MATH 570 Project

Kassidy Chaikin, Savvas Giannaris, Ruihan Jiang

2022-05-25

Abstract: Our research is focused on a subset of stocks in the S&P500 data from April 2018 to April 2022. We extracted the data from the S&P500 using the tidyquant package. The S&P500 tracks stocks of 500 large-cap companies. Naturally, this data set is quite large, so we decided to narrow down which stocks we would look at. In order to do so we selected four different sectors from the S&P500. In specific, we selected the Technology, Financial, Consumer Discretionary, and Health Care Sectors. We then selected five different stocks within each of these sectors that are all on the NYSE (not NASDAQ). The resulting data frame contained 20,140 observations. We partitioned this data into four different data frames, each containing 5,035 observations (1007 per stock). We selected the years 2018-2022 to perform our analysis using the most recent data as possible to understand the impact of the COVID-19 pandemic on the NYSE. In this research, we hope to understand how the COVID-19 pandemic impacted mean return patterns in different sectors. Furthermore, we hope to what hidden factors may be shared between different sectors of the exchange, and how these factors influence different sectors. Finally, we searched to find clear clusters of stocks within our subset of the S&P500. Do we find that different stocks of varying sectors cluster together, or are they clearly distinguished?

Introduction: The New York Stock Exchange (NYSE) is the oldest American exchange and largest equities-based exchange in the world based on the total market capitalization of its listed securities, and the Nasdaq is a global electronic market for buying and selling trading securities. The NYSE is an auction market uses designated market makers (DMMs) to regulate the stock market, whereas, the Nasdaq is a dealer market with many market makers in competition with one another. Market makers at the Nasdaq use their own accounts in transactions with individual customers and other dealers using maintained inventories of stock to buy and sell. Market makers at the Nasdaq will state the bid and ask price for a security that they are making a market in, and over 260 market-making firms provide liquidity for Nasdaq-listed stocks. DMMs on the NYSE are what gives the NYSE a more personal touch, they are real people that sit on the trading floor and are the point of contact for the listed company on the floor. DMMs provide stability by taking the other side of the trade when imbalances occur, buying when investors are selling, and vice versa. They run the opening and closing auctions, using human input and algorithms to help promote price discovery when the volume is typically at its highest. We chose to work with S&P500 stocks exclusively on the NYSE. Stocks are split into sectors based on the characteristics of the company and what type of service or good they supply. We decided to work with four sectors more closely; Technology, Health, Financials, and Consumer Discretionary. Within each sector we selected five stocks. We decided to do so to be able to narrow down the sample size and try to present more coherent data that could then be replicated for other stocks, sectors, and those also on the Nasdaq.

Multivariate Paired Comparisons Hypothesis Test

In our first analysis of S&P500 sectors, we searched to find changing mean return patterns among different sectors in the years 2018-2021. To accomplish this, we selected 5 stocks from 4 sectors of the New York Stock Exchange (NYSE). Specifically, we investigated TWTR, HPQ, TEL, NOW, and ORCL from the Technology Sector, we investigated NKE, MCD, CMG, AAP, and CCL from the Consumer Discretionary Sector, we investigated MA, WFC, MS, BRO, and CB from the Finance Sector, and we investigated UNH, LLY, PEN, DVA, and CVS from the Health Sector.

We organized the daily returns of these stocks into 5 data frames of dimension 1007 x 6, where the columns are the returns for each stock and the rows represent the daily return for each stock every day over the years 2018-2021, and removed 88 rows per sector due to missing values. The first 5 columns of our data are the tickers we selected, the last column is the date, represented by the year in which the calendar day lies.

```
##   TWTR_returns  HPQ_returns  TEL_returns NOW_returns ORCL_returns Year
## 1   0.00000000  0.000000000  0.000000000 0.000000000  0.000000000 2018
## 2  -0.00244794  0.002345196  0.016890690 0.008957778  0.023161193 2018
## 3  -0.01881395  0.007486793  0.004642323 0.007072470  0.009851323 2018
## 4   0.01375573  0.010218408  0.022797352 0.008666335  0.006019139 2018
## 5   0.01110197  0.004138169  0.002108242 0.002148048  0.010521631 2018
## 6  -0.01708011 -0.001373688 -0.009918635 0.001108603  0.005534185 2018
```

```
##     NKE_returns   MCD_returns  CMG_returns  AAP_returns  CCL_returns Year
## 1  0.0000000000  0.0000000000  0.000000000  0.000000000  0.000000000 2018
## 2 -0.0001574385 -0.0042142236  0.054787463  0.009048992  0.002247202 2018
## 3 -0.0006301648  0.0070144588 -0.006116553  0.036898443 -0.000747469 2018
## 4  0.0085120391  0.0020151743  0.021751243  0.010630675 -0.008227221 2018
## 5  0.0089090242 -0.0006896502  0.014914407 -0.007042227 -0.003770610 2018
## 6 -0.0071262925 -0.0022420127  0.002825993 -0.008079648  0.023920943 2018
```

```
##    MA_returns  WFC_returns    MS_returns  BRO_returns   CB_returns Year
## 1 0.000000000  0.000000000  0.0000000000  0.000000000  0.000000000 2018
## 2 0.012573024  0.007693546  0.0026825373  0.011354637  0.007870766 2018
## 3 0.012937311  0.012508108  0.0152875809  0.002709887  0.004043732 2018
## 4 0.020730169  0.006738198 -0.0003765455  0.009459533  0.001805662 2018
## 5 0.003022975 -0.011314579 -0.0041422036 -0.003059740 -0.005060174 2018
## 6 0.001444081  0.003545872  0.0077519210  0.008056723  0.016163912 2018
```

```
##    UNH_returns  LLY_returns  PEN_returns DVA_returns  CVS_returns Year
## 1  0.000000000  0.000000000  0.000000000 0.000000000  0.000000000 2018
## 2  0.010489897  0.005432179  0.002213669 0.005100675 -0.004352564 2018
## 3  0.004340484  0.004463225 -0.065709606 0.026608063  0.026366062 2018
## 4  0.019068894  0.012277898  0.006501229 0.009084850  0.044190119 2018
## 5 -0.017356692 -0.005082827  0.018203112 0.001721237 -0.008667890 2018
## 6  0.004983108 -0.000812628  0.034602077 0.023922772 -0.001028734 2018
```

Data Organization: To understand the mean return patterns over these 3 years, we organized the data into 4 data frames of dimension 251 x 15, where each column shows the difference between a specific stocks difference between the year of investigation (2019, 2020, 2021) and 2018. Below shows the data structure for our difference matrix for the technology sector. The remaining data structures are in the code, but were not outputted for reading convenience.

```
##   D2019.TWTR_returns D2019.HPQ_returns D2019.TEL_returns D2019.NOW_returns
## 1       -0.002435595     -0.0083088871       0.005156596      -0.001516450
## 2        0.026014366      0.0362763372       0.054237791       0.060550485
## 3       -0.088838991     -0.0321517718      -0.021589834      -0.052707576
## 4       -0.032654918     -0.0008822547       0.003020448      -0.032844585
## 5       -0.003575723     -0.0082721344       0.005406732      -0.005566116
## 6       -0.031231089     -0.0013736878      -0.030037936      -0.011437248
##   D2019.ORCL_returns D2020.TWTR_returns D2020.HPQ_returns D2020.TEL_returns
## 1       -0.001550298       -0.007800312      -0.011678997      -0.006990844
```

```
## 2        0.032891308        0.021700637       0.014851245       0.028702966
## 3       -0.033248197       -0.022621021       0.003590037       0.011562878
## 4       -0.009823377       -0.014689339       0.007792559       0.002102526
## 5        0.001459426       -0.004570982      -0.008930539      -0.008236131
## 6        0.007622840       -0.022223896      -0.016184951      -0.010942423
##   D2020.NOW_returns D2020.ORCL_returns D2021.TWTR_returns D2021.HPQ_returns
## 1      -0.031595292      -0.0183088480       -0.007017488        0.01952006
## 2       0.009438426       0.0266830450        0.009472068       -0.01217151
## 3       0.000992123       0.0046428451       -0.007306838       -0.01295478
## 4       0.008529725       0.0037987371        0.031217167       -0.02143212
## 5      -0.007206327       0.0066269566        0.027345084        0.01268168
## 6      -0.011270913       0.0009155885        0.047022450       -0.01116610
##   D2021.TEL_returns D2021.NOW_returns D2021.ORCL_returns
## 1       0.006359984      0.0428755270        0.014530850
## 2      -0.002311259     -0.0007986028        0.035553321
## 3      -0.024066599      0.0535412037        0.012242896
## 4       0.002738750      0.0004652429       -0.001971906
## 5      -0.011959916     -0.0334985187        0.005606490
## 6      -0.009918635      0.0135321678        0.011056435
```

