Jahangirnagar University

Department of Statistics

Masters in Applied Statistics and Data Science (ASDS)

Spring 2023

Assignment

Course Code: WM_ASDS06
Course Title: Multivariate Analysis

Submitted To

Prof. Dr. Tapati Basak

Submitted By

Name : Sayed Hossain Khan

Student ID : 20229068

Section : B

Batch : 9^{th}

 # At First Include the necessary library
library(ggplot2)
library(car)
library(matrixcalc)
library(corrplot)

 # Set Working Directory
setwd("E:/Essentials/Jahangirnagar University/Semester_02_Lecture/PM-ASDS06 Multivariate Analysis/Assignment")

 # Load data in dataframe
data <- read.table("T12-4.DAT")

 # class() function in R is used to return the values of the class attribute
class(data)

```
> class(data)
[1] "data.frame"
```

 # Let's check first few value using head()
head(data)

```
> head(data)
    V1   V2  V3   V4   V5    V6   V7    V8       V9
1 1.06  9.2 151 54.4  1.6  9077  0.0 0.628  Arizona
2 0.89 10.3 202 57.9  2.2  5088 25.3 1.555   Boston
3 1.43 15.4 113 53.0  3.4  9212  0.0 1.058  Central
4 1.02 11.2 168 56.0  0.3  6423 34.3 0.700   Common
5 1.49  8.8 192 51.2  1.0  3300 15.6 2.044 Consolid
6 1.32 13.5 111 60.0 -2.2 11127 22.5 1.241  Florida
```

 # For checking how many rows and column in the data set dim() function is used.
Dim(data)

```
> dim(data)
[1] 22  9
```

 # str() is used to show the structure of the object
str(data)

```
> str(data)
'data.frame':	22 obs. of  9 variables:
 $ V1: num  1.06 0.89 1.43 1.02 1.49 1.32 1.22 1.1 1.34 1.12 ...
 $ V2: num  9.2 10.3 15.4 11.2 8.8 13.5 12.2 9.2 13 12.4 ...
 $ V3: int  151 202 113 168 192 111 175 245 168 197 ...
 $ V4: num  54.4 57.9 53 56 51.2 60 67.6 57 60.4 53 ...
 $ V5: num  1.6 2.2 3.4 0.3 1 -2.2 2.2 3.3 7.2 2.7 ...
 $ V6: int  9077 5088 9212 6423 3300 11127 7642 13082 8406 6455 ...
 $ V7: num  0 25.3 0 34.3 15.6 22.5 0 0 0 39.2 ...
 $ V8: num  0.628 1.555 1.058 0.7 2.044 ...
 $ V9: chr  "Arizona" "Boston" "Central" "Common" ...
```

 # Here column names are defined in a variable
(collected: https://www.solver.com/hierarchical-clustering-example)
 # 22 U.S. public utility companies for the year 1975
 # V1: Fixed - charge covering ration (income/debt)
 # V2: Rate of return of capital
 # V3: Cost per KW capacity in place
 # V4: Cost per KW capacity in place
 # V5: Peak KWH demand growth from 1974 to 1975
 # V6: Sales (KWH use per year)
 # V7: Percent Nuclear
 # V8: Total

 # Assign the column name of dataset

```
thislist <- list("Fixed - charge covering ration (income/debt)",
                 "Rate of return of capital",
                 "Cost per KW capacity in place",
                 "Cost per KW capacity in place",
                 "Peak KWH demand growth from 1974 to 1975",
                 "Sales (KWH use per year)",
                 "Percent Nuclear",
                 "Total")
```

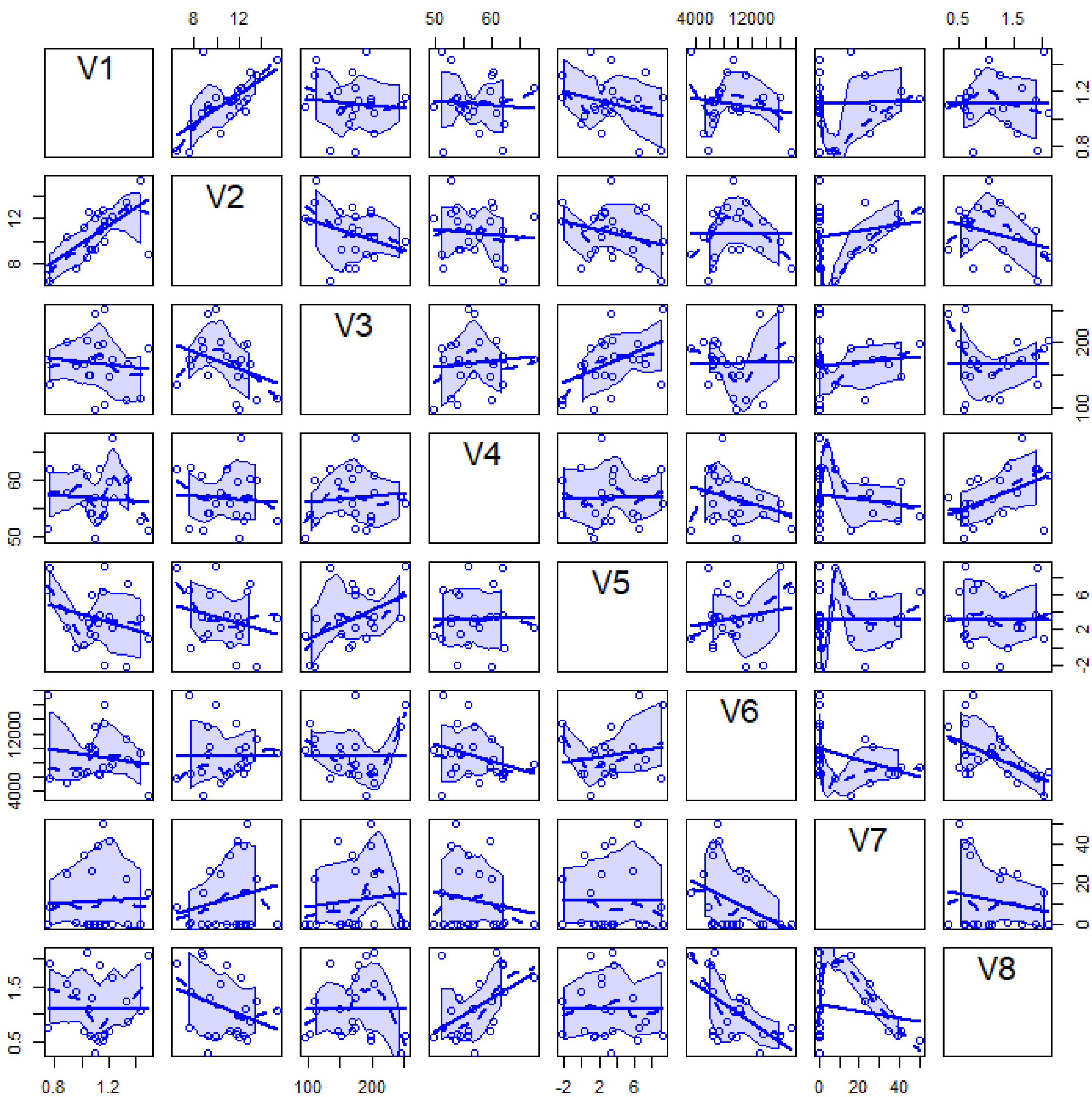
Question 1: Plot the raw data and make comment on the characteristics.

