Omkar Gurav
BE IT
8048

Rajdhani
DATE / /

Assignment 2

- Title :- Finding Principal Components, Variance and Standard Deviation calculations of principal components. (Using R)

- Theory:-

1. Explain Dimensionality Reduction.

→ Dimensionality reduction refers to techniques for reducing the number of input variables in training data.
It is series of techniques in machine learning & statistics to reduce the number of random variables to consider.
It envolves feature selection & feature extraction.

① Feature selection:-

Feature selection techniques find a smaller subset of a many dimensional data set to create a data model.
It is a technique of finding k features of the d dimensions that give us the most information & discard the other (d-k) dimensions.
Subset selection is one of the widely used method for as a feature selection method.

② Feature extraction:-

Feature extraction involves transforming high dimensional data into spaces of fewer dimensions.

Feature extraction is a technique of finding a new set of k dimensions that are combinations of the original d dimensions.
These methods may be supervised or unsupervised depends on whether or not they use the output information.
The best known & most widely used feature extraction method is Principal Components Analysis (PCA).

2. Explain Forward selection and Backward selection Methods.

→ Forward selection:-

It starts with no variables or null model.
In next step, it will add one by one feature which is not already considered before.
At each step after adding one feature the error is checked.
The process is continuing until it will find the subset of features that decreases the error the most, or until any further addition does not decrease the error

Backward selection:-

It starts with all variables.
In next step it remove feature one by one.
At each step removing the feature, the error is checked.
The process is continuing until it will find the

subset of features that decrease the error most, or until any further removal increases the error significantly.
In either case, checking the error should be done on a validation set which distinct from the training set.
With more features, generally training error can be reduced, but validation error may not be reduced.

4. Explain Principal Component Analysis.

→ PCA is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity.
Because smaller data sets are easier to explore & visualize & make analyzing data much easier & faster for ML algorithms without extraneous variables to process.

- Conclusion:- Principal Component Analysis is implemented using R..