Analysis: Upon creation of our Difference data frames (D), we can now perform a Multivariate Paired Comparisons Hypothesis Test under the Null Hypothesis

$$H_0 : \delta = 0$$

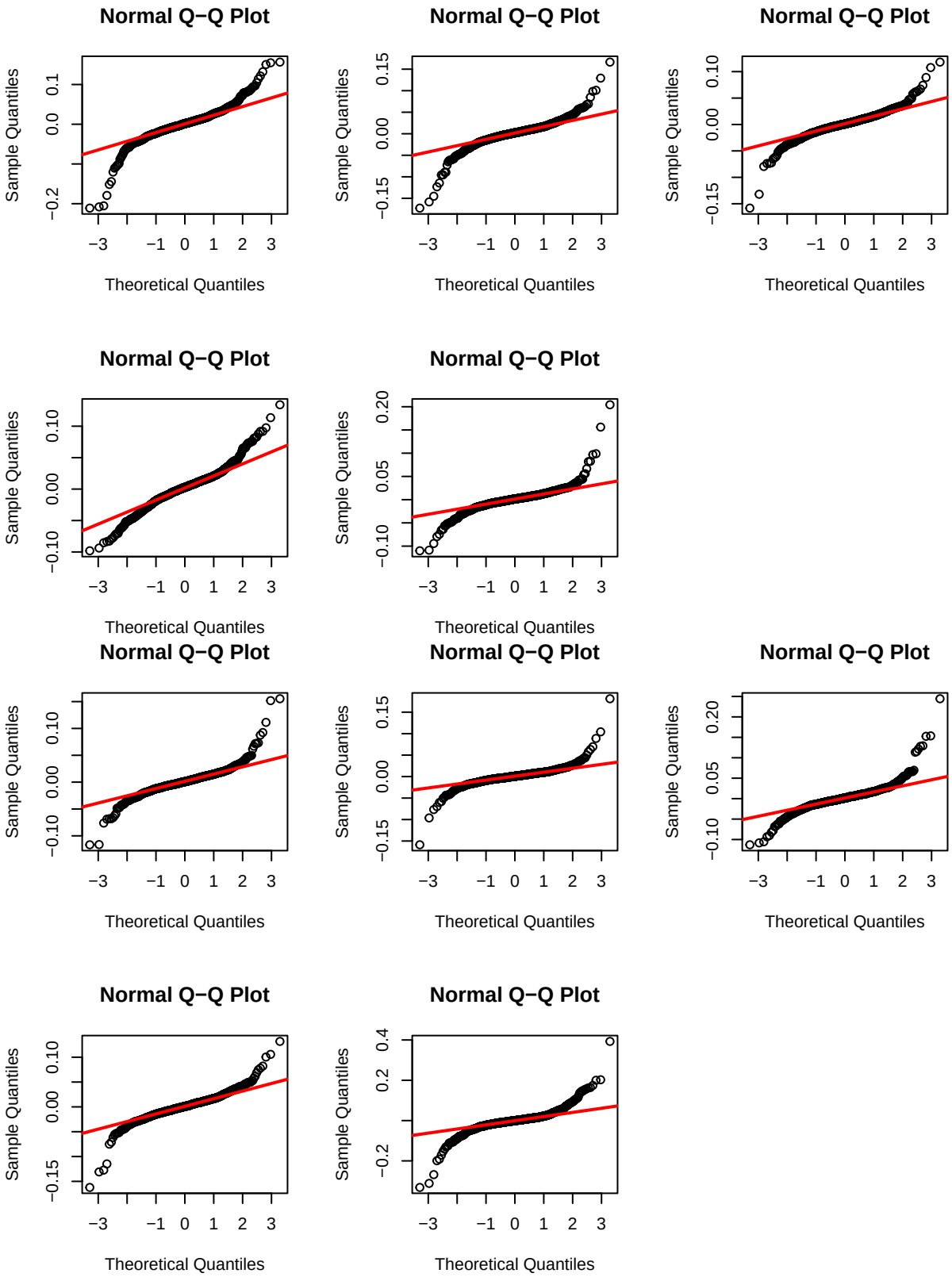
Because the data is large and not normally distributed, we replace the F-quantile value,

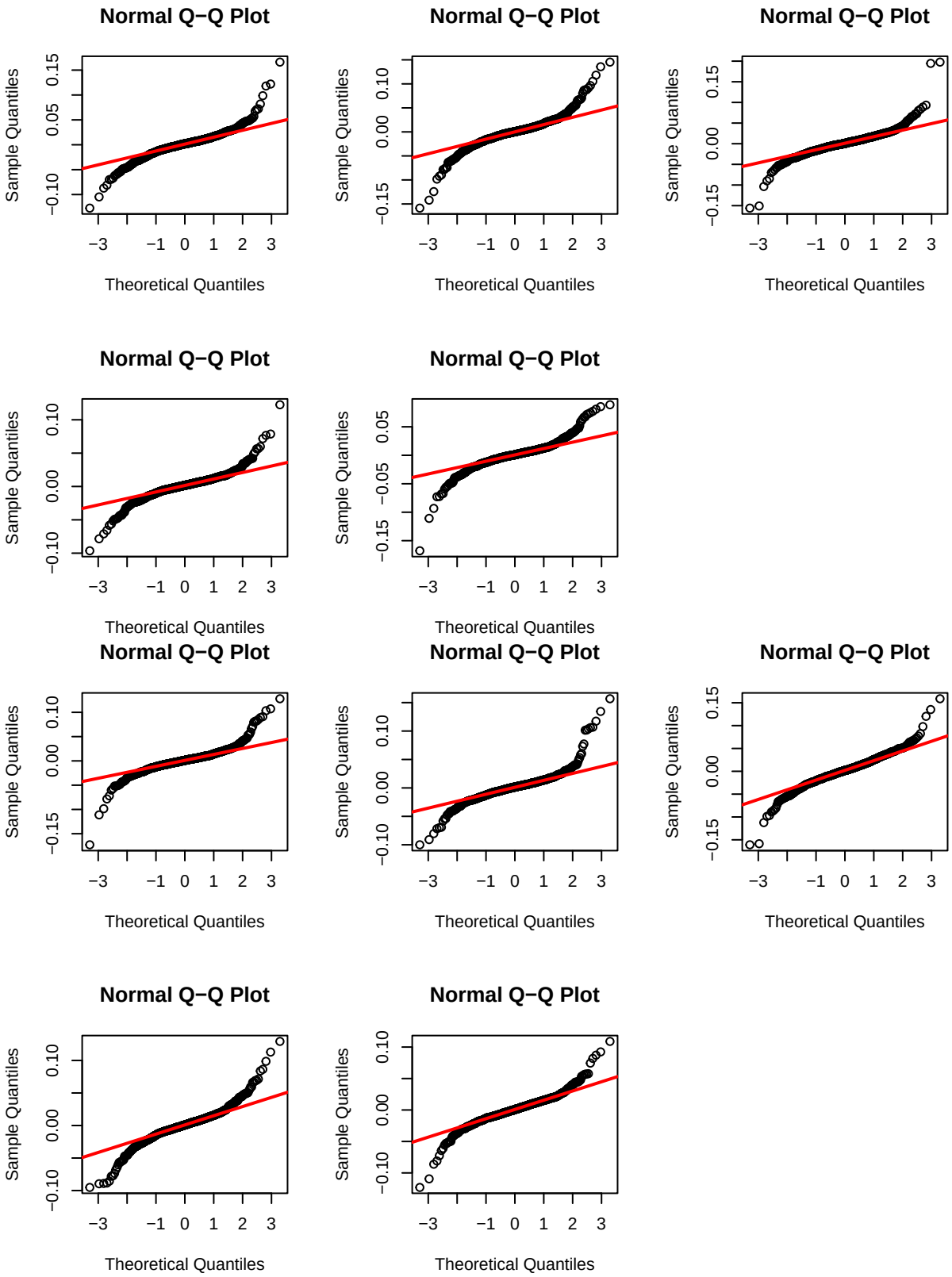
$$\frac{(n-1)*p}{(n-p)}F_{p,n-p}$$

with

$$\chi^2_{\alpha,p}$$

in our Hypothesis Test. For our tests, we will use a significance level of $\alpha = .05$. We know the data is not normally distributed from the QQ plots below.


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.2 0.0 0.1
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.15 0.00 0.15
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.15 0.00 0.10
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.10 0.00 0.10
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.10 0.05 0.20
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.10 0.00 0.10
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.15 0.00 0.15
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.10 0.05 0.20
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.15 0.00 0.10
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3 −1 0 1 2 3
−0.2 0.2 0.4



Interpretation: It is clear that the QQ-plots above are heavily tailed, henceforth not Normal. So, we can safely replace the F-quantile with $\chi_{\alpha,p}$.

Analysis For the first step of our Hypothesis Test, we computed the sample mean and sample covariance for

each of our difference data frames. Furthermore, we computed our n and p values, which were 251 and 15 respectively.

As the dimension of all our difference matrices is the same, we compute the same Chi-Square Quantile for each of our tests. This quantile is our Q2 below

```
## Quantile
## 7.260944
```

We now conduct our hypothesis test. We reject $H_0 : \delta = 0$ when

$$n(\bar{D} - 0)'(S_D)^{-1}(\bar{D} - 0) = n(\bar{D})'(S_D)^{-1}(\bar{D}) > \chi^2_{.05,15}$$

```
##                       [,1]              [,2] [,3]
## Tech Test Statistic "7.8130098815429" ">"  "7.26094392767003"
```

```
## [1] "Reject The Null Hypothesis H0"
```

```
##                                [,1]               [,2] [,3]
## Consumer Disc Test Statistic "6.58394270945454" "<"  "7.26094392767003"
```

```
## [1] "Accept The Null Hypothesis H0"
```

