K-Means Cluster Analysis of Stock Financials

Kian Ankerson

Abstract

In this paper, cluster analysis is explored as a potential strategy to find a smaller subset of individual stocks to invest, or conduct further research into. Companies' financial results alone are used as the underlying data because companies are required to report earnings every quarter. As a result of the analysis, 3,372 stocks listed on the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotations (NASDAQ) are grouped into 15 and 4 clusters. Within these clusters the returns of the stocks are examined to determine if the clusters formed have a meaningful relationship with the returns over the period the earnings reports covered.

Keywords: Stock Financials, K-Means Clustering, Investing

1 Introduction

In this paper K-Means Cluster Analysis is performed on financial data of stocks. Investing is an activity where one buys a financial asset at one point in time, with the hopes that it provides them a return at or on later date(s). Educated stock investing, in particular, involves researching individual companies, and the corresponding stocks, to determine if the company is worthy of an investment.

When it comes to investing in individual stocks, it can a challenge to find the right companies to invest in. A major step of the investing research process involves finding a company, and looking at the past financial results. Doing this for even just one company can take a significant amount of time, as companies release earnings reports every quarter of the year, with a multitude of different numbers to dig through. To make things even harder, at a given point in time, there can be over 6,300 stocks to choose from [1], and it is not possible for an individual to do research on all of these companies. This is where K-Means Clustering is looked upon to help research these companies.

2 Literature Review

In their work [2], Da Costa, Cunha, and Da Silva perform cluster analysis on stocks according to risk and return criterion. The variables used for their investigation include return, risk, several accounting ratios, and dividend yield.

The work used hierarchical clustering, and included 476 different stocks. As part of their hierarchical clustering work, the researchers arbitrarily chose to cut the clusters at a level of 20, which left them with 10 clusters. The analysis also reports that an investor should also consider the country in which a stock is located in to better their investing choices. They concluded that an investor could use hierarchical clustering to balance their portfolios in terms of profit and risk management. As part of the work, they reexamined the best performing clusters for a year after the data was grouped on, and found the clusters continued to profit or minimize losses during that time.

In another work, cluster analysis was performed various financial indicators for companies in the Borsa Istanbull 100 Index (BIST) [3]. The data used in this study included variables for price to earnings ratio, market value to book value ratio, dividend yield, return on assets, return on equity, return, and return on average and risk. Many of these variables, specifically the ratios, are related to the underlying price of the stock, and change on a daily basis according to how the price of the security moves. For the analysis, both hierarchical and non-hierarchical clustering was performed. Out of 100 companies examined, 12 different clusters were created, and 3 were found to be preferred, using the returns from the latest year when the study was done, and when looking at returns over a 3 year period, 4 clusters could be preferred, with those clusters being the same clusters as the 1 year period with one cluster added.

In both of these researches, cluster analysis was performed to analyze and select the best stocks to invest in. Both sets of researchers use hierarchical clustering methods for analysis, but used different data sets with differing variables. One thing that was not thoroughly researched in these papers, was clustering specifically based on the underlying companies' financial results. Numbers reported in the quarterly earnings report by companies are one of, if not the, most important information individuals can find in doing research, and neither study utilized the available data fully. In the two different data sources, only a very small subset of the financial reports data was used to cluster the different companies. There can be upwards of 50 different available numbers reported by companies, which could contain valuable information in differentiating a good investment from a poor one. In addition, the data sets used in each of the studies did not contain the large magnitude of data available. The first study mentioned contained only 476 stocks, and the second study only used data from 100 companies. As mentioned previously, there are well over 6,000 different companies an investor can choose from.

3 Research Question

The specific question this research aims to answer is:
Can stocks be categorized into different meaningful groups using data from annual financial statements in which certain clusters are created that clearly produce better performance than others?

