Thus we can look at which label is more common within the cluster, and use the appropriate metric when performing our analysis.

Figure 2: J.P. Morgan Chase  Co. Example Comparable Companies Graph




Once we have each derived price, we take the min and max of the group of five to then export in the form of a comma separated value table back to R. In R we then represent the min-max data in the form of a bar graph with the bounds of the bars being the min and max themselves. The visualization is created using the reshape2 and ggplot2 library packages in R. An example of the graph produced can be seen in Figure 2. The y axis displays the price per share while each bar represents the price range derived from the corresponding x-axis ratio label for the companies comparables. The final bar that represents the 52 week high/low however displays data exclusively for the company of interest. This measurement shows the range that the company's stock has experienced in the most recent 52 weeks of trading. Considering in this example case that the J.P. Morgan stock was trading around 110 dollars, one may say that according to the first three bars, the stock may be over valued. While according to the last two bars, the price of the stock seems on target. The graph as a whole is then saved as a PDF for a user. Using the graph a user can then see how the current price of the company of interest compares to the predicted ranges from the variety of metrics.

# 5 Results

Upon completion of our program, we tested its effectiveness on numerous companies across a variety of industries. As a whole, our program generates comparables that are accurate and in line with what analysts would have chosen by hand. The program further displays the analysis in a clean and efficient way. Further, the final output mimics those generated by analysts at investment firms. We compared our results to those of NASDAQ.com and found that our output frequently produced more accurate and appropriate comparables while also not producing errors like that of the website. Our program also produces a visual output that current online platforms do not provide.

Figure 3 displays an example of the output comparable companies generated for J.P. Morgan Chase  Co. Overall we see that other banks with similar functionality appear as comparables. J.P. Morgan is not only an investment bank but also offers credit cards, wealth management services and retail banking. In that regard, Citi Bank and Bank of America would be two of the best comparables and they both appear in our output. We provide two other example in Figure 4 and Figure 5 which display a peer group for Facebook and Amgen respectively. Again, in both cases, we see a variable but accurate comparable company set. However, in the real non graphical output, one would see something more like Figure 6 which not only lists the comparable companies but also shows their respective metrics.

Figure 3: Comparables Generated for J.P. Morgan Chase  Co.

```
JPM: JPMorgan Chase & Co
Company 1: Bank of America Corp
Company 2: Wells Fargo & Co
Company 3: Citigroup Inc
Company 4: Cintas Corp
Company 5: Bank of New York Mellon Corp/The
```

Figure 4: Comparables Generated for Facebook

```
FB: Facebook Inc
Company 1: MSFT/Microsoft Corp
Company 2: STX/Seagate Technology PLC
Company 3: VZ/Verizon Communications Inc
Company 4: GOOG/Alphabet Inc
Company 5: AMZN/Amazon.com Inc
```

Figure 5: Comparables Generated for Amgen

```
AMGN: Amgen Inc
Company 1: LLY/Eli Lilly & Co
Company 2: BMY/Bristol-Myers Squibb Co
Company 3: DHR/Danaher Corp
Company 4: CVS/CVS Health Corp
Company 5: MDT/Medtronic PLC
```

Figure 6: Complete Output for Morgan Stanley

```
MS: Morgan Stanley
Company 1: Goldman Sachs Group Inc/The
Company 2: American Express Co
Company 3: US Bancorp
Company 4: Willis Towers Watson PLC
Company 5: Bank of New York Mellon Corp/The

Company 1: GS/Goldman Sachs Group Inc/The
Price: 237.0
Market cap: 9.2979722448E10
P/E ratio: 225.5298483
52 week high: 275.29
52 week low: 209.66
bv: 237.0
evBV: 492.186185

Company 2: AXP/American Express Co
Price: 99.74
Market cap: 8.5812526327E10
P/E ratio: 88.76220763
52 week high: 102.95
52 week low: 75.98
bv: 99.73999786
evBV: 126.7831701

Company 3: USB/US Bancorp
Price: 50.33
Market cap: 8.2665004603E10
P/E ratio: 43.74746834
52 week high: 58.5
52 week low: 49.035
bv: 50.33000183
evBV: null
```

Lastly, we created a web-app to allow a user to request a comparable company analysis on a particular firm either in the SP 500 or the NASDAQ index. This app is unfinished and could not interact with R, and Java yet, however it has the simple user interface elements that we wanted. We built this application using HTML to create interactive elements and JavaScript to connect it to other languages.

# 6 Discussion

Over the course of the project we learned a great deal about the different methods of clustering as we progressed to eventually choose a hierarchical method that utilizes euclidean distance in order to group the data. We both became significantly more comfortable working in R in order to cluster our data and present it graphically to a user. We were also able to refresh and supplement our financial skill set. In order to extract our data in the first place we had to utilize the Bloomberg terminal and subsequently learn the excel shortcuts to pull data into a spreadsheet format. When determining the correct ratios to then present to the user, we performed extensive research and also reverted back to work we did in the finance industry over the past few summers.

While we believe our program does a complete and thorough job of determining comparable companies and presenting a comparable analysis to a user,

we do recognize that there are a few weaknesses in our implementation. For one, we would like to address our user experience. Currently, we have elements across different platforms that would require the user to follow roughly six different steps in order to generate a complete output. We attempted to tackle this issue with numerous tactics but none improved our user experience because they require additional downloads and software packages. We have begun to develop a simple web-app that will communicate with Bloomberg, Excel, Java and R and most importantly the user. Although passing information through JavaScript may pose security risks, company information is public domain and security may not be a pressing concern.

In terms of our data, we see room to improve our extraction of Bloomberg data by establishing a real time connection. Currently our data is from a particular instance in time. Therefore our data is prone to being effected by the market shocks of the particular day we took the data from. Ideally, our program should have a real time connection with Bloomberg and have the ability to retrieve live data. Getting the data every time the user inputs a company would slow down our app, so ideally, we would collect the information from the previous day's stock market close. There are some issues with this as it requires the costly connection to a Bloomberg terminal. Perhaps we could look into a more economical data source. Further, Bloomberg does have missing data for numerous companies. We currently handle this by removing companies with more than one missing value. While missing data is likely not a problem exclusive to Bloomberg, we could seek out other databases to supplement our excel spreadsheet. Removing so many companies could prove problematic as one runs the risk of losing comparables that may be best for particular companies.

# 7 Acknowledgments

In conclusion, we would like to thank Prof. Grant for his assistance and advice throughout the semester and especially in regards to using built-in clustering algorithm provided by R. We would also like to acknowledge Graham Booth who provided feedback on our project midway through the semester and helped us make the decision to use a clustering algorithm over machine learning.

# References

[1] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, Aug 1999.

[2] X. Li, H. Xie, R. Wang, Y. Cai, J. Cao, F. Wang, H. Min, and X. Deng. Empirical analysis: stock market prediction via extreme learning machine. *Neural Computing and Applications*, 27(1):67–78, Jan 2016.

[3] P. Michaud. Clustering techniques. *Future Generation Computer Systems*, 13(2):135 – 147, 1997. Data Mining.

[4] S. Nanda, B. Mahanty, and M. Tiwari. Clustering indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12):8793 – 8798, 2010.

[5] J. Patel, S. Shah, P. Thakkar, and K. Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1):259 – 268, 2015.

[6] K. Sim, V. Gopalkrishnan, C. Phua, and G. Cong. 3d subspace clustering for value investing. *IEEE Intelligent Systems*, 29(2):52–59, Mar 2014.