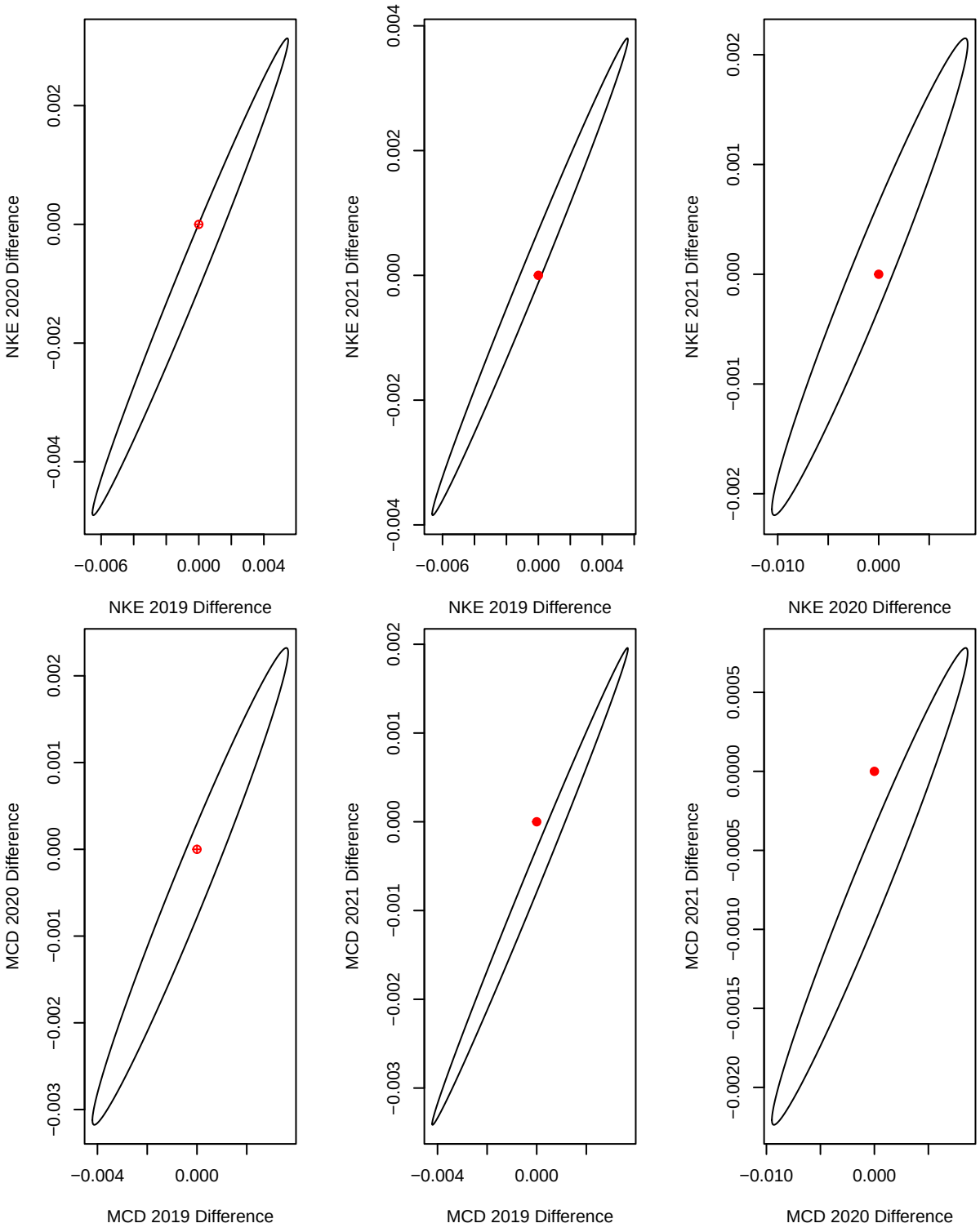
```
##                          [,1]               [,2] [,3]
## Finance Test Statistic "15.7692576203837" ">"  "7.26094392767003"
```

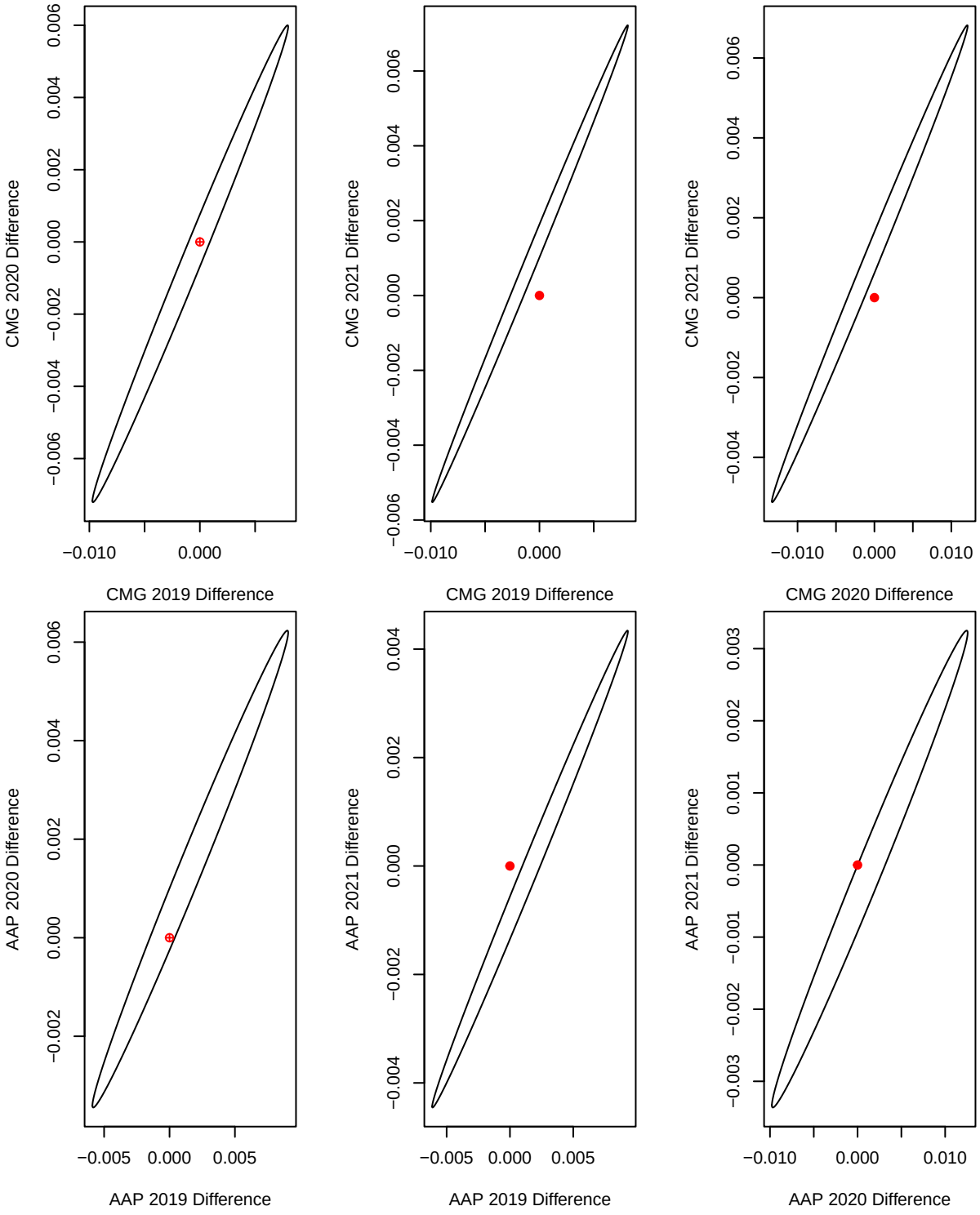
```
## [1] "Reject The Null Hypothesis H0"
```

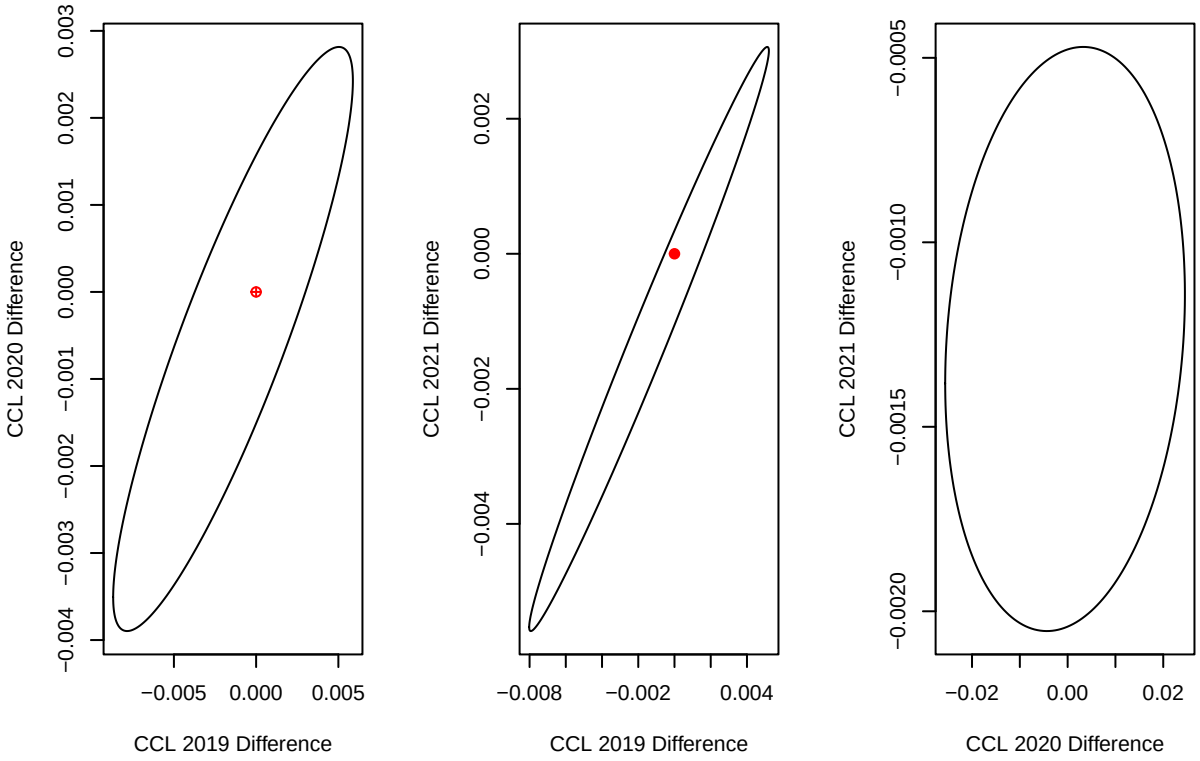
```
##                         [,1]              [,2] [,3]
## Health Test Statistic "7.8130098815429" ">"  "7.26094392767003"
```

```
## [1] "Reject The Null Hypothesis H0"
```

Answer 1: From the results of this test, it is clear that we reject H_0 for the technology, finance and health sectors of the market. On the other hand, we find that upon conducting this test on the Consumer Discretionary Sector, we fail to reject $H_0 : \delta = 0$. This result is quite interesting. To further this investigation, we perform an analysis on each of the stocks highlighted in this difference data frame. For each ticker, we computed the 95% confidence ellipsoid for d2019 vs. d2020, d2019 vs. d2021, and d2020 vs d2021.