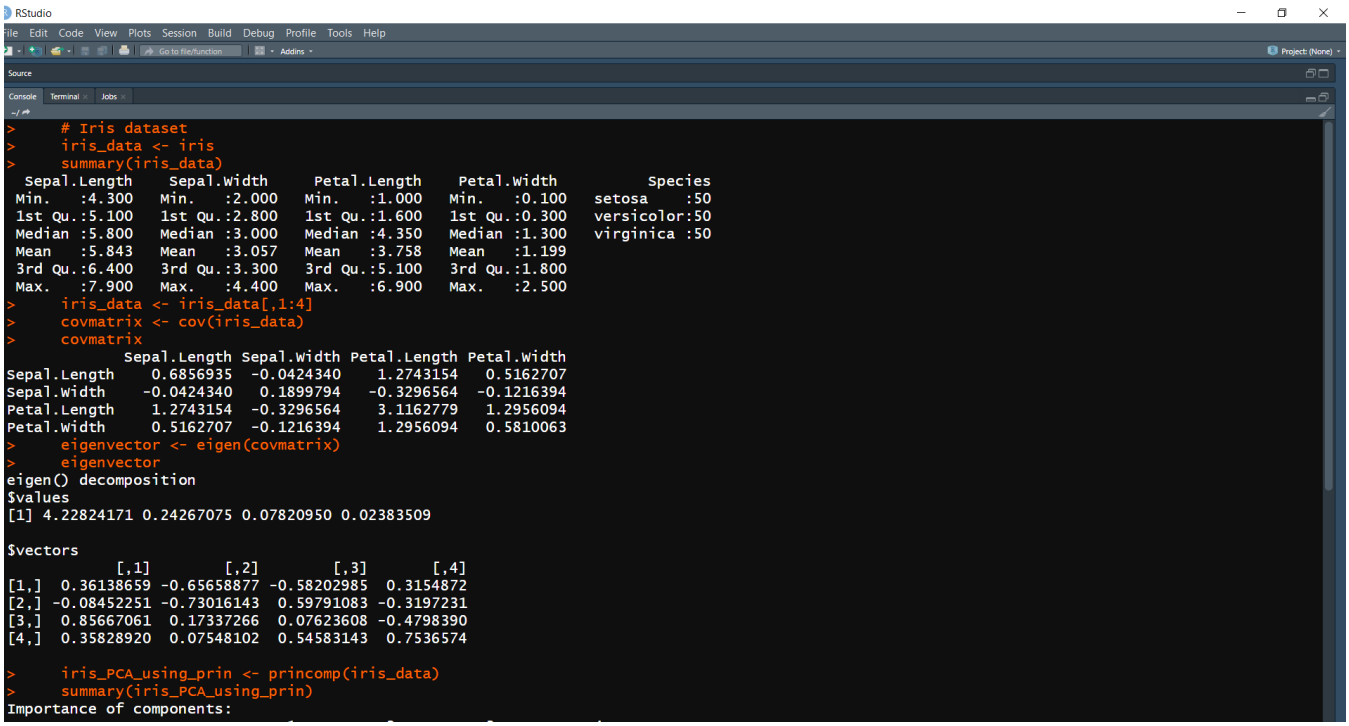
PCA.R

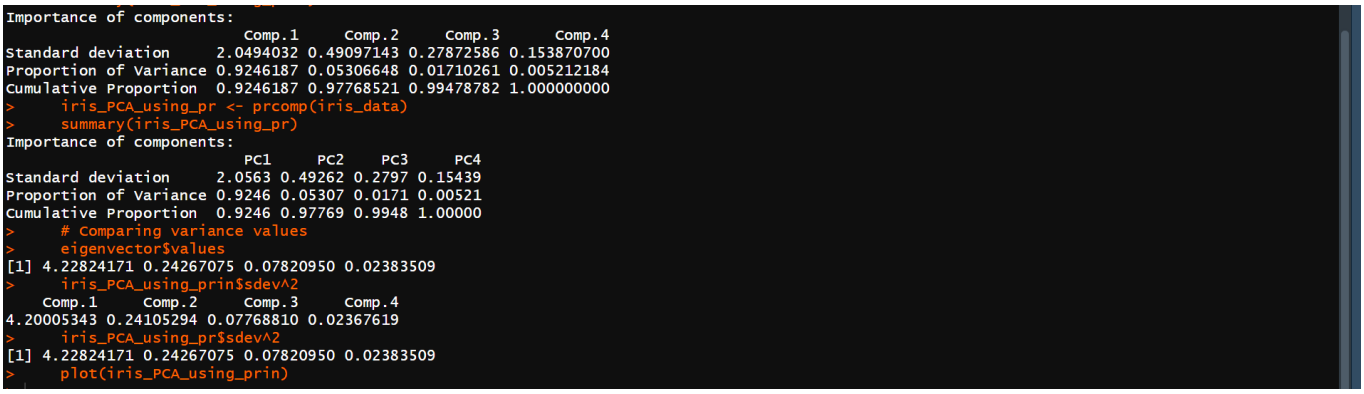
```
# Iris dataset
iris_data <- iris

summary(iris_data)
iris_data <- iris_data[,1:4]

covmatrix <- cov(iris_data)
covmatrix

eigenvector <- eigen(covmatrix)
eigenvector

iris_PCA_using_prin <- princomp(iris_data)
summary(iris_PCA_using_prin)

iris_PCA_using_pr <- prcomp(iris_data)
summary(iris_PCA_using_pr)

# Comparing variance values
eigenvector$values
iris_PCA_using_prin$sdev^2
iris_PCA_using_pr$sdev^2

plot(iris_PCA_using_prin)
screeplot(iris_PCA_using_prin, type = "lines")
biplot(iris_PCA_using_prin)
```