4 Data and Data Sources

The data used in the research was gathered from Yahoo Finance[4] and contains numeric variables that can be found on companies quarterly earnings reports. This data set consisted of 26 numeric variables which are common accounting figures that are required to be reported by the Securities and Exchange Commission. The entire list of variables includes cash, total current assets, good will, total assets, accounts payable, long term debt, total liabilities, total current liabilities, retained earnings, total stockholder's equity, net income, total cash flows from investing activities, capital expenditures, dividends pad, total cash from financing activities, change in cash, repurchase of stock, issuance of stock, total revenue, cost of revenue, gross profit, research and development, total operating expenses, operating income, interest expense and earnings before interest and taxes (ebit). All of these variables can be found on one of the three financial statements, the income statement, the balance sheet, or the statement of cash flows. These variables are very common to individuals with an accounting background, and generally understood by individuals familiar with stock and financial research. A great, detailed explanation for each of the individual financial metrics can be found on in this Investopedia article by Chris B. Murphy[5]. Data was collected through the use of the yfinance R package and API calls. Data was collected over the course of a 3 year period, on an annual basis for 2019 and 2021.

Data on the prices of the stocks investigated comes from eoddata[6]. The prices. Eoddata allows for easy bulk download of historical prices for entire stock exchanges at once, in a comma separated value format. The data used from this source consisted of the adjusted daily close prices for 2019-12-31 and for 2022-01-03. The adjusted daily close price represents the price the stock closed the day at, after adjusting for any stock splits that may have changed the amount the stock was valued at, without affecting any of the company's fundamentals. This price data allowed for calculating the total return each stock produced during the researched period.

5 Methodology

The first step in performing cluster analysis on this data was cleaning the data. The data originally contained more columns and more observations than the ending 3,372 observations and 26 variable columns. The raw data set however, had many rows with missing values for certain columns. There was only a very tiny proportion of data that contained no missing data for any column, so a balance was struck to remove as few columns from the data while keeping as many observations as possible that contained all of the variables. This dramatically reduced the number of observations, but was a necessary step to preserve as many metrics from the financial results as possible.

The next step in preparing the data was converting the data to a percent change representation from the starting 2019 annual report to the 2021 report.

Here is the mathematical formula that was applied to every column, including the to find this number:

$$(X_{2022} - X_{2019})/X_{2019} \tag{1}$$

Where X_{2022} is the value from the 2022 data, and X_{2019} is the value from the 2019 data. This same formula was used to find the change in price of the stocks over the period, but this number was not used as a variable for clustering. This value was used only as a way of examining the individual clusters after performing K-Means.

An extra cleaning process done during this step including dealing with invalid values. When a percent change resulted in an NaN value as a result of 0 / 0, then percent change was set to 0. When the percent change resulted in an infinite value as a result of x_{2022} / 0, and $x_{2022} \neq 0$ the result was set to the non-infinite maximum for the column. This data was then scaled so each column had the same importance for distance calculation.

For the clustering method, K-Means was chosen as the method for determining clusters. K-Means Clustering is a technique where the desired number of clusters are predetermined, and to start random observations are chosen as the centers for the cluster. From there, points are repeatedly assigned to the closest cluster center, where the cluster center is the average coordinate for all points in the clusters. The method to determine the distance between points and cluster centers was the Euclidean Distance Formula. This formula works because the data has already been scaled, so columns are not over-represented in the resulting distance. The formula for finding the distance between a point and a cluster center using Euclidean distance is a follows:

$$\sqrt{\sum_{i=1}^{p}(x_{point_i} - x_{center_i})^2} \tag{2}$$

where p is the number of variables, x_{point_i} is variable for the point being assigned to a cluster, and x_{center_i} is the value for corresponding variable for the center of the cluster whose distance is being calculated.

In order to find the best number of clusters, the NbClust package from R was used. This package accepts a data set, a clustering method, and a distance calculation method, and using different methods of finding the best number of clusters, returns the the number of clusters with the most number of methods supporting it.

Once a best number of clusters was determined, K-Means clustering was performed on the data using the suggested best number of clusters. After assigning the stocks to a cluster, the returns of each cluster were calculated. The return for a single stock, the maximum return for a single stock, the median return, the average return, the percent of stocks in the cluster who had a positive return, and the percent of stocks in the cluster who outperformed the SP500 Index was calculated using the stock prices data not used in determining cluster assignments or the best number of clusters.