NKE 2020 Difference
0.002
0.000
−0.002
−0.004
−0.006
0.000
0.004
NKE 2019 Difference
NKE 2021 Difference
0.004
0.002
0.000
−0.002
−0.004
−0.006
0.000
0.004
NKE 2019 Difference
NKE 2021 Difference
0.002
0.001
0.000
−0.001
−0.002
−0.010
0.000
NKE 2020 Difference
MCD 2020 Difference
0.002
0.001
0.000
−0.001
−0.002
−0.003
−0.004
0.000
MCD 2019 Difference
MCD 2021 Difference
0.002
0.001
0.000
−0.001
−0.002
−0.003
−0.004
0.000
MCD 2019 Difference
MCD 2021 Difference
0.0005
0.0000
−0.0005
−0.0010
−0.0015
−0.0020
−0.010
0.000
MCD 2020 Difference

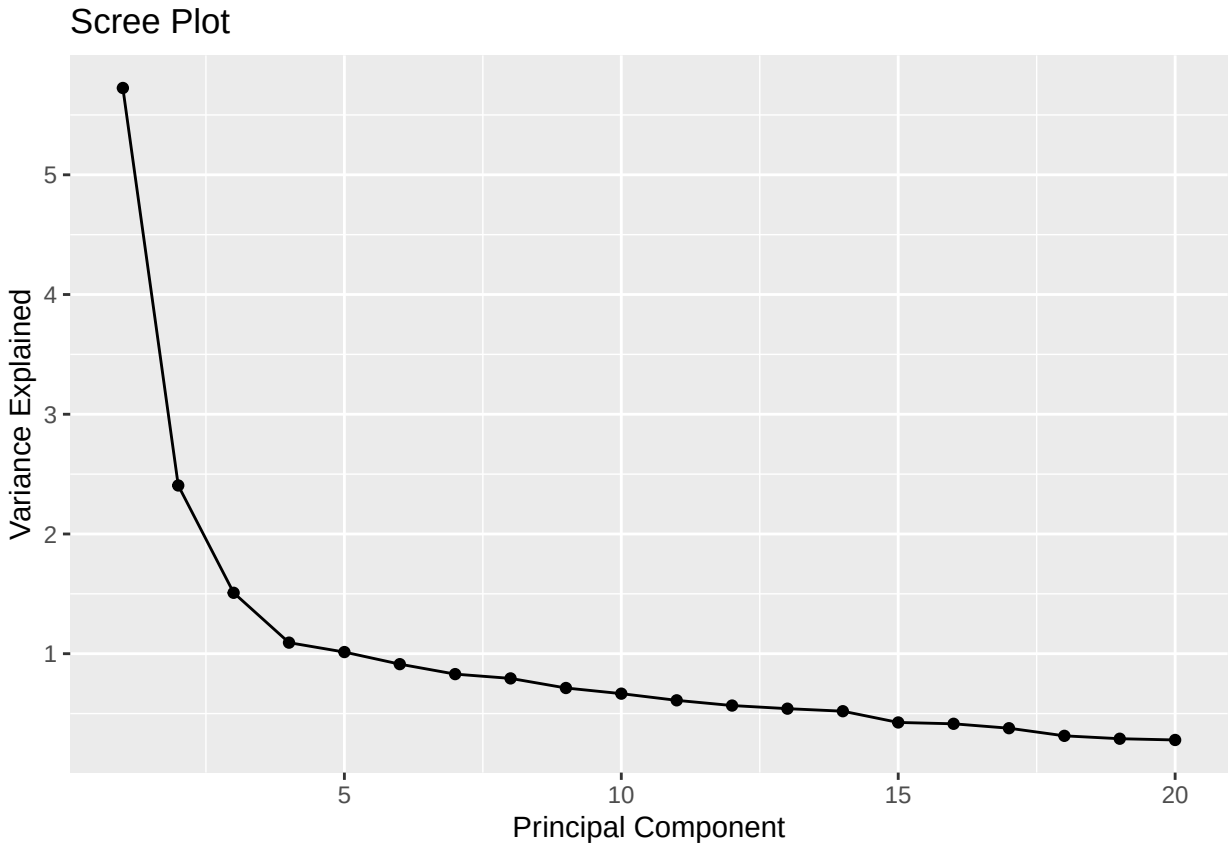

CMG 2020 Difference
CMG 2019 Difference
CMG 2021 Difference
CMG 2019 Difference
CMG 2021 Difference
CMG 2020 Difference
AAP 2020 Difference
AAP 2019 Difference
AAP 2021 Difference
AAP 2019 Difference
AAP 2021 Difference
AAP 2020 Difference



Conclusion 1: From the results of these 95% confidence ellipsoids, it is clear that the difference in return of the stocks that we selected in the Consumer Discretionary Sector from 2018 to 2021 was approximately 0. Initially, we were curious to see if the mean return patterns for these Sectors was influenced by the COVID-19 pandemic. It is an interesting result that we do not reject H_0 for the Consumer Discretionary Sector, as this sector contains large consumer businesses like Nike(NKE), Starbucks(SBUX), McDonalds(MCD), Chipotle(CMG), etc. The mean return pattern for this sector of the NYSE could be explained by many of the societal influences in 2018-2021. For example, during the COVID-Pandemic, the population of the United States was forced into quarantine for an extended period of time. As a result, national consumption was halted, as citizens remained in their homes and saving money, as many jobs went remote and for others, reliable income halted too. From the results of our hypothesis test, 0 is contained by a 95% confidence interval when testing the difference in mean returns for the Consumer Discretionary. For the remaining sectors, we reject the Null hypothesis, and confirm that there was a difference in mean returns for each sector between 2018 and 2021.

Factor Analysis

Do stocks in different sectors of the NYSE share similar hidden factors? (Factor Analysis)




```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
## # A tibble: 20 x 8
##    Factor1   Factor2  Factor3  Factor4   Factor5 communality uniqueness Stock
##      <dbl>     <dbl>    <dbl>    <dbl>     <dbl>       <dbl>      <dbl> <chr>
##  1  -0.334 -0.562    -0.0145   0.00640  0.0852         0.435      0.565 Twitter
##  2  -0.570 -0.0262   -0.127   -0.271   -0.0516         0.418      0.582 HP Inc
##  3  -0.745 -0.170    -0.0956  -0.215    0.0219         0.639      0.361 Te Conn~
##  4  -0.340 -0.718     0.249    0.0415  -0.000956       0.695      0.305 Service~
##  5  -0.377  0.213     0.372   -0.00109 -0.503          0.579      0.421 Oracle ~
##  6  -0.651 -0.243     0.0860   0.232   -0.170          0.572      0.428 Masterc~
##  7  -0.641  0.310    -0.352   -0.246   -0.103          0.702      0.298 Wells F~
##  8  -0.712  0.127    -0.364   -0.163   -0.0643         0.687      0.313 Morgan ~
##  9  -0.688 -0.0239    0.130    0.211   -0.201          0.576      0.424 Brown a~
## 10  -0.715  0.365    -0.0842   0.0213  -0.0986         0.662      0.338 Chubb L~
## 11  -0.528  0.296     0.466    0.138    0.229          0.656      0.344 United ~
## 12  -0.136  0.0226    0.555   -0.677   -0.0151         0.785      0.215 Eli Lil~
## 13  -0.264 -0.471     0.0246  -0.326    0.112          0.411      0.589 Penumbra
## 14  -0.387  0.285     0.123    0.0876   0.663          0.694      0.306 DaVita
## 15  -0.449  0.468     0.291   -0.0790   0.231          0.564      0.436 CVS
## 16  -0.487 -0.267    -0.00598  0.0805  -0.126          0.331      0.669 Nike
## 17  -0.604  0.000390  0.0933   0.395    0.00783        0.529      0.471 McDonal~
## 18  -0.465 -0.605     0.135    0.0899   0.190          0.645      0.355 Chipotle
## 19  -0.618  0.366    -0.0313   0.119   -0.0561         0.534      0.466 Advance~
## 20  -0.479 -0.109    -0.584   -0.0515   0.208          0.629      0.371 Carniva~
```