Answer:

Here are solve plot that are used to visualize the data.

 # Plot 01

```
dev.new(width = 15,    # Create new plot window
        height = 15,
        noRStudioGD = TRUE)
scatterplotMatrix(data[1:8], diagonal = F)
```

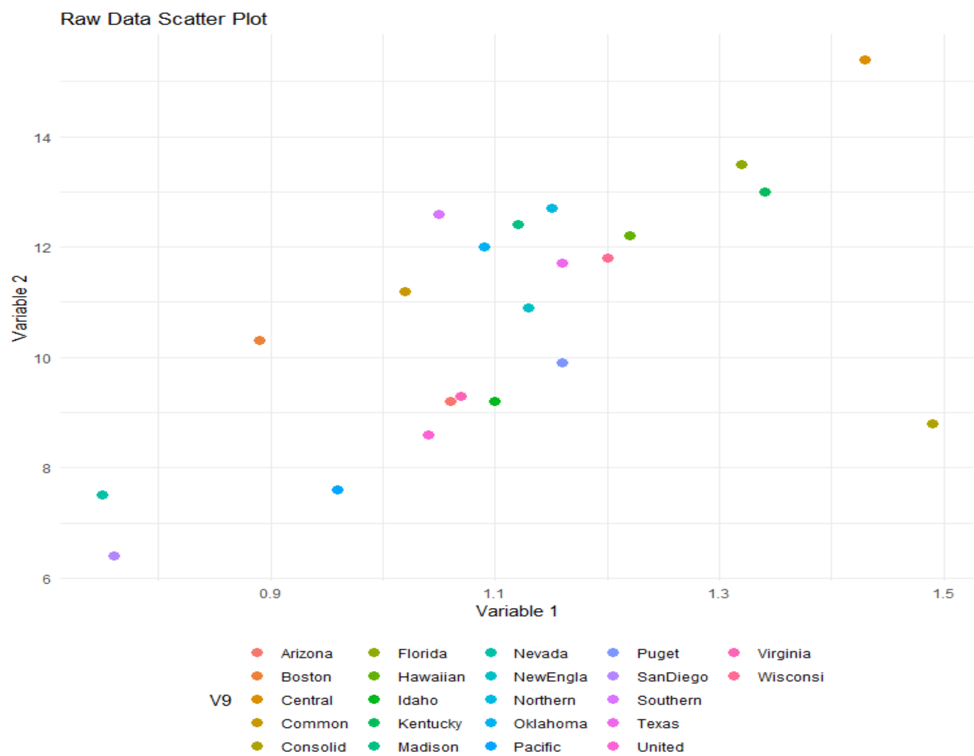


Interpretation:
Variable V1 shows a weakly negative relationship with V3, but a somewhat positive relationship with V2.
Variable V2 and V3 have a significant negative relationship and a somewhat favorable relationship, respectively. Additionally, it only slightly affects other variables.
Variable V3 is strongly correlated with variable V2 and somewhat correlated with variable V1. It only slightly positively affects other variables.

Weak to moderate relationships exist between variables V4, V5, and V6. The relationships between variables V7 and V8 are moderately positive but weak with respect to other variables.