After clustering was done, the sizes of the clusters and the calculated returns data was examined to see if clustering had resulted in any meaningful groupings. Additionally, a visualization according to the top 2 principal components of the data set was done to see what the points and corresponding cluster shapes looked like. This easily visualized stocks that were put in a cluster by itself, identifying it as an outlier of the data set. Because clustering is very sensitive to outliers, when an outlier was found it was removed from the data, the best number of clusters was recalculated and clustering of the data was performed again. This was an iterative process, which was repeated numerous times, with the goal of getting rid of outliers and producing relatively meaningfully sized clusters. This iterative process reduced the final number of observations being clustered to the 3,372 stocks.

6 Results

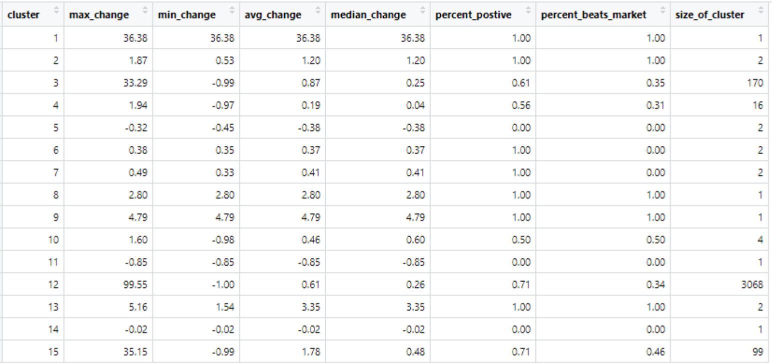
6.1 Descriptive Statistics

Table 1: Descriptive Statistics of Variables

variable	mean	std_dev
accountsPayable	-0.0003012	1.1047278
capitalExpenditures	0.0034075	1.1061408
cash	-0.0003338	1.0992523
changeInCash	0.0030492	1.1060622
costOfRevenue	-0.0184869	0.901348
dividendsPaid	0.0004762	1.0846882
ebit	0.0158912	0.4286334
goodWill	-0.0734698	0.8238097
grossProfit	0.0142299	0.0468187
interestExpense	-0.0095828	1.0218628
issuanceOfStock	-0.0158131	0.0041444
longTermDebt	-0.0378187	0.9325828
netIncome	-0.0014821	0.778935
operatingIncome	0.0161162	0.4291329
repurchaseOfStock	0.0051125	1.1058532
researchDevelopment	-0.068833	0.5385778
retainedEarnings	-0.0308564	0.6754392
totalAssets	-0.0199384	0.0256679
totalCashflowsFromInvestingActivities	-0.0153493	0.7699214
totalCashFromFinancingActivities	0.0096313	0.1959818
totalCurrentAssets	-0.0451573	0.4453539
totalCurrentLiabilities	-0.0249216	0.1478787
totalLiab	-0.0292974	0.2606193
totalOperatingExpenses	-0.07213	0.7091398
totalRevenue	-0.0238747	0.4959572
totalStockholderEquity	0.0174227	0.6317934

In the above table, the variables used in clustering are displayed. The variables were calculated by taking the percent change from the 2019 data to the 2021 data, and then normalized so each variable could not artificially have more importance in determining distance because it's values were normally higher than another's.

6.2 Main Results



cluster	max_change	min_change	avg_change	median_change	percent_postive	percent_beats_market	size_of_cluster
1	36.38	36.38	36.38	36.38	1.00	1.00	1
2	1.87	0.53	1.20	1.20	1.00	1.00	2
3	33.29	-0.99	0.87	0.25	0.61	0.35	170
4	1.94	-0.97	0.19	0.04	0.56	0.31	16
5	-0.32	-0.45	-0.38	-0.38	0.00	0.00	2
6	0.38	0.35	0.37	0.37	1.00	0.00	2
7	0.49	0.33	0.41	0.41	1.00	0.00	2
8	2.80	2.80	2.80	2.80	1.00	1.00	1
9	4.79	4.79	4.79	4.79	1.00	1.00	1
10	1.60	-0.98	0.46	0.60	0.50	0.50	4
11	-0.85	-0.85	-0.85	-0.85	0.00	0.00	1
12	99.55	-1.00	0.61	0.26	0.71	0.34	3068
13	5.16	1.54	3.35	3.35	1.00	1.00	2
14	-0.02	-0.02	-0.02	-0.02	0.00	0.00	1
15	35.15	-0.99	1.78	0.48	0.71	0.46	99

Figure 1: K = 15 Clustering Results

In this figure 1 are the cluster, the returns statistics for the cluster, and the size of the clusters for K = 15 clusters.