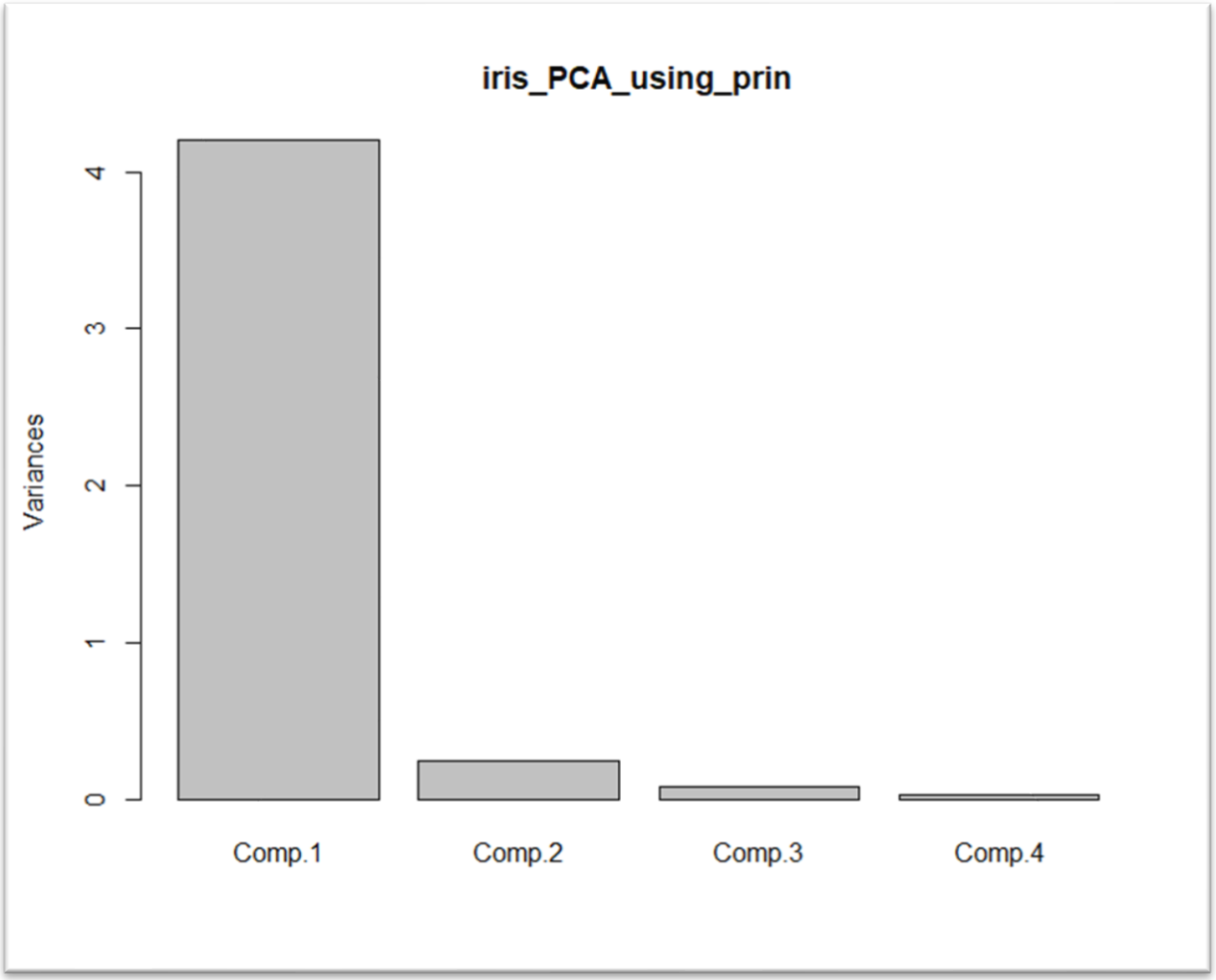
```
plot(iris_PCA_using_pr)
screeplot(iris_PCA_using_pr, type = "lines")
biplot(iris_PCA_using_pr)



# gsp dataset

gsp_data <- read.csv('C:/Users/DELL/Downloads/pca_gsp.csv')

summary(gsp_data)
gsp_data <- gsp_data[,2:14]

covmatrix1 <- cov(gsp_data)
covmatrix1

eigenvector1 <- eigen(covmatrix1)
eigenvector1

gsp_PCA_using_prin <- princomp(gsp_data)
summary(gsp_PCA_using_prin)

gsp_PCA_using_pr <- prcomp(gsp_data)
summary(gsp_PCA_using_pr)

# Comparing variance values
eigenvector1$values
gsp_PCA_using_prin$sdev^2
gsp_PCA_using_pr$sdev^2
```