 # Plot 02

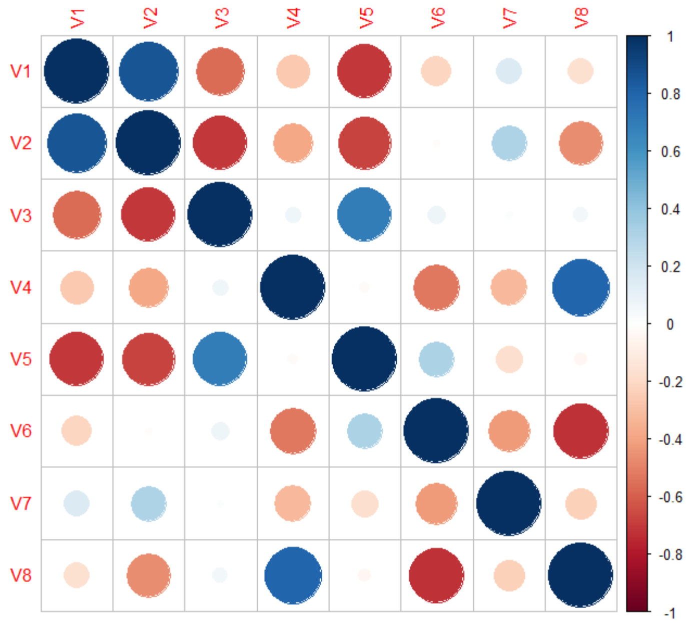
```
ggplot(data, aes(x = V1, y = V2, color = V9)) +
  geom_point(size = 3) +
  labs(x = "Variable 1", y = "Variable 2", title = "Raw Data Scatter Plot") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Interpretation:
This plot shows the Fixed - charge covering ration (income/debt) (V1) and Rate of return of capital (V2) of 22 U.S. public utility firms

 # Plot-3

co_data<-cor(data[,1:8])
corrplot(cor(co_data))

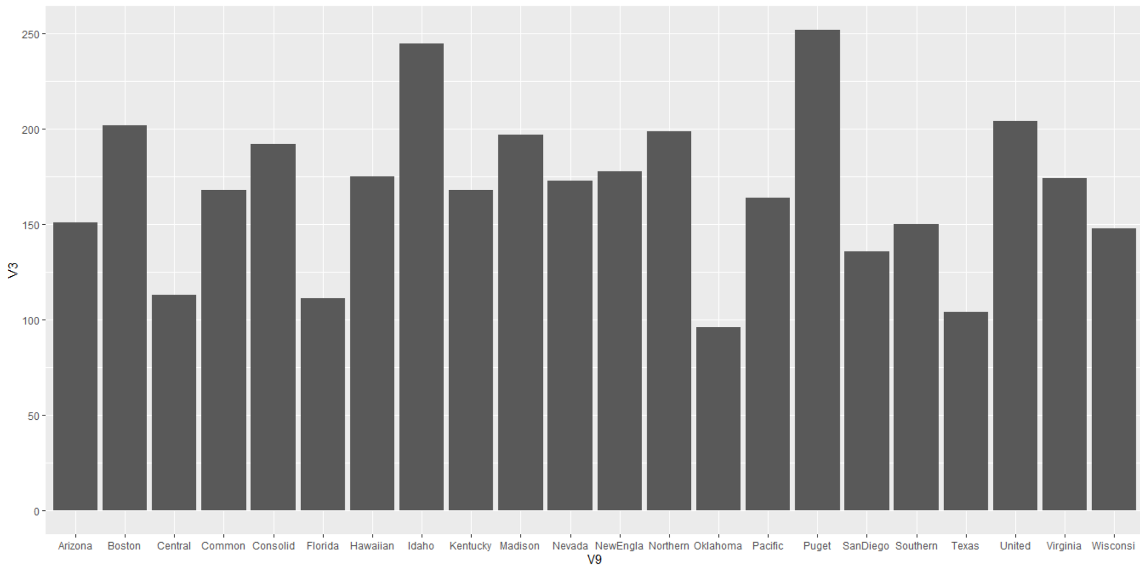


Interpretation:

The correlation matrix and correlation plot reveal relationships between variables in a dataset. The matrix contains coefficients ranging from -1 to 1, with -1 indicating strong negative relationships and 1 indicating strong positive relationships. The plot visually represents these correlations, revealing clusters or groups with high correlations, suggesting potential relationships or dependencies.

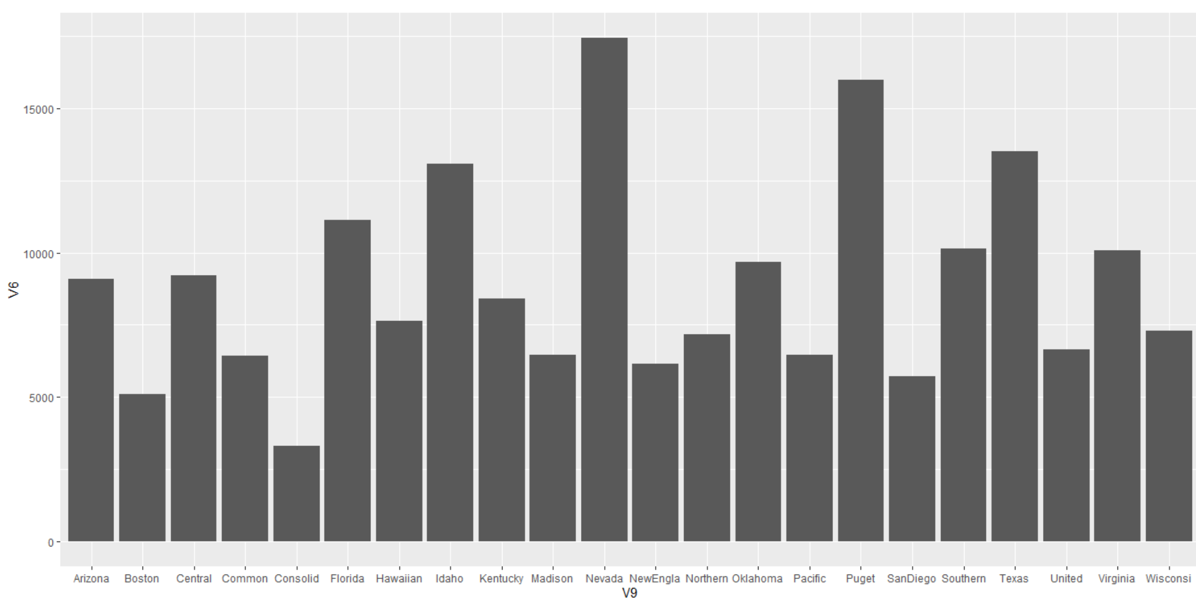
 # Plot 04

```
dev.new(width = 30,   # Create new plot window
        height = 15,
        noRStudioGD = TRUE)

ggplot(data, aes(x=V9, y=V3)) +
  geom_bar(stat = "identity")
```



Interpretation:
The bar graph is a visual representation of the cost per KW capacity distribution among 22 U.S. public utility firms in 1975.

