Oguzcan Adabuk
11/28/2020

MAT 599 Final Project
Principal Component Analysis for Stock Portfolio Management
by Oguzcan Adabuk

1. **Goal**

The primary goal of this paper is to implement Principal Component Analysis to analyze 30 stocks that make up the Dow Jones Industrial Average stock market index (ticker: ^DJI) and replicate this index. The secondary goal is to analyze stocks further and make inferences about different sectors and how they were affected by the events of year 2020.

2. **Data**

Daily closing returns for the last year of stocks that make up DJI has been downloaded from Yahoo Finance website. This data includes closing price for 252 trading days between November 27 2019 and November 27 2020 for the stocks below;
AXP,AMGN,AAPL,BA,CAT,CSCO,CVX,GS,HD,HON,IBM,INTC,JNJ,KO,JPM,MCD,MMM,MRK,MSFT,NKE,PG,TRV,UNH, CRM,VZ,V,WBA,WMT,DIS,DOW
Additionally, the *DJI* data for the last 252 days is also obtained.
In order to better compare these closing prices, the data is converted to percentage returns.

3. **Model**

All eigenvectors of a square symmetric matrix Q are orthogonal to each other, these eigenvectors define a space. A matrix E of normalized eigenvectors of a matrix Q as its columns, can take any vector and rotate into its eigenspace. Because E is a rotation matrix, transpose of E is equal to its inverse, therefore when a vector is rotated into eigenspace by multiplying by E, it can be rotated out of the eigenspace by multiplying it with E^T. The principal component analysis reduces dimensions by stretching data along axes that have more variation and getting rid of axes with small variation as these axes do not contain significant amount of information about the data. Stretching a vector along the axes is performed by multiplying the vector by a diagonal matrix that has the eigenvalues along its diagonal D. Therefore any square matrix Q can be decomposed as $Q=EDE^T$. Because variation along axes are proportional to how much information each axis contain, the eigendecomposition is applied to the covariance matrix of the centered data matrix to find the principal components.

The data matrix A is rotated so that the data lies along the directions of maximum variation. To perform this rotation, A is multiplied with a rotation matrix R^T.
$Y=R^TA$, R is chosen to make sure that the covariance matrix of Y is diagonal.
$Cov(Y) = Cov(P^TX) =$
$[\lambda_1\ 0\ 0\ \ldots\ 0]$
$[0\ {}_\lambda 2\ 0\ \ldots\ 0]$
$[\ldots\ ..\ .\ .\ .\ ..\ ..]$
$[0\ 0\ \ldots\ldots\lambda_N]$
$Cov(Y) = E[YY^T]$
$=E[(R^TY)(R^TX)^T]$

$=E[(R^TX)(X^TR)]$

$=E[(R^TX)(X^TR)], \quad E[R] = R$

$=R^TE(XX^T)R$

$=R^TCov(X)R,$

Because R is a rotation matrix, and inverting a rotation means that rotating it in the opposing direction by the same amount, $R^T=R^{-1}$.