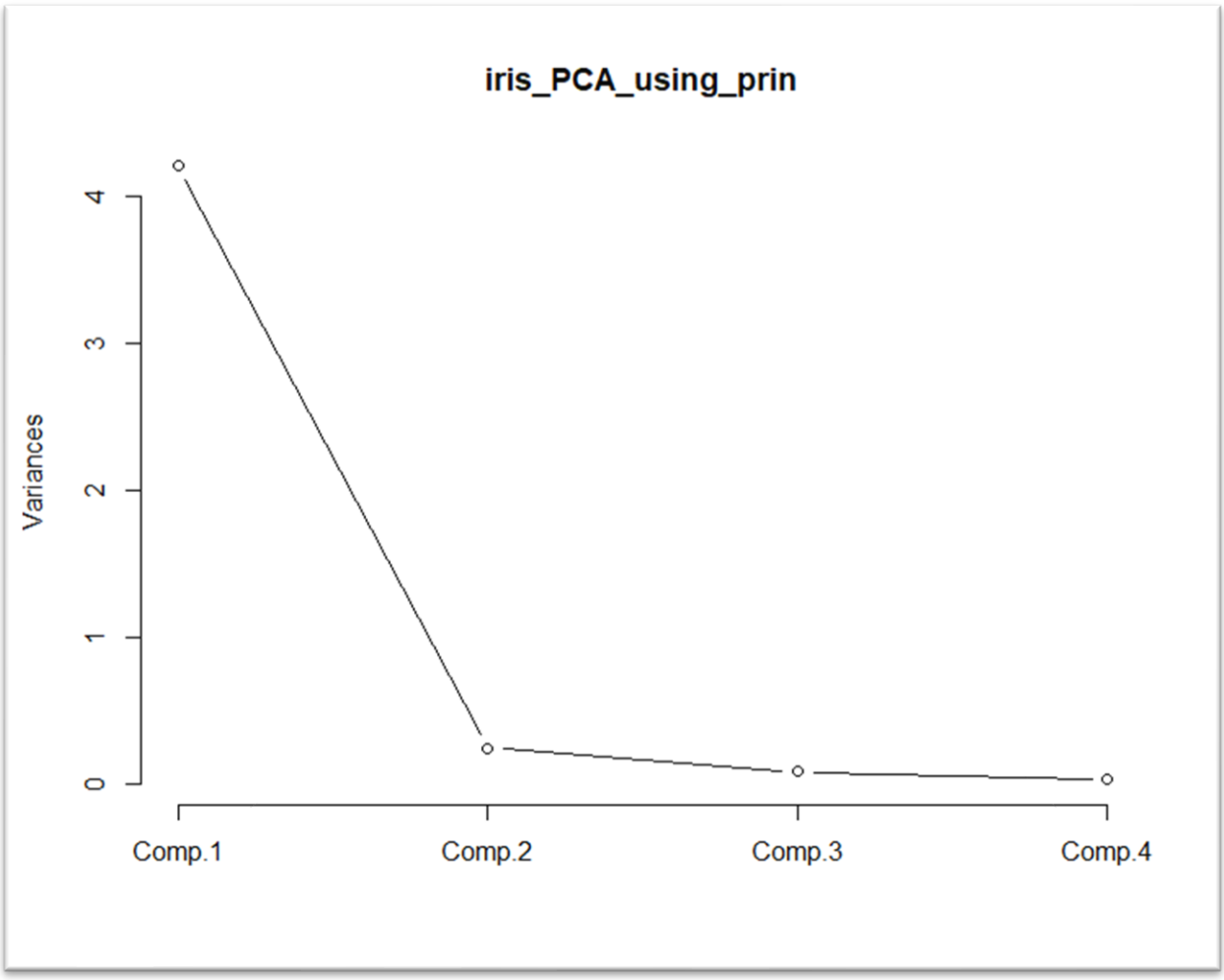
```
plot(gsp_PCA_using_prin)

screeplot(gsp_PCA_using_prin, type = "lines")

biplot(gsp_PCA_using_prin)


plot(gsp_PCA_using_pr)

screeplot(gsp_PCA_using_pr, type = "lines")

biplot(gsp_PCA_using_pr)
```