```
ggplot(data, aes(x=V9, y=V6)) +
  geom_bar(stat = "identity")
```



Interpretation:
The bar graph is a visual representation of the Sales (KWH use per year) among 22 U.S. public utility firms in 1975.

Question 2: Obtain the summary measures and hence visualize them if necessary.

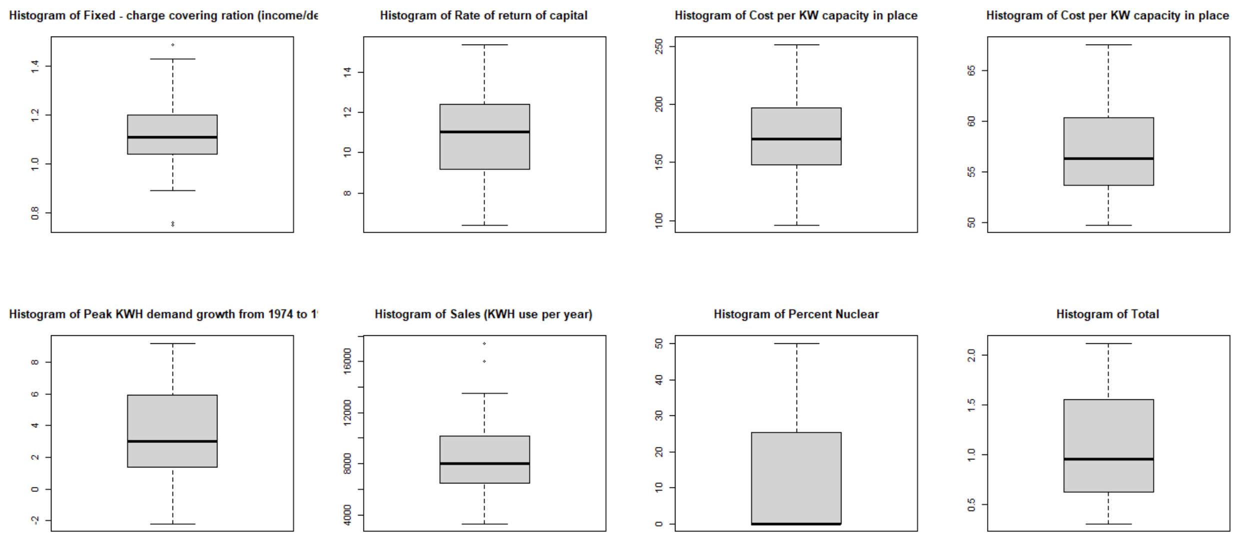
summary(data[,1:8])

```
       V1               V2               V3               V4               V5
 Min.   :0.750   Min.   : 6.40   Min.   : 96.0   Min.   :49.80   Min.   :-2.200
 1st Qu.:1.042   1st Qu.: 9.20   1st Qu.:148.5   1st Qu.:53.77   1st Qu.: 1.450
 Median :1.110   Median :11.05   Median :170.5   Median :56.35   Median : 3.000
 Mean   :1.114   Mean   :10.74   Mean   :168.2   Mean   :56.98   Mean   : 3.241
 3rd Qu.:1.190   3rd Qu.:12.35   3rd Qu.:195.8   3rd Qu.:60.30   3rd Qu.: 5.350
 Max.   :1.490   Max.   :15.40   Max.   :252.0   Max.   :67.60   Max.   : 9.200
       V6               V7              V8               V9
 Min.   : 3300   Min.   : 0.0   Min.   :0.309   Length:22
 1st Qu.: 6458   1st Qu.: 0.0   1st Qu.:0.630   Class :character
 Median : 8024   Median : 0.0   Median :0.960   Mode  :character
 Mean   : 8914   Mean   :12.0   Mean   :1.103
 3rd Qu.:10128   3rd Qu.:24.6   3rd Qu.:1.516
 Max.   :17441   Max.   :50.2   Max.   :2.116
```

Interpretation:

The dataset contains 22 observations of 9 variables related to U.S. public utility companies. The summary provides insights into the minimum, maximum, quartiles, mean, and count of each variable, offering a comprehensive overview of the data distribution. The variables include financial and operational metrics such as fixed-charge covering ratio, rate of return, cost per KW capacity, peak KWH demand growth, sales, percent nuclear, and location.

```
# Boxplot
dev.new(width = 30,   # Create new plot window
        height = 15,
        noRStudioGD = TRUE)

par(mfrow=c(2,4))

for (i in 1:8)
{
  boxplot(data[i],
         main=paste("Histogram of", thislist[i],sep=" "))
}
```



Interpretation:

The code creates boxplots for each variable in the dataset, displaying the distribution of values. The boxplots show median values, interquartile ranges, whiskers, and outliers. These plots provide insights into variable distribution and variability, enabling comparisons and identifying potential outliers. Variables V5 and V6 contain outliers, indicating potential differences in data.

Question 3: Check the positive definite property of the variance-covariance matrix.

cov(data[1:8])