$R\ Cov(Y) = RR^TCov(X)R=Cov(X)R$

Since $Cov(Y)$ is diagonal, and R can be written as a set of column vectors,

$RCov(Y)=[\lambda_1 r_1, \lambda_2 r_2,\ldots, \lambda_N r_N]$

$Z=Cov(X),$

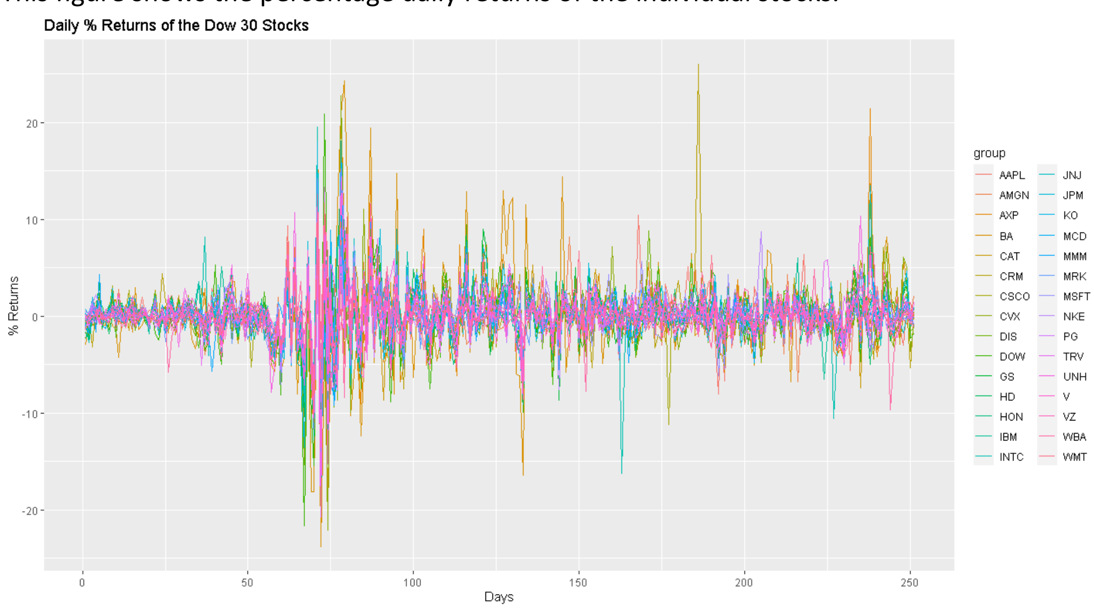
$\lambda r_i=Zr_i$ for each r_i, because λ is a column vector and Z is a matrix,

Matrix R doesn't actually rotate or twist Z but merely scales it. Therefore the eigenvectors and eigenvalues are obtained.

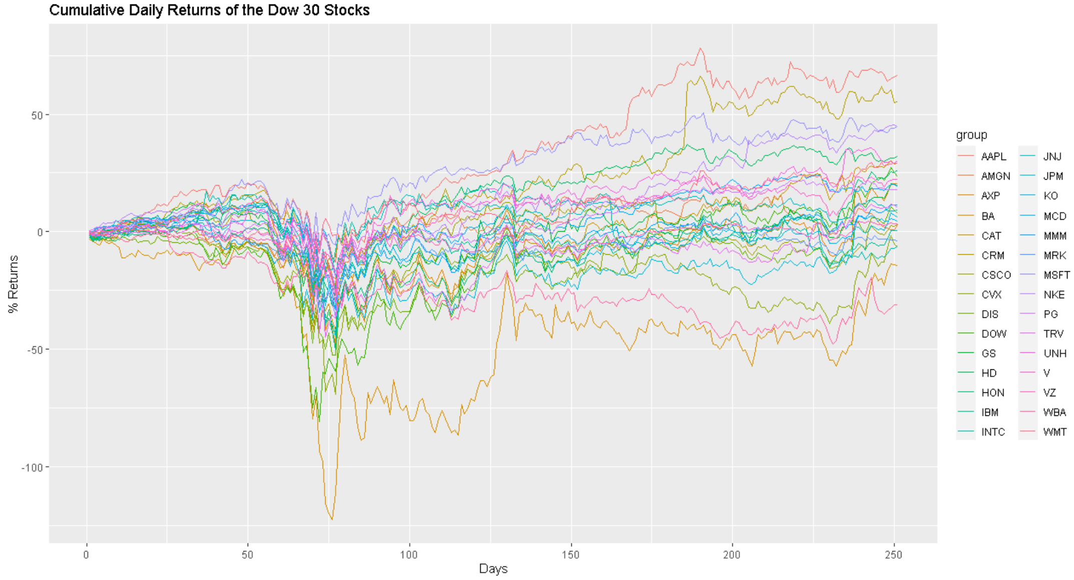
- Gather returns for each stock as column vectors and create a data matrix A. The daily returns make up the rows of A. For e.g. row one is the returns for all 30 stocks for the first day. The data matrix A has 251 rows and 30 columns. N=251, M=30. (We love 1 day for calculating percentage returns (S(T)-S(T-1)/S(T-1)) * 100.
- The data is centered by subtracting the mean of each column from column values. Centered data is placed in a new matrix B.
- The covariance matrix C is computed by $(1/N\text{-}1)\ B^TB$.
- Compute eigenvectors and eigenvalues of C.
- Magnitude of eigenvalues provide level of importance for each principal component, largest eigenvalue makes the associated principal component the one that containing most information.