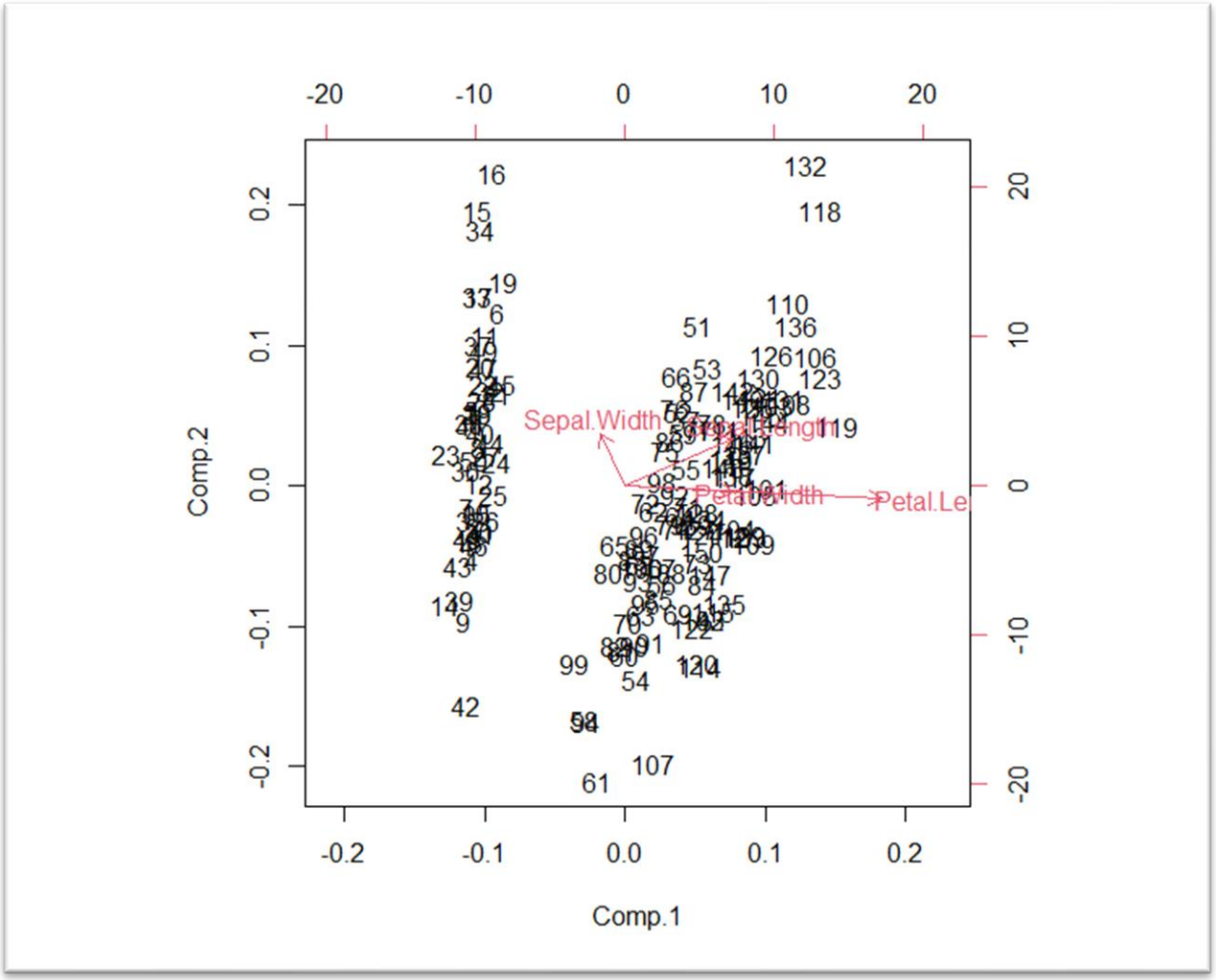
Output :