```
              V1          V2           V3            V4          V5
V1   0.034044372   0.2661299   -0.7812554 -6.752165e-02  -0.14908009
V2   0.266129870   5.0357576  -32.1259740 -8.643723e-01  -1.82012987
V3  -0.781255411 -32.1259740 1696.7272727  1.843290e+01  55.92077922
V4  -0.067521645  -0.8643723   18.4329004  1.990184e+01   0.46573593
V5  -0.149080087  -1.8201299   55.9207792  4.657359e-01   9.72348485
V6 -99.346385281 -76.6160173 4092.5151515 -4.560037e+03 1952.87424242
V7   0.138809524   7.9676190   79.3095238 -1.229762e+01  -1.00142857
V8  -0.001372165  -0.4088848    0.1195758  1.204446e+00  -0.01236926
              V6            V7            V8
V1 -9.934639e+01  1.388095e-01 -1.372165e-03
V2 -7.661602e+01  7.967619e+00 -4.088848e-01
V3  4.092515e+03  7.930952e+01  1.195758e-01
V4 -4.560037e+03 -1.229762e+01  1.204446e+00
V5  1.952874e+03 -1.001429e+00 -1.236926e-02
V6  1.260239e+07 -2.227602e+04 -1.106557e+03
V7 -2.227602e+04  2.819686e+02 -1.728324e+00
V8 -1.106557e+03 -1.728324e+00  3.092451e-01
```

Interpretation:
The covariance matrix from R code cov(data[1:8]) shows positive linear relationships between variables V1 and V2, with a 0.266 positive covariance. Negative linear relationships were observed between V1 and V3, with a -0.781 negative covariance and -32.126 negative covariance. The diagonal elements represent the variances of each variable, while off-diagonal elements represent the covariances between pairs of variables. The matrix provides insights into the variability of each variable.

cor(data[1:8])

```
            V1          V2           V3          V4           V5           V6
V1  1.00000000  0.642744766 -0.102793192 -0.08203019 -0.259111089 -0.151671159
V2  0.64274477  1.000000000 -0.347550467 -0.08634194 -0.260111168 -0.009617468
V3 -0.10279319 -0.347550467  1.000000000  0.10030926  0.435367718  0.027987098
V4 -0.08203019 -0.086341943  0.100309264  1.00000000  0.033479746 -0.287935594
V5 -0.25911109 -0.260111168  0.435367718  0.03347975  1.000000000  0.176415568
V6 -0.15167116 -0.009617468  0.027987098 -0.28793559  0.176415568  1.000000000
V7  0.04480188  0.211444212  0.114661857 -0.16416254 -0.019125318 -0.373689523
V8 -0.01337310 -0.327655318  0.005220183  0.48550006 -0.007133152 -0.560526327
            V7           V8
V1  0.04480188 -0.013373101
V2  0.21144421 -0.327655318
V3  0.11466186  0.005220183
V4 -0.16416254  0.485500063
V5 -0.01912532 -0.007133152
V6 -0.37368952 -0.560526327
V7  1.00000000 -0.185085916
V8 -0.18508592  1.000000000
```

Interpretation:
The correlation matrix displays pairwise correlations between variables in the data. It indicates a moderate positive correlation (0.64) between variables V1 and V2, indicating higher values in V1 are associated with higher values in V2. A moderate positive correlation (0.44) exists between variables V3 and V5, indicating higher values in V3 are associated with higher values in V5. A strong negative correlation (-0.56) exists between variables V6 and V8, indicating an inverse relationship. Other correlations show weak or negligible associations. Overall, the correlation matrix provides insights into relationships among variables, highlighting which pairs are positively correlated, negatively correlated, or unrelated.

```
# Absolute Value
s<-as.matrix(abs(data[1:8]))
class(s)
dim(s)

# Sigma Transformation
Sigma=t(s)%*%s
Sigma
```

is.symmetric.matrix(as.matrix(Sigma))
is.positive.definite(as.matrix(Sigma))

Interpretation:

The Sigma transformation calculates the covariance matrix, Sigma, based on the data matrix's absolute values. The resulting matrix provides information on relationships and variances among variables. The 'is.symmetric.matrix' function checks if Sigma is symmetric, meaning the correlation between variables remains constant regardless of their order. The 'is.positive.definite' function checks if Sigma is positive definite, indicating positive eigenvalues and well-behaved matrix. By performing the Sigma transformation and evaluating its symmetry and positive definiteness, statistical analysis can assess the validity and properties of the covariance matrix.

Question 4: Perform eigenvalue decomposition and report the results.

r1 <- eigen(Sigma)

r1$values # lamda sign (λ)