4. Implementation

The returns of 30 stocks that make up Dow 30 can be hard to analyze. If there were more stocks this would be even more overwhelming. Therefore reduction of dimensionality is very useful to analyze these stocks.
This figure shows the percentage daily returns of the individual stocks.



Here you can see their cumulative returns of the individual 30 stocks:



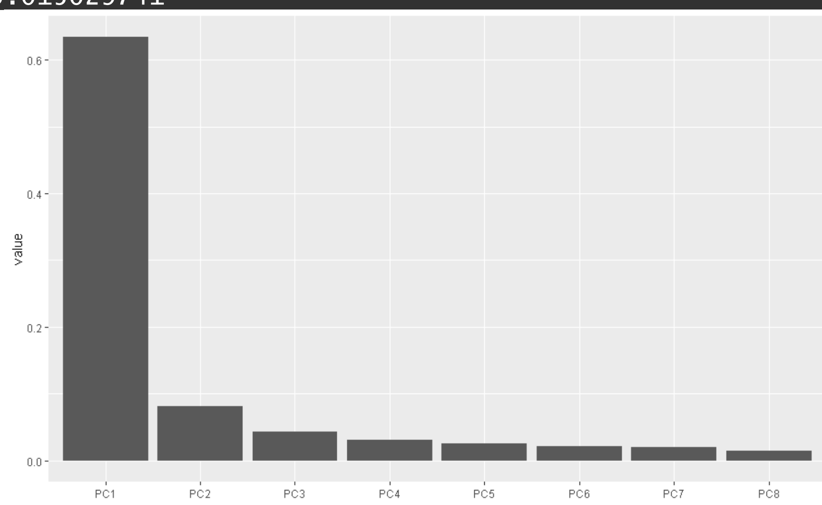
Principal Component Analysis applied to the stocks that make up Dow 30. The first principal component approximates the returns of DJI.

Eigenvectors and eigenvalues of the covariance matrix C are obtained to perform PCA. Eigenvalues of C gives us how much variance is packed by each principal component. Because our covariance matrix is 30x30, there are 30 eigenvalues. First eight are shown in figure below.

```
> eigen.val
 [1] 168.3762956  21.5994850  11.3983008   8.2172034   6.8290321   5.7281485   5.4106489
 [8] 3.9913487
```

Percentage of contribution of each principal component:

```
> eigen.val/sum(eigen.val)
 [1] 0.634034356 0.081334582 0.042921210 0.030942534 0.025715264 0.021569800 0.020374230
 [8] 0.015029741
```