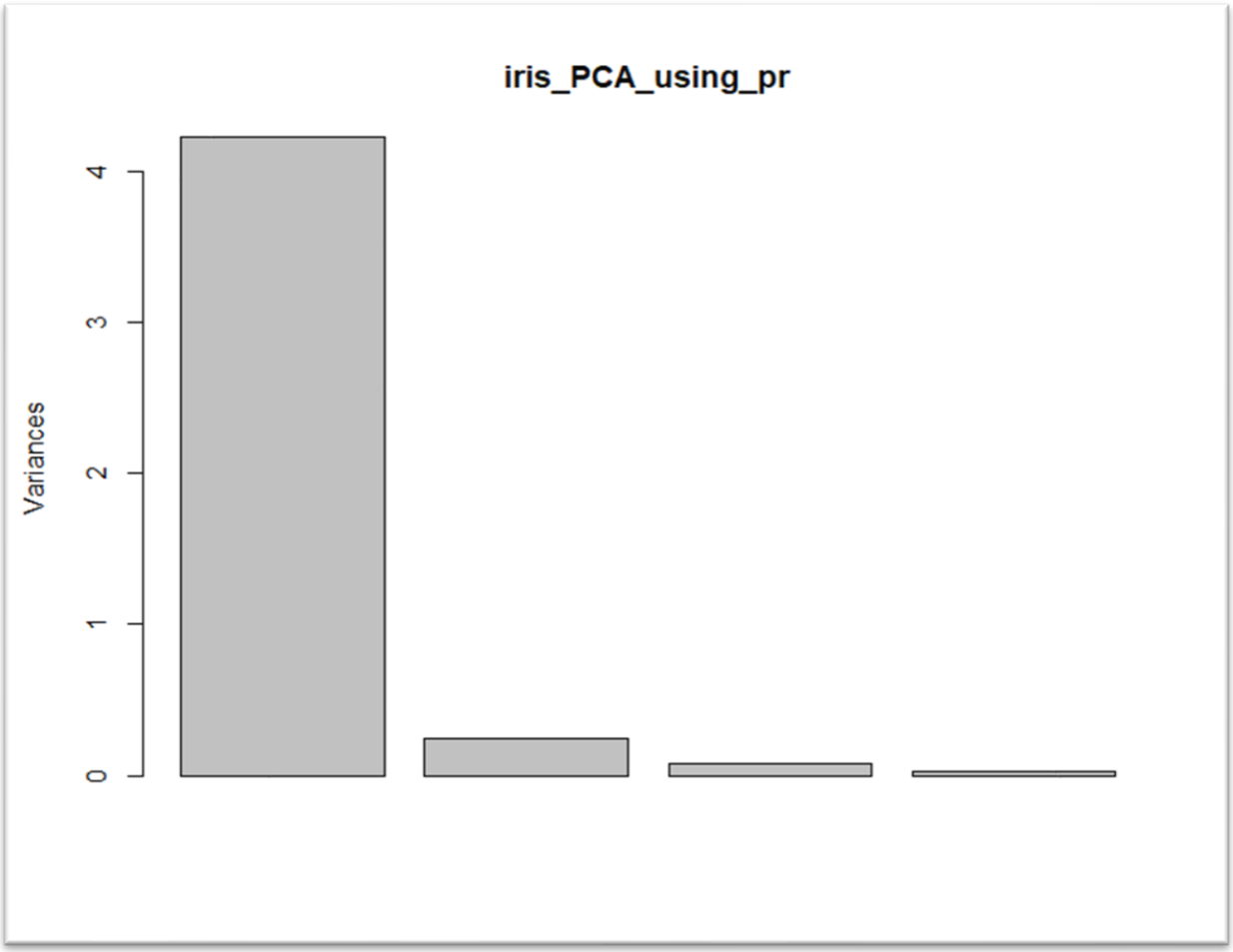