```
> r1$values # lamda sign (λ)
[1] 2.013383e+09 1.245739e+05 5.356062e+03 3.234548e+03 1.281462e+02 8.055365e+01
[7] 2.637460e+00 3.581538e-01
```

Interpretation:

The Sigma has eigenvalues representing the variance explained by each principal component. The largest eigenvalue is 2.013383e+09, indicating the first principal component explains significant variance in the data. The second eigenvalue is 1.245739e+05, contributing less to overall variability. The third eigenvalue is 5.356062e+03, indicating a decrease in variance explained. The fourth eigenvalue is 3.234548e+03, contributing less variance than the previous eigenvalues. The fifth eigenvalue is 1.281462e+02, indicating a relatively small amount of variance explained. The sixth eigenvalue is 8.055365e+01, contributing less overall variability. The seventh eigenvalue is 2.637460e+00, indicating a further decrease in variance explained. The eighth principal component has the least variance explained at 3.581538e-01. The eigenvalues provide insight into the relative importance of each principal component in explaining the variability in the data.

r1$vectors

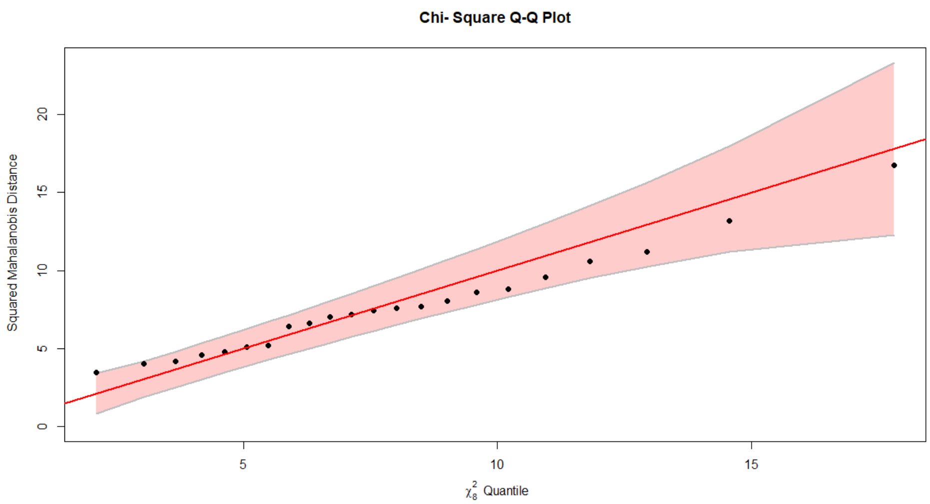
```
> r1$vectors
              [,1]         [,2]         [,3]         [,4]          [,5]
[1,] -1.075006e-04 -0.004757275 -0.005630282 -0.016850964 -0.0395382221
[2,] -1.045152e-03 -0.039463083 -0.101561081 -0.179864638 -0.5037855913
[3,] -1.642756e-02 -0.957556068  0.186259875  0.217882474 -0.0233994074
[4,] -5.503282e-03 -0.251675159 -0.185659839 -0.927464996  0.1189481584
[5,] -3.892474e-04 -0.010367467 -0.001783652  0.023505263  0.8523761736
[6,] -9.998488e-01  0.017289126 -0.001032556  0.001483597 -0.0000994445
[7,] -9.367698e-04 -0.133036816 -0.959404251  0.240671291  0.0247203900
[8,] -9.588995e-05 -0.007799397  0.005528696 -0.034975672  0.0527670489
              [,6]          [,7]          [,8]
[1,]  0.0555802722  1.337209e-01  9.884978e-01
[2,]  0.8334369001  3.943398e-02 -7.618139e-02
[3,]  0.0100605349  5.205594e-04 -1.407030e-03
[4,] -0.1604888056 -4.538418e-02  1.841124e-03
[5,]  0.5215944709 -2.563649e-02  8.574516e-03
[6,] -0.0003057868  9.173754e-05 -5.313298e-06
[7,] -0.0567078483  9.670186e-03  8.669369e-04
[8,] -0.0340474693  9.888139e-01 -1.303410e-01
```

Interpretation:

The eigenvectors in the data represent various directions of variation, with each column corresponding to a specific eigenvector. Eigenvector 1 is the most influential, with high coefficients for variables V9 and V8. Eigenvector 2 focuses on variables V6 and V3, while eigenvectors 3 and 4 show distinct patterns related to V2 (Rate of return of capital), V4 (Cost per KW capacity in place), and V6 (Sales and Percent Nuclear). Eigenvectors 5, 6, 7, and 8 provide additional directions of variation, but their coefficients do not show strong patterns or specific interpretations. Overall, the eigenvectors offer insights into the underlying structure and patterns of variation in the data.

Question 5: Check the multivariate normality of the dataset and take necessary steps if the data is non-normal.

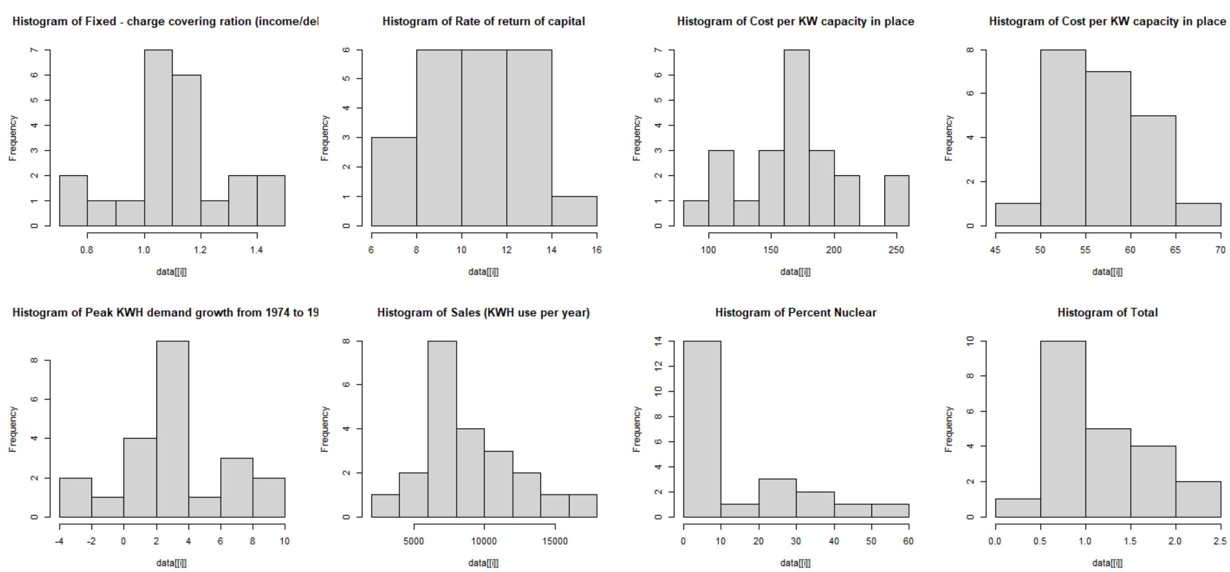
```
dev.new(width = 30,   # Create new plot window
        height = 15,
        noRStudioGD = TRUE)