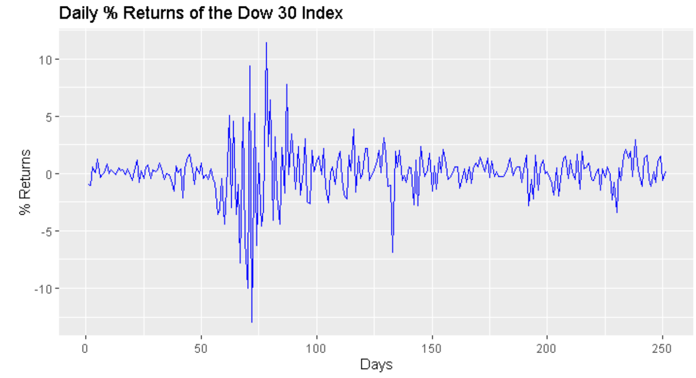
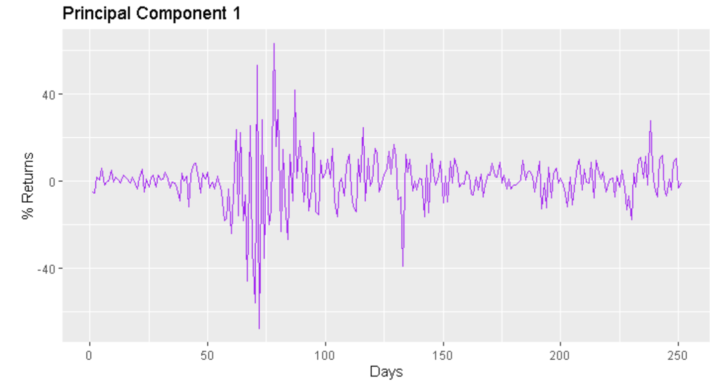
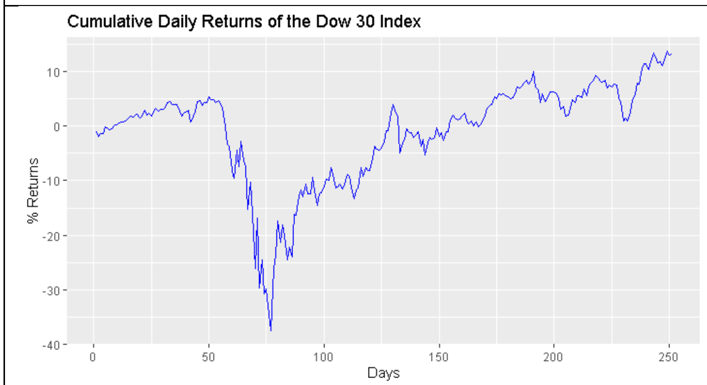
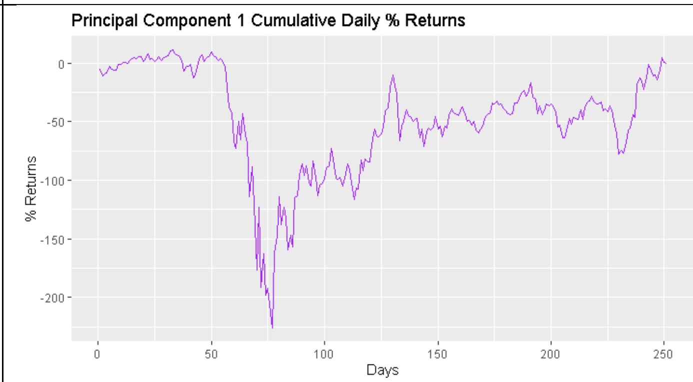


This screeplot tells us that more than 70% of the variation is packed within the first two principal components. Therefore using the first two components we can replicate Dow 30 accurately.

Principal Components are calculated by multiplying the centered data matrix B with the eigenvectors of the covariance matrix C.

The first eigenvector of C gives us how much each stock contributes to the principal component 1. Therefore all the stocks should have the same sign and coefficients for the first principal component should be positive.

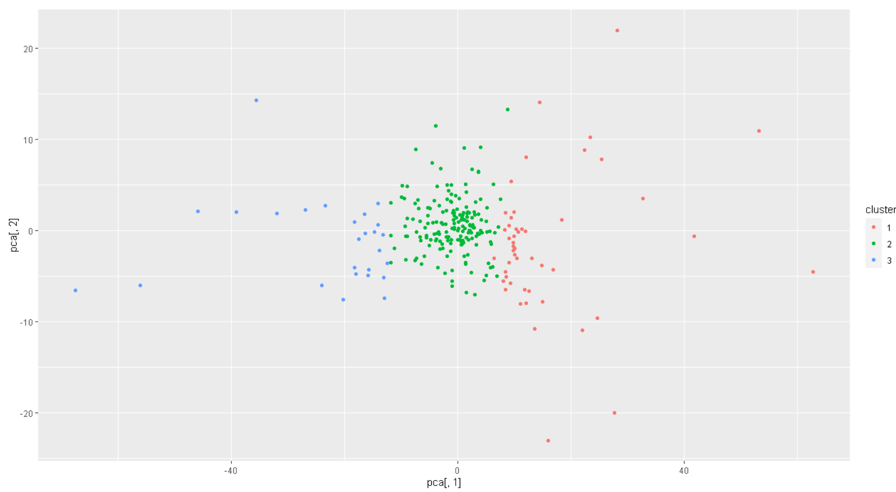
Let’s compare DJI returns and Principal Component 1 returns.