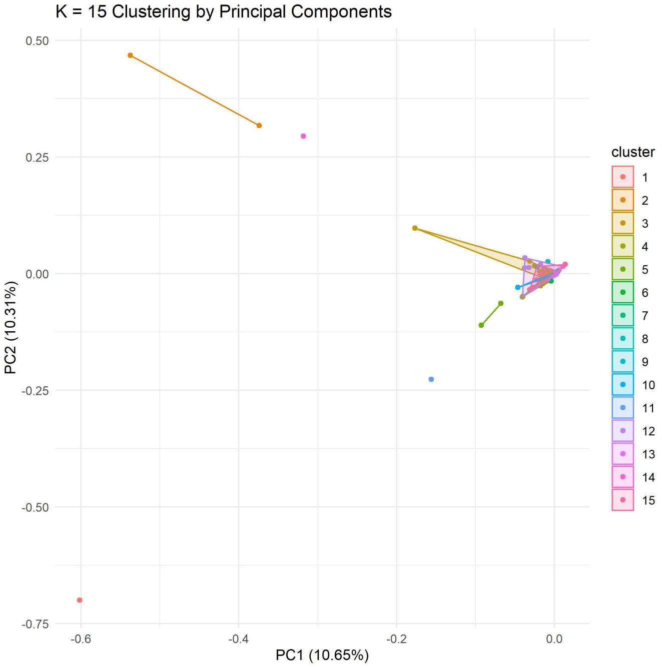
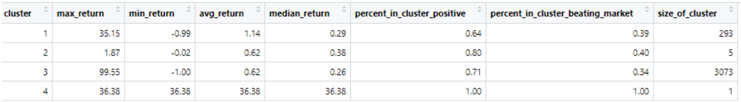


Figure 2: K = 15 Cluster Visualization by PC1 and PC2

This figure 2 shows the cluttered aspect and undesirable quality of the cluster shapes according to the top principal components.



cluster	max_return	min_return	avg_return	median_return	percent_in_cluster_positive	percent_in_cluster_beating_market	size_of_cluster
1	35.15	-0.99	1.14	0.29	0.64	0.39	293
2	1.87	-0.02	0.62	0.38	0.80	0.40	5
3	99.55	-1.00	0.62	0.26	0.71	0.34	3073
4	36.38	36.38	36.38	36.38	1.00	1.00	1

Figure 3: K = 4 Clustering Results

In this figure 3 are the cluster, the returns statistics for the cluster, and the size of the clusters for K = 4 clusters.