# Chi Square Plot
cqplot(data[,1:8], main = "Chi- Square Q-Q Plot")
```



Interpret:

The code generates Q-Q plots for variables in columns 1 to 8 of the dataset. The plots assess if the data follows a chi-square distribution. Linearity suggests adherence, while deviations indicate departure from the expected distribution.

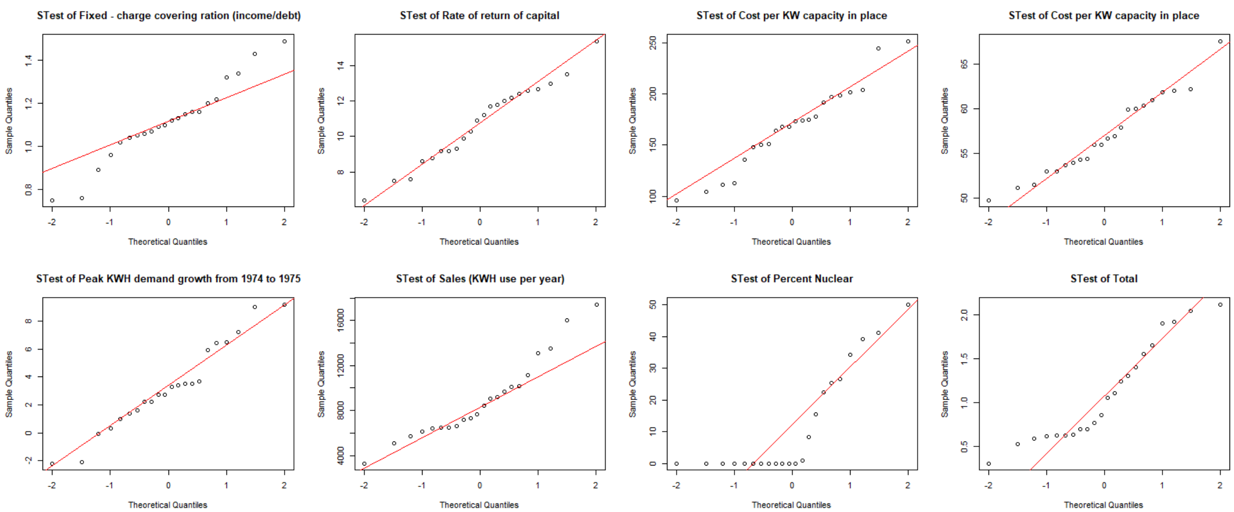
```
# Histogram
par(mfrow=c(2,4))
for (i in 1:8)
{
 hist(data[[i]], main=paste("Histogram of", thislist[i],sep=" "))
}
```



Interpretation:

A positively skewed distribution is observed in variable V7. But other variable are approximately normally distributed.

```
# Shapiro Wilk Test

for (i in 1:8)
{
  shapiro.test(data[[i]])
  qqnorm(data[[i]], main=paste("STest of", thislist[i],sep=" "))
  qqline(data[[i]],col="red")
}
```



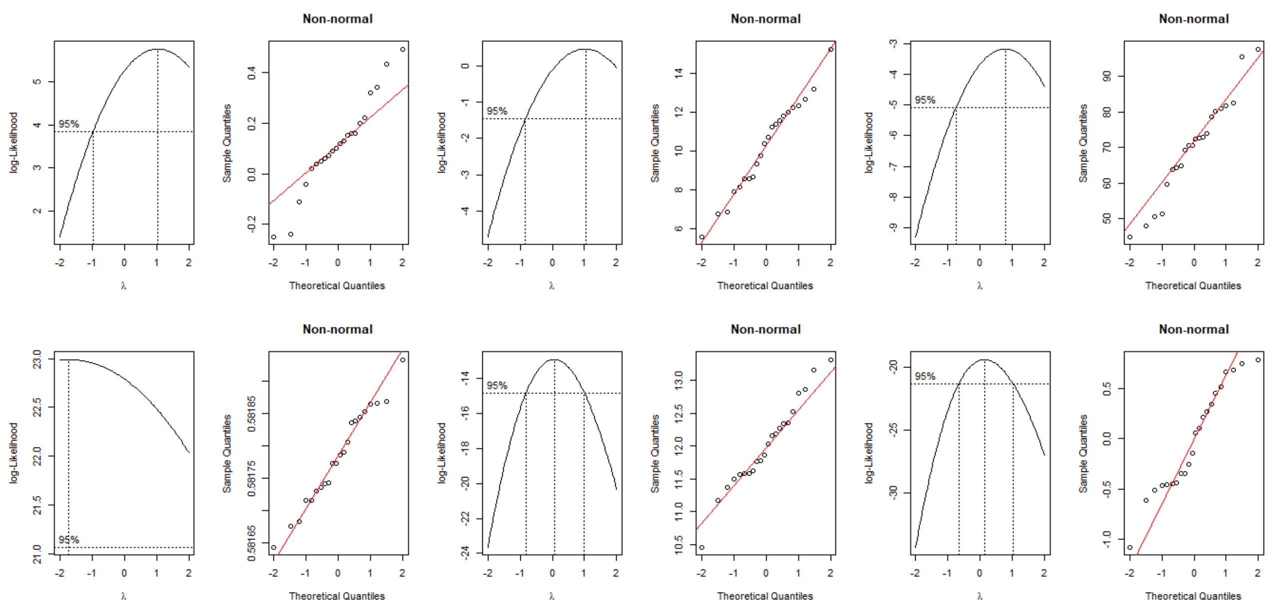
Interpretation:

In summary, the Shapiro-Wilk test provides a statistical hypothesis test to determine if a dataset follows a normal distribution, while the QQ plot and QQ line are graphical tools to visually assess the departure from normality. These methods are useful for checking the assumptions of normality in statistical analyses that rely on the normal distribution, such as parametric tests.

If the data points in the QQ plot closely align with the QQ line, it suggests that the data is nearly or approximately normally distributed except Percent Nuclear. In this case, the points fall along or near the line, indicating a good fit to the expected distribution.