DJI	PC1
	
	

As it can be seen from the graphs, the PC1 approximates DJI closely. Further analysis can be performed to improve returns. In order to replicate this index, we can look at the coefficients of the first column of eigenvector matrix. This will show us the ratios of each stock we need to have in our portfolio to replicate DJI.

```
          Contribution Tickers                Sectors
4   0.335434978351458      BA            Industrials
1   0.276712507114268     AXP             Financials
7   0.249257109265766     CVX                 Energy
30  0.242328528634008     DOW        Basic Materials
15  0.236314136034833     JPM              Financial
8   0.227072734324167      GS              Financial
10  0.196546746950243     HON            Industrials
29  0.188576146294369     DIS Communication Services
22  0.188233606332374     TRV              Financial
26  0.188058384810624       V              Financial
12  0.186328069948546    INTC             Technology
23  0.182079154090808     UNH             Healthcare
5   0.179318619283976     CAT            Industrials
9   0.176126471201838      HD      Consumer Cyclical
11  0.169594491801716     IBM             Technology
19  0.165664404430487    MSFT             Technology
3   0.164321952797199    AAPL             Technology
6   0.161172115634416    CSCO             Technology
16  0.159417591608575     MCD      Consumer Cyclical
20  0.158861897176624     NKE      Consumer Cyclical
24  0.155986367880176     CRM             Technology
27  0.146030186799376     WBA             Healthcare
17  0.145631324939858     MMM            Industrials
14  0.136980237921026      KO     Consumer Defensive
2   0.119265005941437    AMGN             Healthcare
18  0.107375302525155     MRK             Healthcare
13  0.106893505514168     JNJ             Healthcare
21   0.10683466173253      PG     Consumer Defensive
25 0.0850267398633831      VZ Communication Services
28 0.0730136147924025     WMT     Consumer Defensive
```

Plotting principal component scores against each other gives us a cluster of returns with many outliers. Because there seems to be one large main cluster centered around 0, main thing can be interpreted is how different trends effected the returns in 2020, mainly being chaotic. Negative outliers can be attributed to COVID19 crisis occurring in march and negatively affecting the market for the next several months. Positive outliers can be attributed to those sectors that actually did well during the pandemic and packages announced by the treasury to save failing giants and stimulus packages.