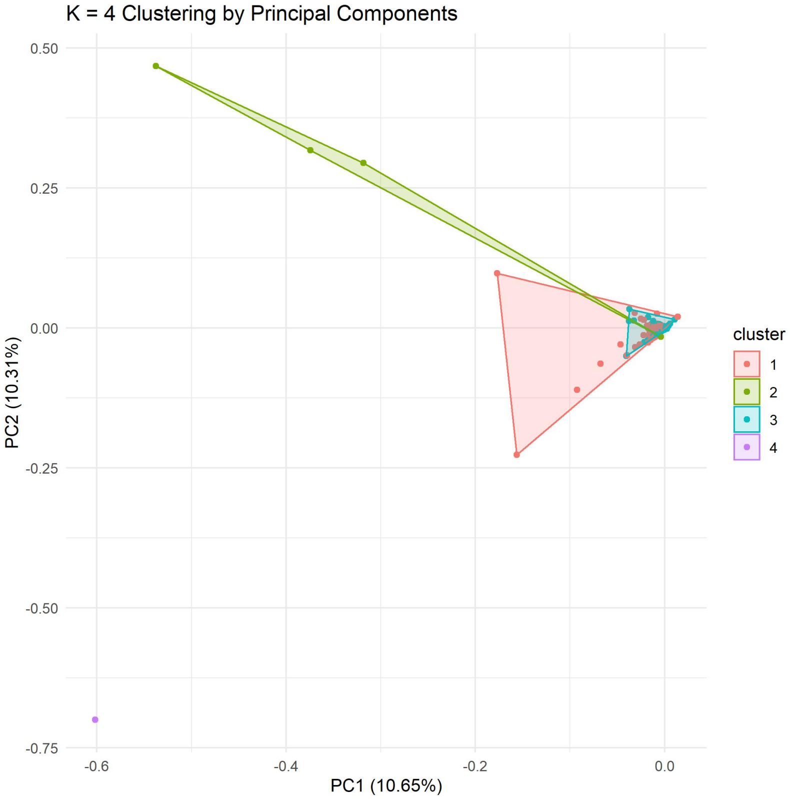


Figure 4: K = 4 Cluster Visualization by PC1 and PC2

This figure 4 shows the how the cluttered cluster shapes according to the top principal components is still present when reducing the number of clusters to 4 for the K-Means.

7 Discussion

The steps of inside the methodology related to the clustering, removal of outliers an re-clustering was repeated many times, as part of the analysis. This process however, only produced new outliers in terms of the clustering, and still resulted in many clusters with 1 or a few observations in it. Because of this continuation, it was decided to stop the outlier removing and use the last iteration of the data set.

With the final data set, the best number of clusters was determined to be tie between K = 15 and K = 4. K-Means clustering was performed, cluster returns were calculated, and cluster visualization was done with both of these K values. The previous figures in the Main Results section show the results from both numbers of clusters.

In figure 1 it is clear that outliers are still persistent in the data. There are 11 clusters, where there are 4 or less stocks in that group. Cluster 3 has 170 stocks, cluster 4 has 16, cluster 12 had 3,068, and cluster 15 had 99. What this says is that the great majority of companies financials look similar to each other, apart from a few outliers. The visualization 2 shows just how similar most of the companies look. The percent_beats_market column is the most important column to determine if the cluster is important, because it represents the proportion of the stocks in the cluster that would be worth an investment over a simple index fund. It is clear that in the clusters with a meaningful amount of stocks, there is no clear cluster with a majority of stocks that would be worthy of an investment because there are still over half of the stocks in the clusters that did worse than the broad index fund. This clustering provides essentially very little help to the investor looking to narrow down the search on time worthy investments, or potential stocks to research.

Even when you reduce the number of clusters to 4, another option determined to be equally the best number of clusters as 15, the success in separating good investments from bad investments still fails. Examining the results of clustering for K = 4 inside figure 3, there are still no clear distinctions of meaningfully sized clusters in which a majority of the results are beating the SP500. For cluster 1, with 293 stocks, the average return is beating the SP500, but the median return is not and only 39 percent of the stocks beat the index fund. For clusters 2 and 4, the clusters only had 5 and 1 stocks respectively, more pointing out outliers than clustering of companies. All of the remaining 3,073 stocks were then placed in cluster 3. This cluster represents the large majority of the stocks, and therefore provides no meaningful insights.

8 Conclusion

In conclusion, using K-Means Clustering on annual stock financial statements data to create clusters and find groups of individual stocks that produce better returns, or are worthy of further research, was not a successful endeavor. The results of clustering did not meaningfully separate the thousands of stocks into smaller sized clusters. Even in clusters that contained a relatively significant number of observations, not only a few points, but not all of the points, there still was no indication that that cluster contained better investments than the others. The majority of stocks in these clusters still did not outperform the general market, and were therefore not worthy of an investment. There are a lot more variables to consider than the 26 captured in the cluster analysis. Stock prices also move with a lag, and can reflect future expectations not explained in the data used. Overall, K-Means Clustering analysis did not create a viable investment strategy, or investment options subset.

For a future research direction, a form of dimension reduction could be applied to better cluster the stocks. Performing Principal Component Analysis could allow the stocks to be described in many fewer variables, while maintaining variance explanation, which could help in producing more equally distributed cluster sizes. It would also be potentially beneficial to perform cluster analysis on a smaller number of observations. An individual does not have to consider every possible investment available when making a decision, the total could be reduced with criterion to produce a smaller subset of options to research, such as company industry or stock share price. Further research could also be applied to data with a different time horizon such as 5 years or 1 year, as opposed to just 3 years.