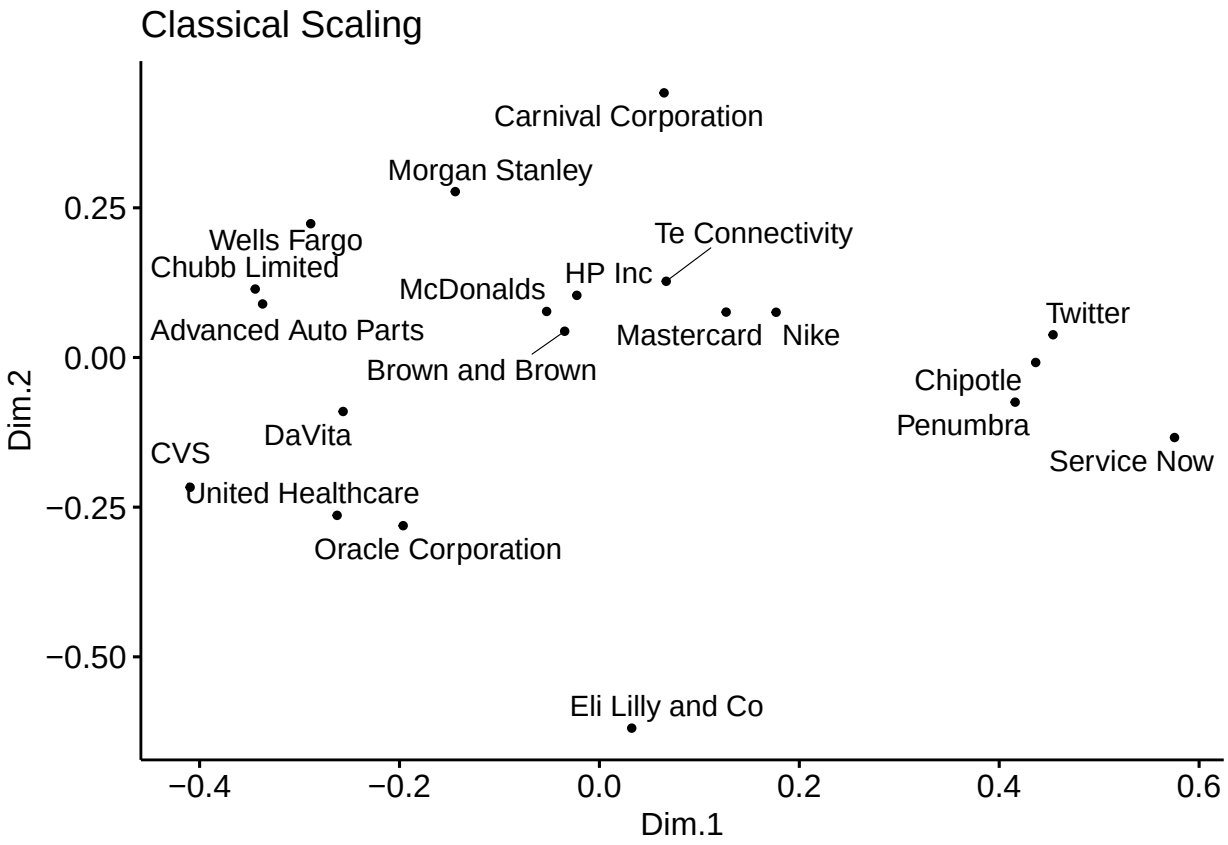
To begin we first collected all the data for each of the stocks we selected to use, and found the return for each stock to then compile them all into a matrix. For the factor analysis we chose to just look at the data from the year 2021 This matrix ends up being a 251x20 matrix with each stock being a column (20) and each day of trade being a row (251). Then we looked at the scree plot and Kaiser's rule of thumb, which suggest using 5 factors to complete the rest of the factor analysis. With five factors it is more clear to see which factor certain stocks load more heavily on versus the other factors which is more ideal when trying to figure out if there are common underlying factors that stocks in different sectors hold. We want to be able to see that stocks load heavily on one factor over the others rather than being directly in the middle of multiple factors in order to come to this conclusion and m=5 factors allows us to do this.

Conclusion : Although it is difficult to truly say what each factor is, we have made some educated guesses based on some further analysis, and would like to share a few of these with reasoning below. We believe factor one could be an independent factor that is related to the pricing of the stock, as all 20 of our stocks load negatively on factor one with stocks in every sector loading extremely negatively (between -0.5 and -0.7) on it, and this would make sense since the pricing of stocks is based on a company by company basis. We believe factor two could be related to consumer investment, as a majority of financial sector stocks load most heavily on this factor (Wells Fargo, Morgan Stanley, and Chubb Limited) all of which have much to do with benefiting an individual whether its homeowners insurance or a bank loan, these companies have dealings with consumer investment in the long term. We believe factor three is driven by the health needs of the population during the covid-19 pandemic as many more people were in need of healthcare insurance and pharmaceuticals with so many people getting sick and being hospitalized and treated for illness and United Healthcare and Eli Lily and Co loaded the most heavily on factor 3. We believe factor four is related to industries that were still profiting during the covid-19 pandemic, this is because we found Mastercard and McDonalds loads the most heavily on this factor and during the pandemic credit card usage still remained the same if not more as in the year 2021 people were finally out of their homes and out of lock down looking to do anything fun and normal which typically means spending money, and fast food restaurants were a great way to get food cheap and distantly while so many were out of work and in need of cheaper food and also for those that were gaining their normal busy lives back and looking to just grab a quick bite. We believe factor five is related to the production of healthcare devices and services that are needed by specific populations both DaVita and Penumbra load heavily on this factor along with United Healthcare and DaVita provides kidney dialysis and Penumbra makes devices for interventional therapies, and United Healthcare provides health insurance which lowers the cost for hospital visits, doctors visits, etc.

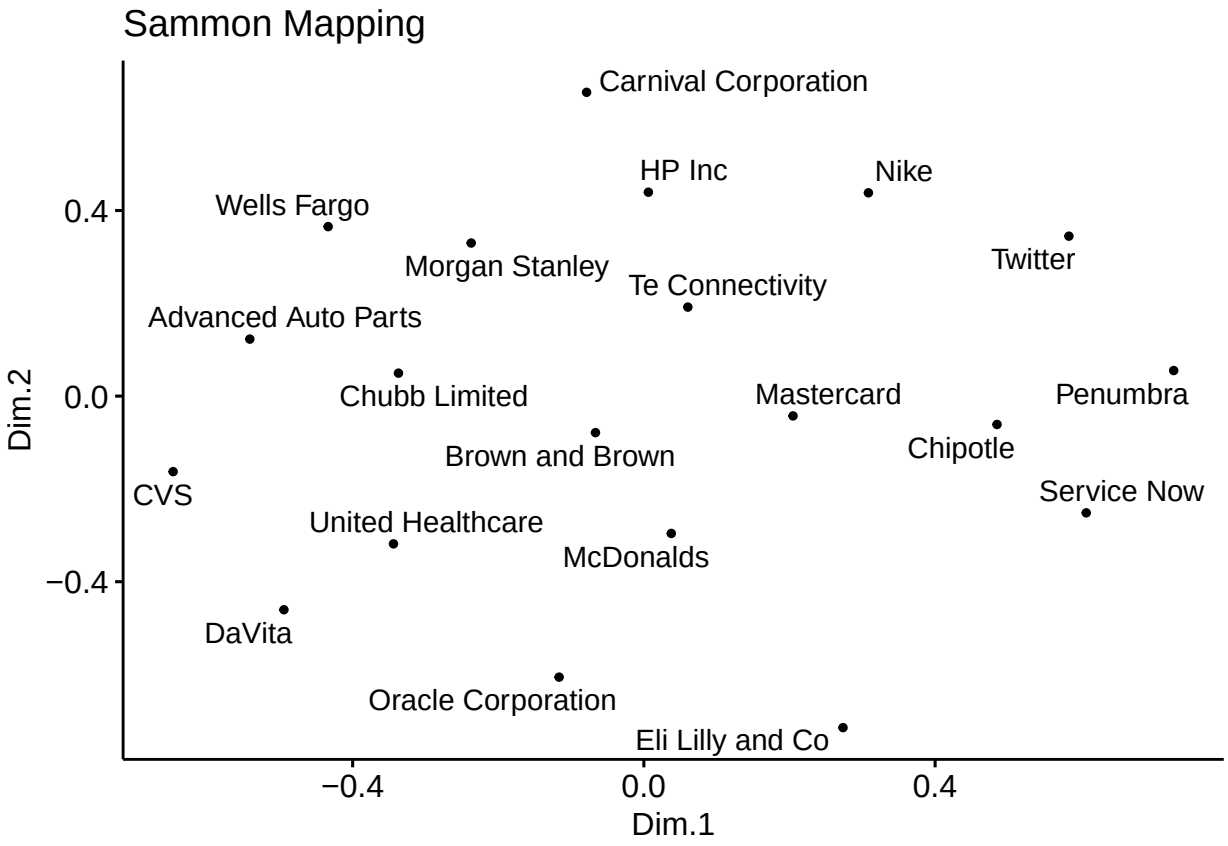
This leads us to our ultimate conclusion that stocks in different sectors of the NYSE share similar hidden factors.

How to embed/visualize stocks in a 2-dimensional or 3-dimensional map so that similar stocks, however it is defined, are close in the map. (multidimensional scaling)