5. Code

```
### 11/29/2020 Authored by Oguzcan Adabuk ###
require(ggplot2)

#Set the current directory to load the data
dn<-dirname(getSourceEditorContext()$path)
setwd(dn)

# Read Daily stock returns and return log returns
get_log_returns<-function(f){
  filename <- paste(f, ".csv", sep="")
  d<-read.csv(filename)
  d<-d$Close
  #d<-d[60:120]
  n<-length(d)
  #d.returns <- 100*log(d[-1]/d[-length(d)])
  d<- ((d[-1]-d[1:n-1])/d[1:n-1]) * 100
  return(d)
}

# Mean center columns of a given matrix
center_apply <- function(x) {
  apply(x, 2, function(y) (y - mean(y)))
}

# Stock tickers and sectors
tickers <-
c("AXP","AMGN","AAPL","BA","CAT","CSCO","CVX","GS","HD","HON","IBM","INTC","JNJ","KO","JPM","MCD","MMM","MRK","MSFT","NKE","PG","TRV","UNH","CRM","VZ","V","WBA","WMT","DIS","DOW")
```

```
sectors <- c("Financials", "Healthcare", "Technology", "Industrials", "Industrials", "Technology", "Energy",
"Financial", "Consumer Cyclical", "Industrials", "Technology", "Technology", "Healthcare", "Consumer Defensive",
"Financial", "Consumer Cyclical", "Industrials", "Healthcare", "Technology", "Consumer Cyclical", "Consumer
Defensive", "Financial", "Healthcare", "Technology", "Communication Services", "Financial", "Healthcare",
"Consumer Defensive", "Communication Services", "Basic Materials")

# Dow Jones Industrial Average Index daily log returns for the last 1 year
dji<-get_log_returns("DJI")


# Create the data matrix A
A <- list()
for(i in 1:length(tickers)){
  A[[tickers[i]]] <- get_log_returns(tickers[i])
}

xrows = length(A[[1]])
xcols = length(A)

# Create a new Matrix B from the centered columns
B<-matrix(sample(0, xcols*xrows, replace=T), nrow=xrows)
B<-center_apply(matrix(unlist(A), nrow=xrows))

# Computer the covariance matrix
C <- (1/(xrows-1)) * t(B) %*% B

# Compute Eigenvectors and Eigenvalues
C.e<-eigen(C)
eigen.vec <- C.e$vectors
eigen.vec[,1] <- eigen.vec[,1] * -1
#eigen.vec[,1] <- eigen.vec[,1]# * -1
eigen.val <- (C.e$values)

# Calculate principal components as linear combinations, store in PCA matrix
pca <- B %*% eigen.vec

# Ratios of stocks we need to keep to replicate DJI
top_cont <- data.frame(cbind(as.numeric(eigen.vec[,1]), tickers, sectors))
colnames(top_cont) <- c("Contribution", "Tickers", "Sectors")
top_cont[order(top_cont$Contribution, decreasing = TRUE),]

# Use built-in princomp() function to validate results from eigendecomposition above
xp <- princomp(matrix(unlist(A), nrow=xrows))

# Plot Contributions of each Principal Component
pcp<-eigen.val/sum(eigen.val)

dfev <- data.frame(
  name=c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8"),
  value=pcp[1:8]
)

# Barplot
```

```
ggplot(dfev, aes(x=name, y=value)) +
  geom_bar(stat = "identity")

# Plots
x<-c(1:xrows)
x_cs<-c(1:xrows)

for(i in 1:xcols){
  x<-cbind(x, A[[i]])
  x_cs<-cbind(x_cs, cumsum(A[[i]]))
}

cn<-append("x", tickers)
df<-data.frame(x)
colnames(df)<-cn

df_cs<-data.frame(x_cs)
colnames(df_cs)<-cn

n<-length(df$AXP)

# Plot % returns
df_g <- data.frame(x=df$x,
                y=c(df$AXP, df$AMGN, df$AAPL, df$BA, df$CAT, df$CSCO, df$CVX, df$GS, df$HD, df$HON, df$IBM,
df$INTC, df$JNJ, df$KO, df$JPM, df$MCD, df$MMM, df$MRK, df$MSFT, df$NKE, df$PG, df$TRV, df$UNH, df$CRM,
df$VZ, df$V, df$WBA, df$WMT, df$DIS, df$DOW),
                group=c(rep("AXP", n),rep("AMGN", n),rep("AAPL", n),rep("BA", n),rep("CAT", n),rep("CSCO",