References

[1] S. R. Department, “Nyse and nasdaq: Listed companies comparison q3 2022,” Nov 2022. [Online]. Available: https://www.statista.com/statistics/1277216/nyse-nasdaq-comparison-number-listed-companies/

[2] N. Da Costa Jr, J. Cunha, and S. Da Silva, “Stock selection based on cluster analysis,” *Economics Bulletin*, vol. 13, no. 1, pp. 1–9, 2005.

[3] B. Tekin and F. B. Gümüş, “The classification of stocks with basic financial indicators: an application of cluster analysis on the bist 100 index,” *Tekin, B. & Gumus, F., B.(2017), International Journal of Academic Research in Business and Social Sciences*, vol. 7, no. 5, 2017.

[4] Y. Finance, “Yahoo finance,” Dec 2022. [Online]. Available: https://finance.yahoo.com

[5] C. B. Murphy, “Financial statements: List of types and how to read them,” Nov 2022. [Online]. Available: https://www.investopedia.com/terms/f/financial-statements.asp#:~:text=Investopedia%20%2F%20Julie%20Bang-,What%20Are%20Financial%20Statements%3F,%2C%20financing%2C%20or%20investing%20purposes.

[6] eoddata.com, “Historical stock price data,” Nov 2022. [Online]. Available: https://eoddata.com/download