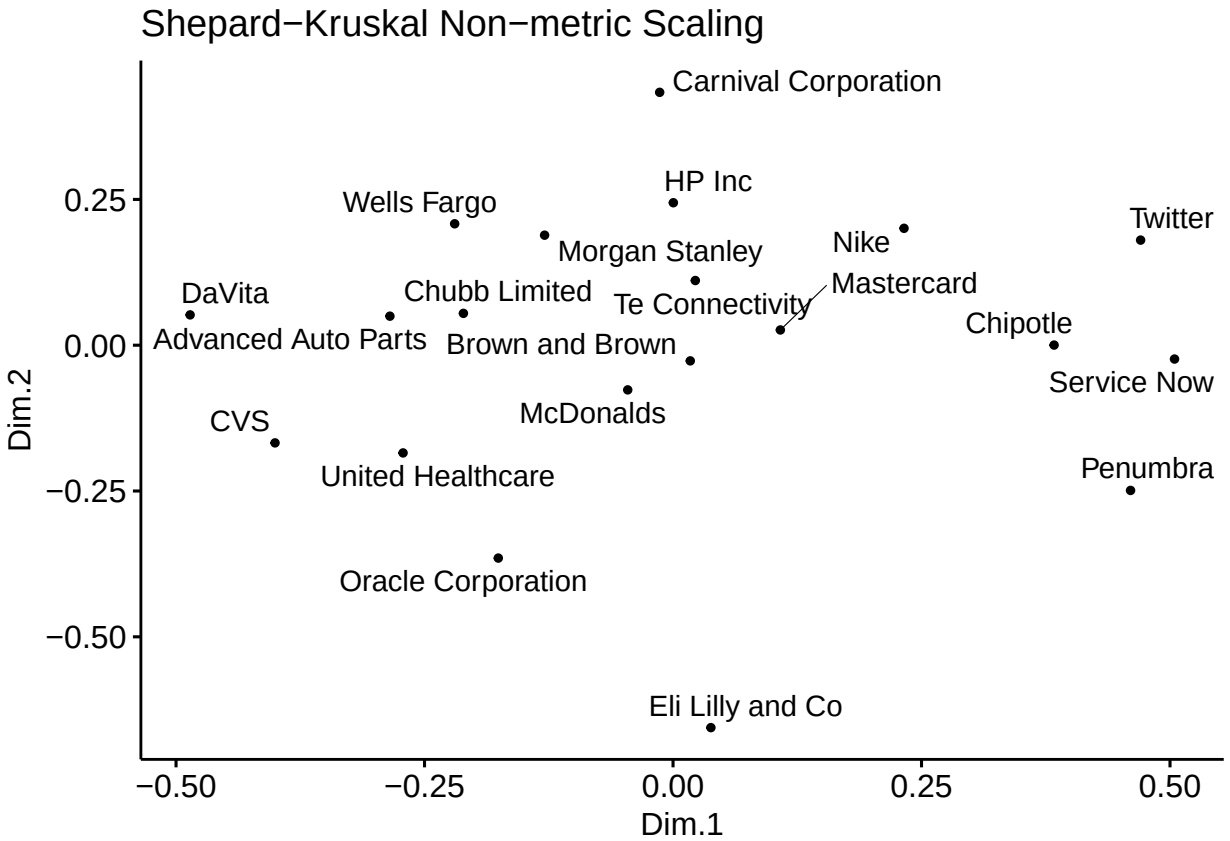
Multidimensional Scaling




```
## Initial stress        : 0.20108
## stress after  10 iters: 0.10079, magic = 0.092
## stress after  20 iters: 0.07955, magic = 0.213
## stress after  30 iters: 0.07355, magic = 0.500
## stress after  40 iters: 0.07329, magic = 0.500
## stress after  50 iters: 0.07322, magic = 0.500
```




```
## initial  value 22.826958
## iter   5 value 15.680302
## final  value 15.514859
## converged
```



First we found find the dissimilarity matrix by finding the correlation matrix and subtracting from one. This gives us a 20x20 matrix where each column is a stock and each row is also a stock and the data is one minus the correlation between each stock. This allows us to complete the MDS methods. Using classical scaling, it seems TWTR, CMG, PEN, and NOW are clustered together. Also, TEL, HPQ, MCD, BRO, MA, NKE are clustered together, and WFC, CB, and AAP are clustered together. LLY and CCL definitely do not fit into any of the clusters it is very far away from all the clusters formed. And there is a cluster formed by CVS, UNH, and ORCL. MS seems to be between two different clusters and DVA looks to be in the middle of three different clusters, so it would be difficult to draw a conclusion about these stocks. Using Sammon mapping, there appears to not be any obvious clusters. The stocks look to be spread out across the entire plot, so it seems that Sammon mapping would not be very helpful for us in visualizing stocks in a 2-dimensional map.

Using Shepard-Kruskal non-metric scaling, we see TWTR, CMG, NOW, and PEN being clustered together just as they were in classical scaling. We also see MCD, BRO, MA, TEL clustered together. There also could be a cluster including MS, HPQ, CCL, WFC, CB, and AAP. DVA could possibly be included in the cluster with AAP, but it is a bit further away and could just not be part of any cluster. There is also a cluster that contains CVS, UNH, and ORCL, also just as with classical scaling. NKE looks like it is between two clusters and it could be difficult to really say which cluster it belongs to. LLY is far away from all clusters formed and definitely cannot be considered in any of them. When comparing the scaling methods, it seems to us that classical scaling is the ideal method to use as when looking at the plot the clusters appear to be more clear and obvious than when looking at the non-metric scaling plot, and it is very clear that the Sammon-mapping gives no clusters at all. This makes sense since classical scaling focuses on approximating the actual dissimilarity and non-metric scaling only uses the rank order among the dissimilarities.

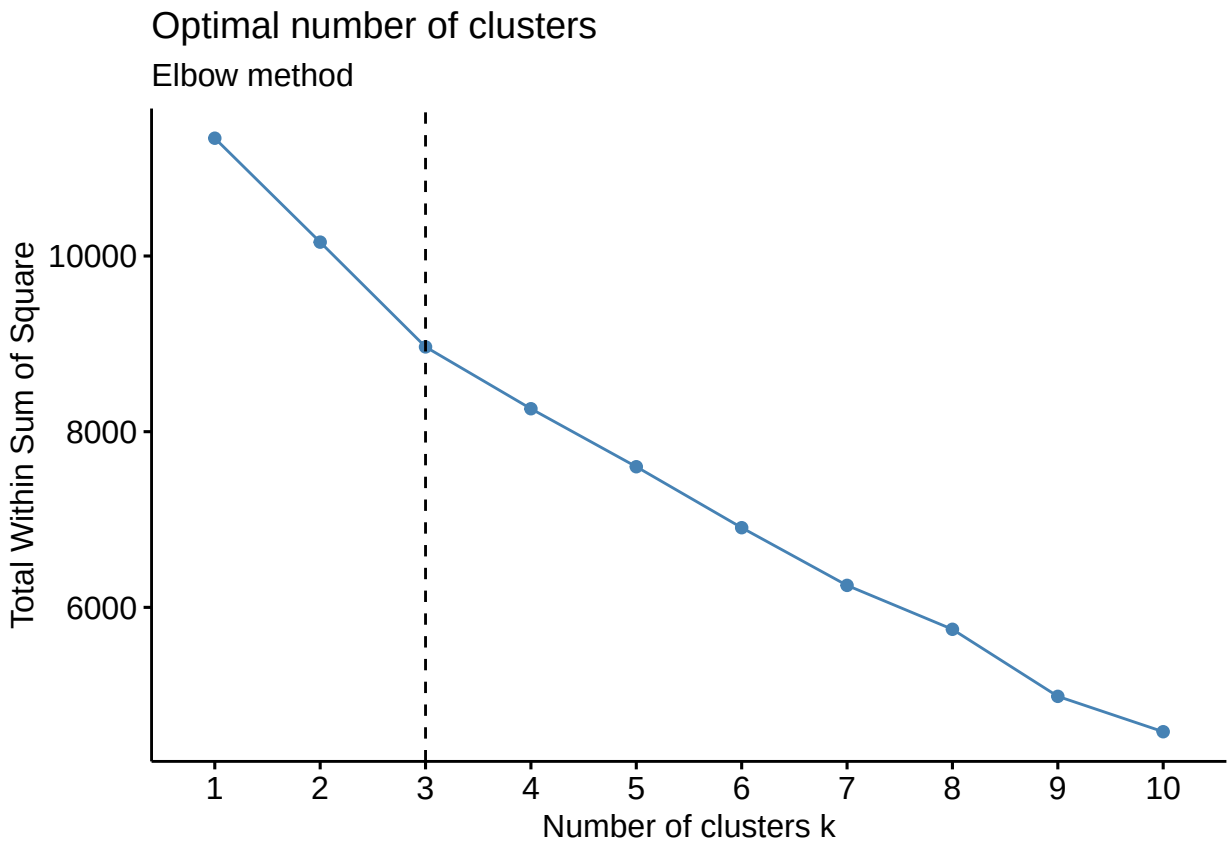
Cluster Analysis & ICA

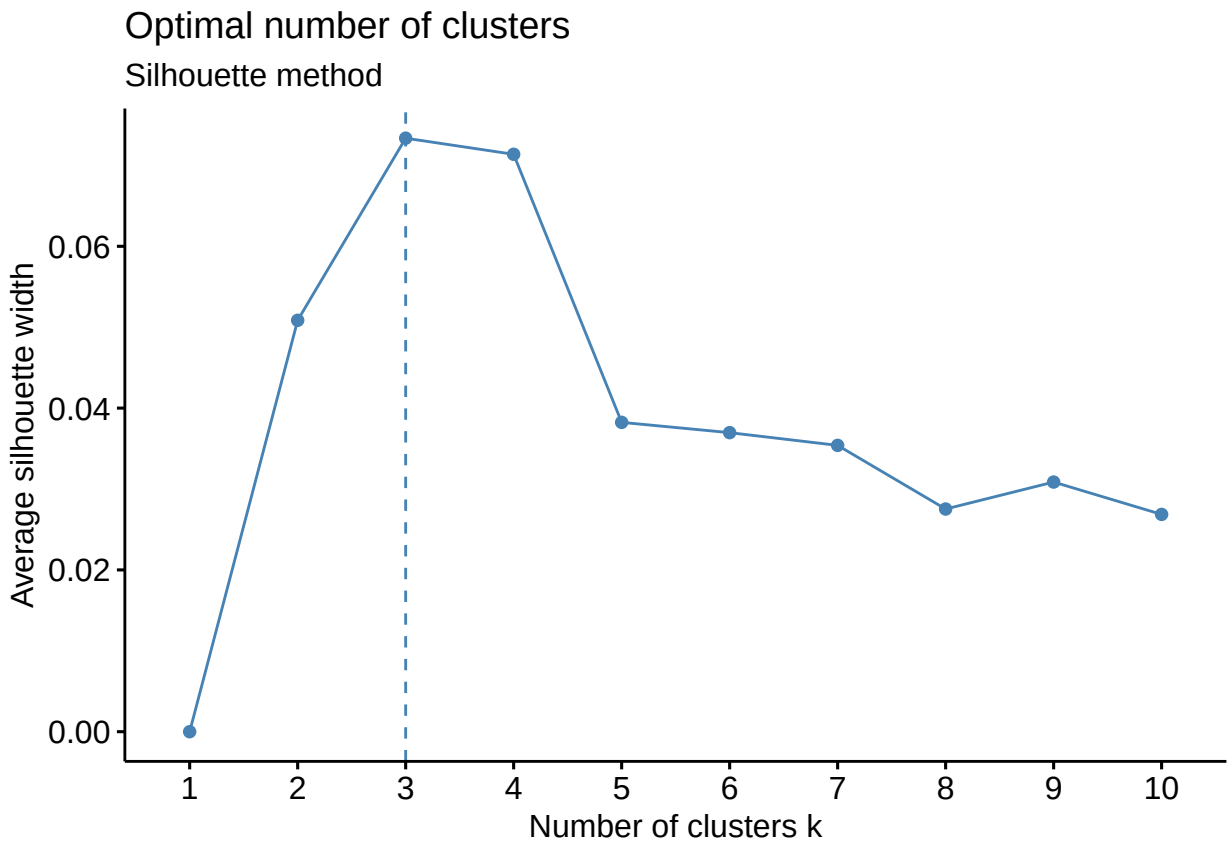
Prepare Data for Cluster Analysis and Independent Components Analysis: Choose the data from 2018 to 2021. Every row is a date and every column is a different ticker.