n),rep("CVX", n),rep("GS", n),rep("HD", n),rep("HON", n),rep("IBM", n),rep("INTC", n),rep("JNJ", n),rep("KO",
n),rep("JPM", n),rep("MCD", n),rep("MMM", n),rep("MRK", n),rep("MSFT", n),rep("NKE", n),rep("PG",
n),rep("TRV", n),rep("UNH", n),rep("CRM", n),rep("VZ", n),rep("V", n),rep("WBA", n),rep("WMT", n),rep("DIS",
n),rep("DOW", n))
                )

ggplot(df_g, aes(x, y, col=group)) + geom_line() +
  labs(title="Daily % Returns of the Dow 30 Stocks", x="Days", y="% Returns")

# Plot cumulative returns
df_g_cs <- data.frame(x=df_cs$x,
                y=c(df_cs$AXP, df_cs$AMGN, df_cs$AAPL, df_cs$BA, df_cs$CAT, df_cs$CSCO, df_cs$CVX, df_cs$GS,
df_cs$HD, df_cs$HON, df_cs$IBM, df_cs$INTC, df_cs$JNJ, df_cs$KO, df_cs$JPM, df_cs$MCD, df_cs$MMM,
df_cs$MRK, df_cs$MSFT, df_cs$NKE, df_cs$PG, df_cs$TRV, df_cs$UNH, df_cs$CRM, df_cs$VZ, df_cs$V,
df_cs$WBA, df_cs$WMT, df_cs$DIS, df_cs$DOW),
                group=c(rep("AXP", n),rep("AMGN", n),rep("AAPL", n),rep("BA", n),rep("CAT", n),rep("CSCO",
n),rep("CVX", n),rep("GS", n),rep("HD", n),rep("HON", n),rep("IBM", n),rep("INTC", n),rep("JNJ", n),rep("KO",
n),rep("JPM", n),rep("MCD", n),rep("MMM", n),rep("MRK", n),rep("MSFT", n),rep("NKE", n),rep("PG",
n),rep("TRV", n),rep("UNH", n),rep("CRM", n),rep("VZ", n),rep("V", n),rep("WBA", n),rep("WMT", n),rep("DIS",
n),rep("DOW", n))
)

ggplot(df_g_cs, aes(x, y, col=group)) + geom_line() +
  labs(title="Cumulative Daily Returns of the Dow 30 Stocks", x="Days", y="% Returns")

# Plot DJI
```

```
dfj<-data.frame(x=1:xrows, y=dji)
ggplot(dfj, aes(x=x, y=y)) + geom_line(color="blue")+
  labs(title="Daily % Returns of the Dow 30 Index", x="Days", y="% Returns")

# Plot DJI cumulative sum
dfp<-data.frame(x=1:xrows, y=cumsum(dji))
ggplot(dfp, aes(x=x, y=y)) + geom_line(color="blue")+
  labs(title="Cumulative Daily Returns of the Dow 30 Index", x="Days", y="% Returns")

# Plot PC1
dfpc<-data.frame(x=1:xrows, y=pca[,1])
ggplot(dfpc, aes(x=x, y=y)) + geom_line(color="purple")+
  labs(title="Principal Component 1", x="Days", y="% Returns")

# Plot PC1 cumulative sum
dfpc<-data.frame(x=1:xrows, y=cumsum(pca[,1]))
ggplot(dfpc, aes(x=x, y=y)) + geom_line(color="purple")+
  labs(title="Principal Component 1 Cumulative Daily % Returns", x="Days", y="% Returns")

# Plot PC2
dfpc<-data.frame(x=1:xrows, y=pca[,2])
ggplot(dfpc, aes(x=x, y=y)) + geom_line(color="green")+
  labs(title="Principal Component 2", x="Days", y="% Returns")

# Plot PC2 cumulative sum
dfpc<-data.frame(x=1:xrows, y=cumsum(pca[,2]))
ggplot(dfpc, aes(x=x, y=y)) + geom_line(color="green")+
  labs(title="Principal Component 2 Cumulative Daily % Returns", x="Days", y="% Returns")



# Plot Score plot
stock_cluster <- data.frame(A)
cls <- kmeans(x=stock_cluster, centers=3)
stock_cluster$cluster <- as.character(cls$cluster)
head(stock_cluster)

ggplot() +
  geom_point(data = stock_cluster,
           mapping = aes(x = pca[,1],
                  y = pca[,2],
                  colour = cluster))
```

6. References

- PRINCIPAL COMPONENT ANALYSIS FOR STOCK PORTFOLIO MANAGEMENT, Giorgia Pasini, International Journal of Pure and Applied Mathematics Volume 115 No. 1 2017, 153-167
- Stock Market Analytics with PCA From Principal Component Analysis to Capital Asset Pricing, Yao Lei Xu,
- Machine Learning An Algorithmic Perspective 2nd Ed., Ch. 6 , Stephen Marsland
- Process Improvement Using Data, Kevin Dunn