```
# Boxcox Transformation
dev.new(width = 30,   # Create new plot window
        height = 15,
        noRStudioGD = TRUE)

par(mfrow=c(2,6))

for (i in 1:8)
{
```

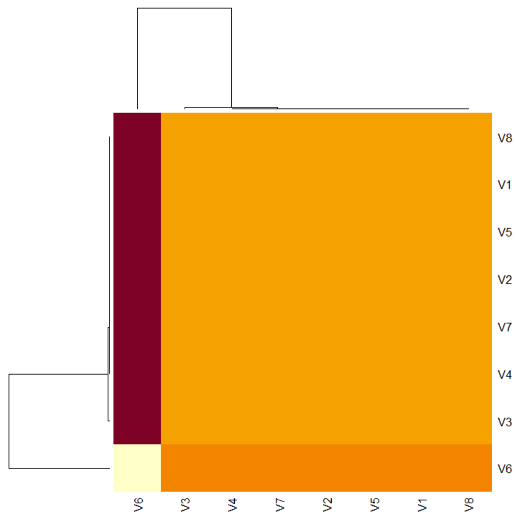
```
  if (0 %in% data[,i] | any(data[,i] < 0)) {
    next
  }
  b <- boxcox(lm(data[,i] ~ 1))  # Perform Box-Cox transformation
  lambda <- b$x[which.max(b$y)]  # Extract the lambda value
  bcx <- (data[,i]^lambda - 1) / lambda  # Apply Box-Cox transformation

  # Check normality of Box-Cox transformed data using QQ-plot and Shapiro-Wilk test
  #  Normal or Not Normal ?
  qqnorm(bcx, main = 'Non-normal')  # Create QQ-plot of transformed data
  qqline(bcx, col = "red")  # Add reference line to QQ-plot
  shapiro.test(bcx)  # Perform Shapiro-Wilk test on transformed data
}
```



Interpretation:
The Box-Cox transformation is used to improve the normality of data by applying a power transformation. In the provided code, each variable is transformed using the Box-Cox transformation. The resulting transformed data is assessed for normality using QQ-plots and the Shapiro-Wilk test. A linear QQ-plot and a high p-value in the Shapiro-Wilk test suggest approximate normality. Deviation from linearity in the QQ-plot or a low p-value indicates departure from normality. The data with negative and zero value cannot be interpret.

Question 6: Group the companies by similarity measures and hence show the grouping information by a plot?

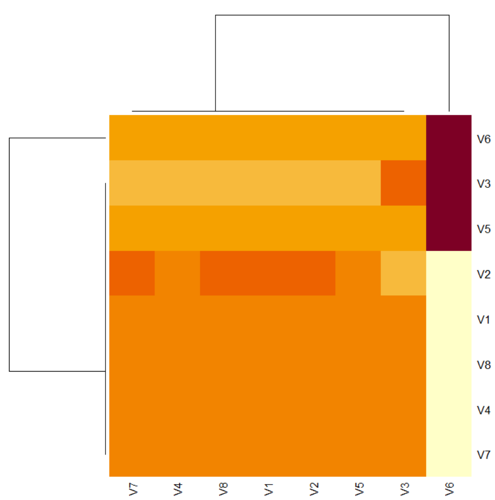
dmat<-dist(t(data[1:8]),method="euclidean",diag=T,upper=T)
heatmap(as.matrix(dmat))



Interpretaion:

A distance matrix is a square matrix that represents the dissimilarity or similarity between objects or observations. It is calculated using the Euclidean distance metric and applied to a transposed subset of data. The resulting distance matrix is visualized using a heatmap, which provides a visual representation of the similarity or dissimilarity between observations. The heatmap reveals color intensity, clustering, outliers, and similarity comparison, helping to identify clusters, outliers, and assess overall similarity or dissimilarity between observations.

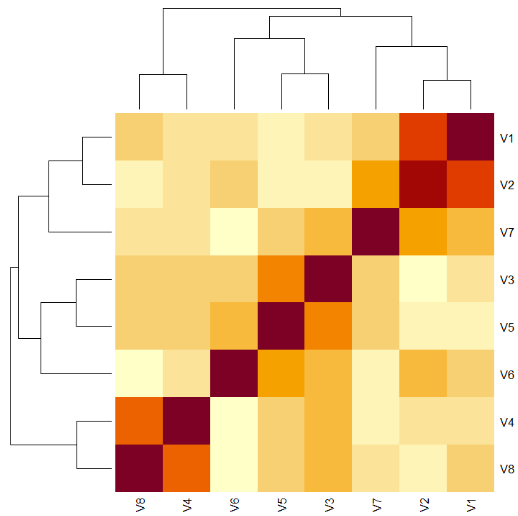
heatmap(cov(data[1:8]))



Interpretation:

The heatmap() function generates a covariance heatmap for similarity measurement, displaying relationships between variables based on their covariance values. The heatmap includes color intensity, clustering, strength of relationship, and patterns of positive or negative covariance. Higher covariance values indicate stronger relationships, while lower values indicate weaker ones. By examining the heatmap, one can identify patterns of positive or negative covariance, allowing for a better understanding of the data's structure and interdependencies.

```
heatmap(cor(data[1:8]))
```



Interpretation:

The correlation heatmap is a tool used for measuring similarity in data. It displays the relationships between variables based on their correlation values, with color intensity representing strength and direction. Clustering patterns reveal similar correlation patterns with other variables, while strength of relationship indicates the magnitude of correlation values. The heatmap helps identify patterns of positive or negative correlation, allowing for a better understanding of linear associations within the data.