K-Means Cluster Analysis

1. Each row name of data represents a stock, each column name of data is a date. Therefore, the data is a 20 by 1007 matrix.
2. Determine Optimal Clusters: From both Elbow and Silhouette methods, it seems that k=3 is good.

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

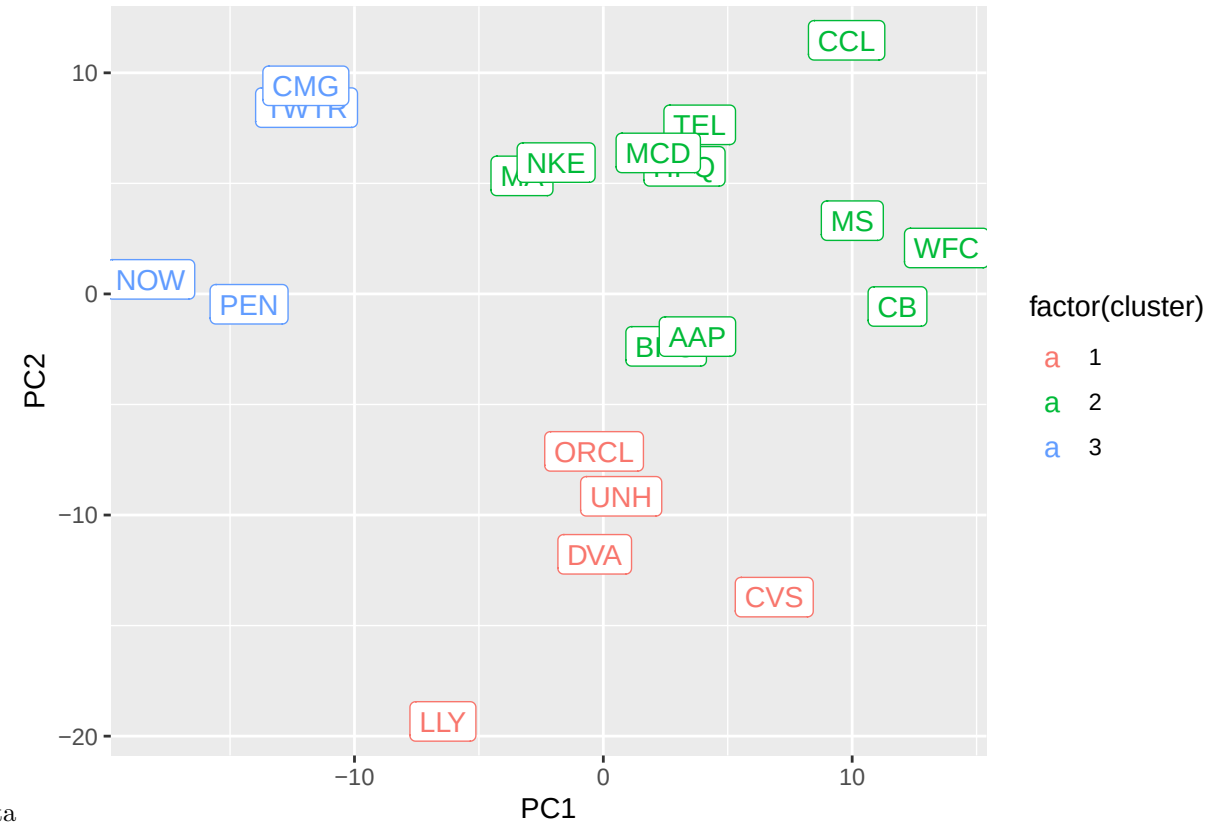




3. Cluster Result:

```
km3 <- kmeans(data_cluster,centers = 3)$cluster
km3
```

```
## TWTR  HPQ  TEL  NOW ORCL   MA  WFC   MS  BRO   CB  UNH  LLY  PEN  DVA  CVS  NKE
##    3    2    2    3    1    2    2    2    2    2    1    1    3    1    1    2
##  MCD  CMG  AAP  CCL
##    2    3    2    2
```



4. Visualize Data

5. Problem: (Conclusion & Interpretation)
   Are there clear clusters of stocks in our subset of the S&P500?
   Answer: Yes, there is three clear clusters of stocks.
   Do we find that different stocks of different sectors cluster together, or are clearly distinguished?
   Answer: A few different stocks of different sectors cluster together with common features and they forms three clusters. Two of them are clearly distinguished, and the other one needs more information to analyze.
   Recall Cluster Result:
   Cluster 1: Twitter (TWTR, Technology), Service Now (NOW, Technology), Penumbra (PEN, Health), Chipotle (CMG, Consumer Discretionary)
   Cluster 2: NKE (Nike, Consumer Discretionary), MCD (McDonald's, consumer Discretionary), AAP (Advance Auto Parts, Consumer Discretionary), CCL (Carnival, Consumer Discretionary), MA (MasterCard, Finance), BRO (Brown & Brown, Finance), MS (Morgan Stanley, Finance), WFC (Wells Fargo, Finance), CB (Chubb Limited, Finance), TEL (TE Connectivity, Technology), HPQ (HP, Technology)
   Cluster 3: LLY (Eli Lilly, Health), UNH (UnitedHealth, Health), DVA (DaVita, Health), CVS (CVS, Health), ORCL (Oracle, Technology)
   Detailed Analysis for the data:
   We have no idea about cluster 1 it may need more information to decided the common feature of stocks here.
   Cluster 2 includes all stocks from consumer discretionary category except Chipotle, all stocks from finance category, and two stocks from technology. Since consumer discretionary and finance can be both consider as economic components, we assume that this cluster is related to people's daily economical behavior. This feature is also shown by TE connectivity and HP. TE connectivity primary focus is on reliable design and enabling people to achieve their potential, helping to drive technological innovation in connected transportation data communications and smart homes, transforming the way people live, work and connect. HP mainly produces computers to the market.
   We assume that cluster 3 is highly related to health, because except for Oracle, other stocks in cluster

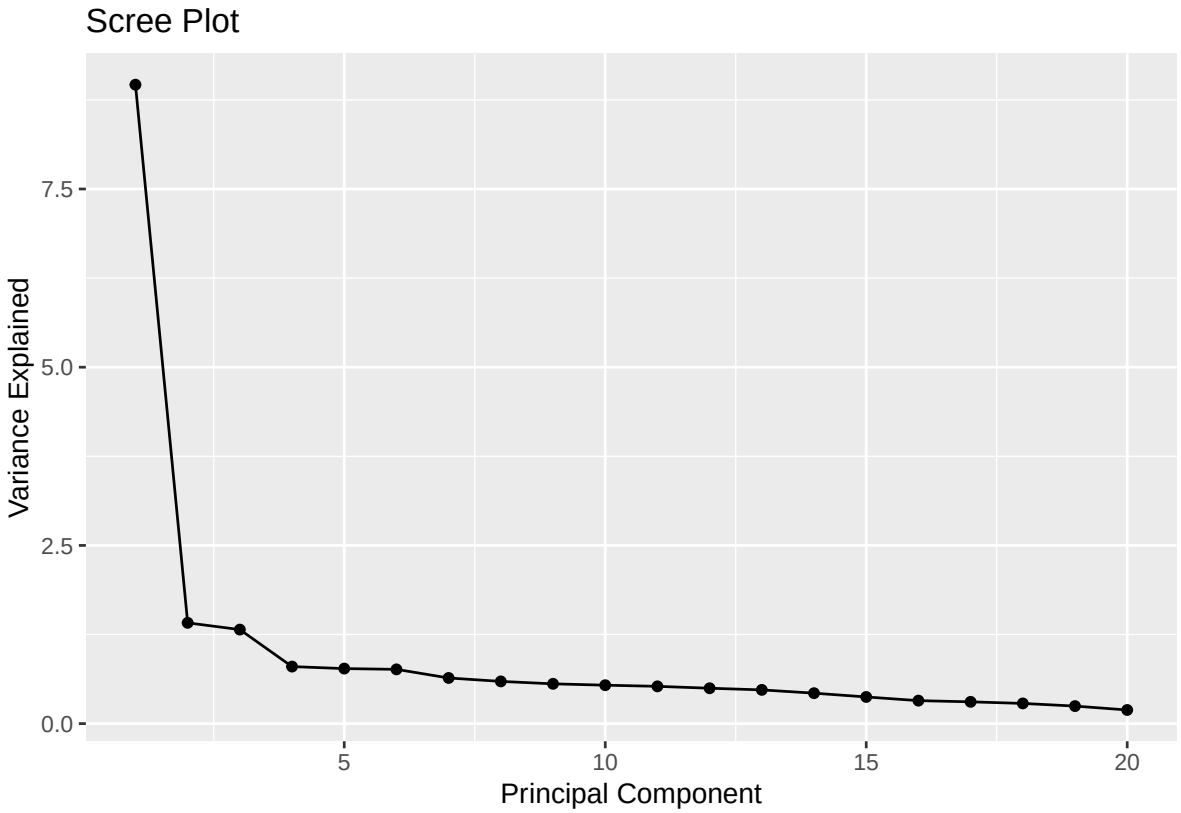
3 are all belong to health category, and Oracle focuses on database which may contain database related to medical field.