9 Codes

~~~
headers
```{r}
library(tidyverse)
library(ggfortify)
library(NbClust)
library(tidyquant)
library(factoextra)
library(ggplot2)
library(yfinance)
options(scipen = 99999999)
Get Data
```{r}
library(yfinance)
tickers <- read_csv("ticker_list.csv")
#get financial data from 2019-2021
get_data_2019_2021 <- function(ticker){
data <- get_financials(ticker)
data <- data %>% filter(str_detect(date, "2019") | str_detect(date,
"2021") )
return(data)
}

get_data_catch_error <- function(ticker){
  financials <- tryCatch(
    {
      ticker_data <- get_data_2019_2021(ticker)
    },
      error =
      {
        ticker_data <- data_frame()
      },
    warn =
      {
      ticker_data <- data_frame()
      },
      finally =
        {
          return(ticker_data)
        }
  )
  return(financials)
}

data <- data_frame()
~~~

```
for(ticker in tickers$ticker){
  financials <- get_data_catch_error(ticker)
  data <- bind_rows(data, financials)
}

write_csv(data, "stock_financials_2019_2021.csv")
‘‘‘
Cleaning Data and Normalizing
‘‘‘{r}
data <- read_csv("stock_financials_2019_2021.csv")
prices_2019_2021 <- read_csv("price_changes_2019_2021.csv")
#some stocks for the same year got split up into multiple rows so add
them back together
cleaned <- data %>% group_by(ticker, date) %>% summarise(across(.fns
= sum))

cleaned <- data %>% group_by(ticker) %>% filter(n() == 2)
#select the columns I want
less_data <- cleaned %>% select(
  ticker, date,
  cash,
  totalCurrentAssets,
  goodWill,
  totalAssets,
  accountsPayable,
  longTermDebt
  , totalLiab, totalCurrentLiabilities,
  retainedEarnings, totalStockholderEquity, netIncome,
  totalCashflowsFromInvestingActivities, capitalExpenditures,
  totalCashflowsFromInvestingActivities, dividendsPaid,
  totalCashFromFinancingActivities, changeInCash, repurchaseOfStock,
  issuanceOfStock, totalRevenue, costOfRevenue, grossProfit,
  researchDevelopment, totalOperatingExpenses, operatingIncome, interestExpense,
ebit
)

#filter to exclude stocks with erroneous and missing data for certain
categories
less_data <- less_data %>% filter( totalRevenue > 0 & !is.na(totalRevenue)
& !is.na(totalAssets) & !is.na(cash) & !is.na(ebit) & !is.na(accountsPayable)
                                  & !is.na(totalOperatingExpenses)
& !is.na(totalLiab) & !is.na(retainedEarnings)  ) %>% group_by(ticker)
%>% filter(n() > 1)


#replace NA with 0
```

```
less_data[is.na(less_data)] <- 0

less_data <- less_data %>% arrange(ticker, date)


three_year_difference <- less_data  %>%
  summarise(across(.cols = (-date), .fns = ~( . - lag(.) )/ lag(.))
)

three_year_difference <- three_year_difference %>% filter(!is.na(cash))

three_year_difference <- ungroup(three_year_difference)

#replace the NaN with 0
#thats the result of 0 divided by 0
three_year_difference[is.na(three_year_difference)] <- 0


finite_maximums <- three_year_difference
#replace infinite with dummy value
is.na(finite_maximums) <- sapply(finite_maximums, is.infinite)
#should i replace the infinite values with the max? other values? how
would that affect scaling/normalizing?
#lets replace with the maximums
replaced_maximums <- finite_maximums %>%
  mutate(across(everything(), ~replace_na(.x, max(.x, na.rm = TRUE))))

#Join with the price changes over this time
normalized_replaced_infinity <- normalized_replaced_infinity %>% left_join(prices_2019_2021,
by = c("ticker" = "symbol")) %>% filter(!is.na(change))
write_csv(normalized_replaced_infinity, "all_data.csv")

‘‘‘


Read in Data
‘‘‘{r}
data <- read_csv("all_data.csv")
‘‘‘
K-Means Clustering Sizes
‘‘‘{r}
#Best Cluster size
Nb.clust.kmeans <- NbClust(data = data_prices_available[3:28], method
= "kmeans")
#How does the voting looks
kmeans_clustersize <- Nb.clust.kmeans$Best.nc %>% as_tibble()
kmeans_clustersize <- kmeans_clustersize[1,] %>% pivot_longer(cols
= everything()) %>% group_by(value) %>% summarise(votes = n())

‘‘‘


Kmeans size 15
‘‘‘{r}
kmeans15 <- kmeans(data_prices_available[3:28], centers = 15, nstart
= 100)
#look at the data by cluster
kmeans15_data <- data_prices_available
kmeans15_data$cluster <- kmeans15$cluster

kmeans15_data <- kmeans15_data %>% mutate(positive = if_else(change
> 0, 1, 0), beats_market = if_else(change > sp500, 1, 0))
kmeans15_data_results<- kmeans15_data %>% group_by(cluster) %>% summarise(max_change
= max(change), min_change = min(change), avg_change = mean(change),median_change
= median(change) , percent_postive = sum(positive) / n(), percent_beats_market
= sum(beats_market)/n(), size_of_cluster = n())

#inside these clusters there are lots of clusters with only 1 or 2
data points

#what do thees clusters look like
autoplot(kmeans15, data = data_prices_available[3:28], frame = T )
+ theme_minimal() + ggtitle("K = 15 Clustering by Principal Components")
ggsave("k15_visual.jpg")
‘‘‘



4 was also proposed as a size, lets try this clustering
‘‘‘{r}
kmeans4 <- kmeans(data_prices_available[3:28], centers = 4, nstart
= 100 )

#look at the data by cluster
kmeans4_data <- data_prices_available
kmeans4_data$cluster <- kmeans4$cluster

kmeans4_data <- kmeans4_data %>% mutate(positive = if_else(change >
0, 1, 0), beats_market = if_else(change > sp500, 1, 0))
kmeans4_data_results<- kmeans4_data %>% group_by(cluster) %>% summarise(max_change
= max(change), min_change = min(change), avg_change = mean(change),median_change
= median(change) , percent_postive = sum(positive) / n(), percent_beats_market
= sum(beats_market)/n(), size_of_cluster = n())

#inside these clusters there are lots of clusters with only 1 or 2
```

```
data points

#what do thees clusters look like
autoplot(kmeans4, data = data_prices_available[3:28], frame = T, )
+ theme_minimal() + ggtitle("K = 4 Clustering by Principal Components")
ggsave("k4_visual.jpg")
```

‘‘‘