ICA

1. Each row name of data represents a date, each column name of data is a stock. Therefore, the data is a 1007 by 20 matrix.

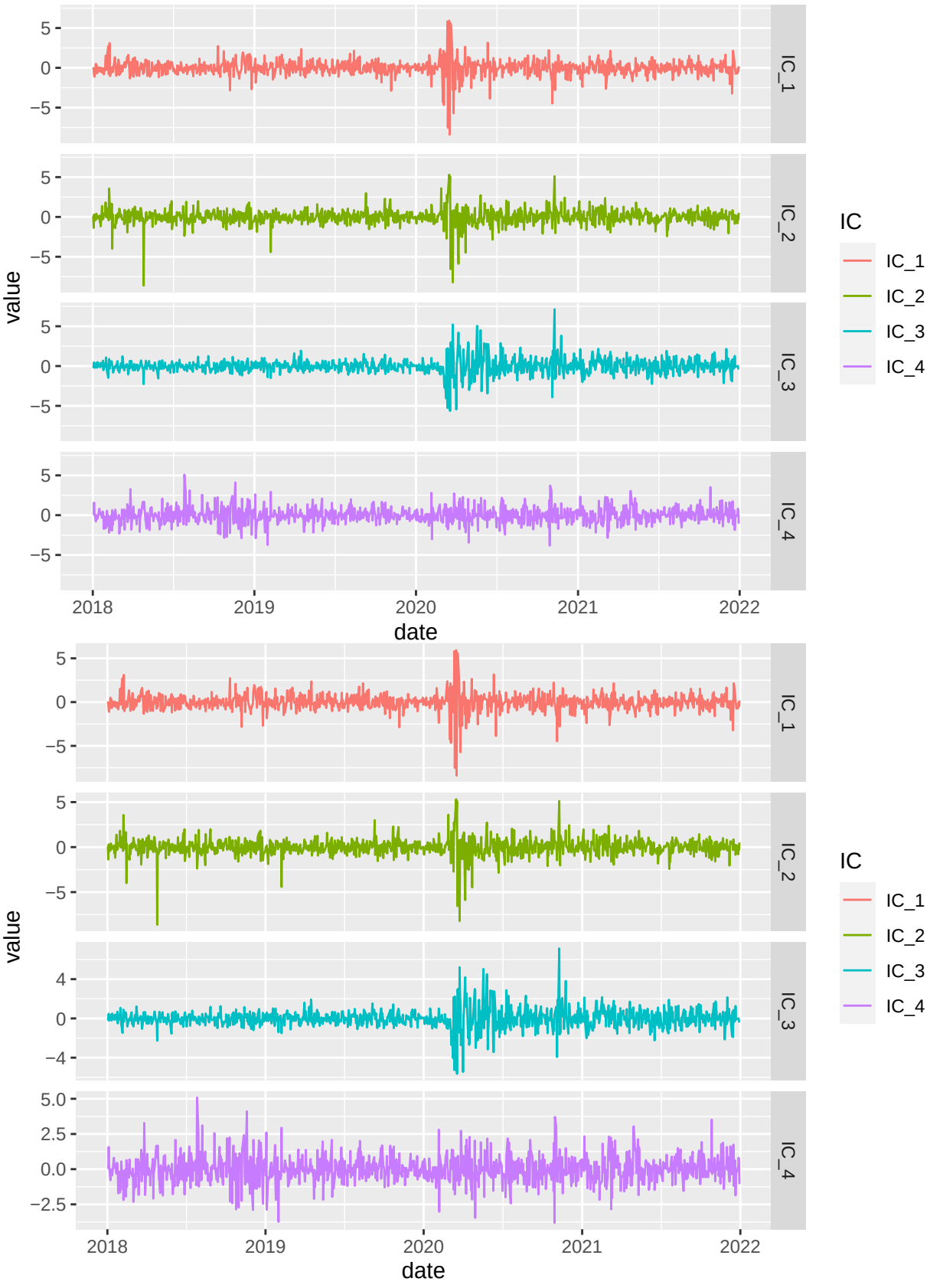
```
##                   TWTR         HPQ         TEL          NOW         ORCL
## 2018-01-02 -0.03474518 -0.04095581 -0.04001154 -0.075415097 -0.04636939
## 2018-01-03 -0.10944788  0.05754943  0.81873348  0.277495234  1.18897918
## 2018-01-04 -0.60888215  0.27354850  0.19603694  0.203219606  0.47906795
## 2018-01-05  0.38503247  0.38826718  1.11905614  0.266013249  0.27467477
## 2018-01-08  0.30404890  0.13284993  0.06718037  0.009211749  0.51483824
## 2018-01-09 -0.55597155 -0.09865305 -0.54430245 -0.031739367  0.24879572
##                    MA          WFC          MS         BRO          CB
## 2018-01-02 -0.05234512 -0.007206241 -0.04356739 -0.07478072 -0.03019724
## 2018-01-03  0.54515985  0.312301795  0.07362484  0.66105321  0.39135923
## 2018-01-04  0.56244163  0.512250522  0.62432774  0.10084708  0.18641481
## 2018-01-05  0.93278728  0.272626757 -0.06001439  0.53823175  0.06651023
## 2018-01-08  0.09131030 -0.477096403 -0.22453770 -0.27306996 -0.30123883
## 2018-01-09  0.01626562  0.140059941  0.29510232  0.44733417  0.83556895
##                   UNH         LLY         PEN         DVA         CVS
## 2018-01-02 -0.05480158 -0.07480879 -0.05595107 -0.03158289 -0.03394194
## 2018-01-03  0.48222213  0.20683917  0.02580954  0.21179294 -0.26716901
## 2018-01-04  0.16740212  0.15658396 -2.48289791  1.23800575  1.37884781
## 2018-01-05  0.92140196  0.56174303  0.18416813  0.40189557  2.33391443
## 2018-01-08 -0.94335174 -0.33832449  0.61637049  0.05054493 -0.49839374
## 2018-01-09  0.20030018 -0.11694503  1.22205693  1.10987859 -0.08905353
##                   NKE         MCD         CMG         AAP         CCL
## 2018-01-02 -0.06128978 -0.04128657 -0.08541396 -0.04971215  0.00332474
## 2018-01-03 -0.06940830 -0.30915876  2.18788602  0.37744553  0.05551060
## 2018-01-04 -0.09378846  0.40461427 -0.33920853  1.69208943 -0.01402991
## 2018-01-05  0.37764153  0.08682103  0.81711187  0.45210876 -0.18774524
## 2018-01-08  0.39812861 -0.08513082  0.53343057 -0.38214696 -0.08424672
## 2018-01-09 -0.42877569 -0.18381885  0.03184516 -0.43112201  0.55886643
```

2. Use Elbow Method on Factor Analysis to Determine Optimal Clusters: It seems that k=4 is resonable.



3. Visualize Data by 4 Independent Components:

```
##                   IC_1        IC_2        IC_3         IC_4       date
## 2018-01-02  0.06596282  0.08630711  0.02669927  0.030924434 2018-01-02
## 2018-01-03 -0.19322698 -1.37256193 -0.15987225  0.118854521 2018-01-03
## 2018-01-04 -0.91282821  0.29566321  0.47023206  1.575689829 2018-01-04
## 2018-01-05 -1.11062492 -0.47283124 -0.21199462  0.189502601 2018-01-05
## 2018-01-08  0.50355644 -0.36902647 -0.13947662 -0.787117732 2018-01-08
## 2018-01-09 -0.46256979  0.40637808  0.02903154 -0.007483776 2018-01-09
```



4.Problem: (Conclusions & Interpretation)

What are some independent factors that drive price or return of a group of stocks? (Independent component analysis)
Answer: we think the factors should be some big real world events, and the possible one to explain our data here is the pandamic of covid-19 happened in March of 2020.
Detailed Analysis for the data: Covid-19 started from March 2020 and from both plots (fix scaled one and free scaled one), we see that the amplitude of IC_1 IC_2 and IC_3 has suddenly changed a lot from that time, whereas the amplitude for IC_4 wasn't change much.So we assume that IC_4 is the noise term. Since the sudden change on the amplitude for IC_1 and IC_2 went back to the trend before the covid-19, so we consider the covid only caused a short term influence on these two independent components. Because the amplitude of IC_3 after the covid-19 was still obviously bigger than it has been before the covid-19, therefore we think that covid-19 has create a long term effect on IC_3.

Appendix

Savvas Giannaris - Multivariate Paired Comparison, Regression (see descripion), Report Organization Kassidy Chaikin - Factor Analysis, Multidimensional Scaling Ruihan Jiang - Cluster Analysis and Independent Component Analysis

Description: Initially, we were attempting to use multivariate regression to predict the return patterns of other stocks within NYSE sectors. Unfortunately, after many attempts I was unable to yield a model that provided valuable information. I could not return a model that had a any correlation to the response variables. With such inconclusive results, we felt that the rest of the report deserved further analysis and attention.