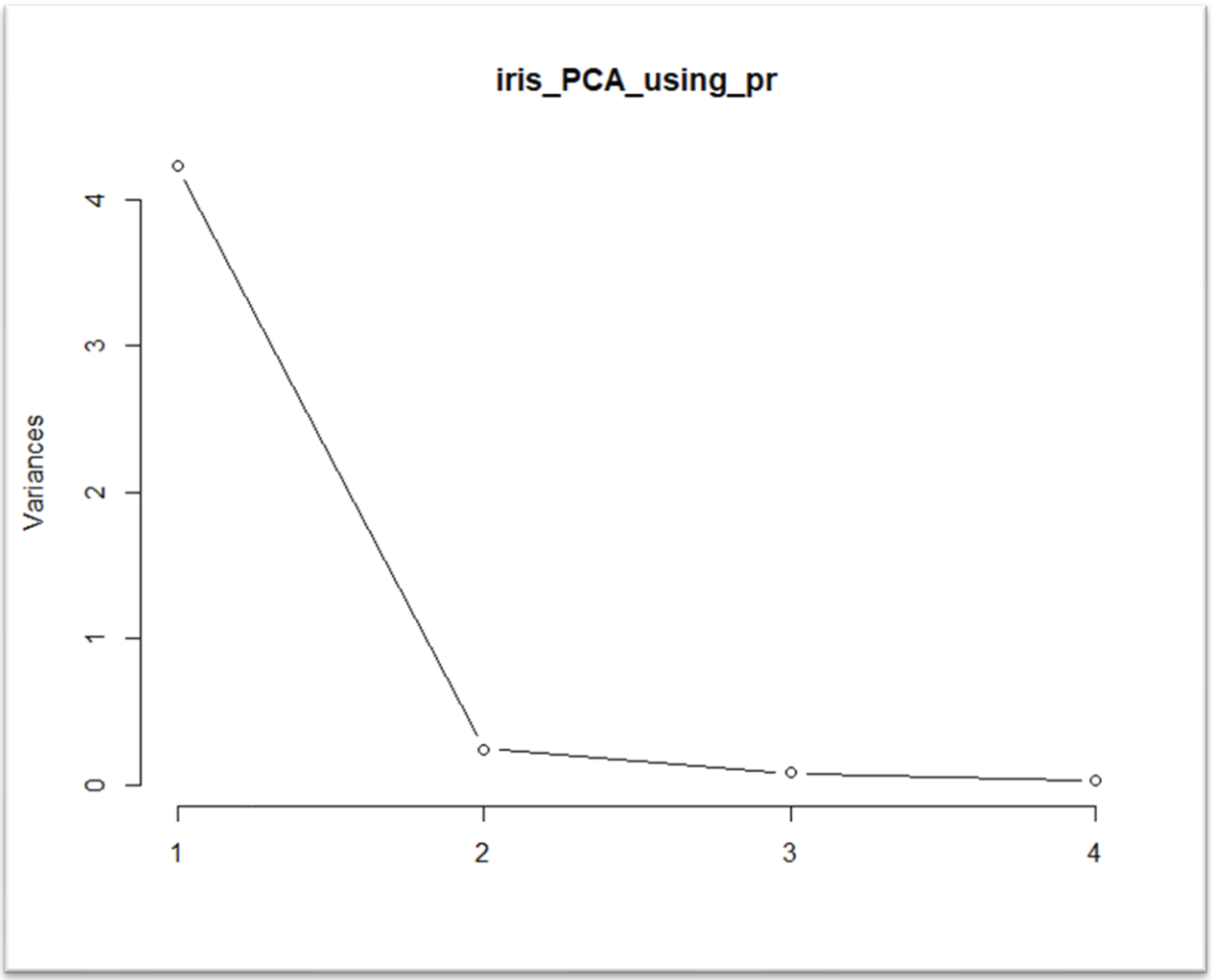


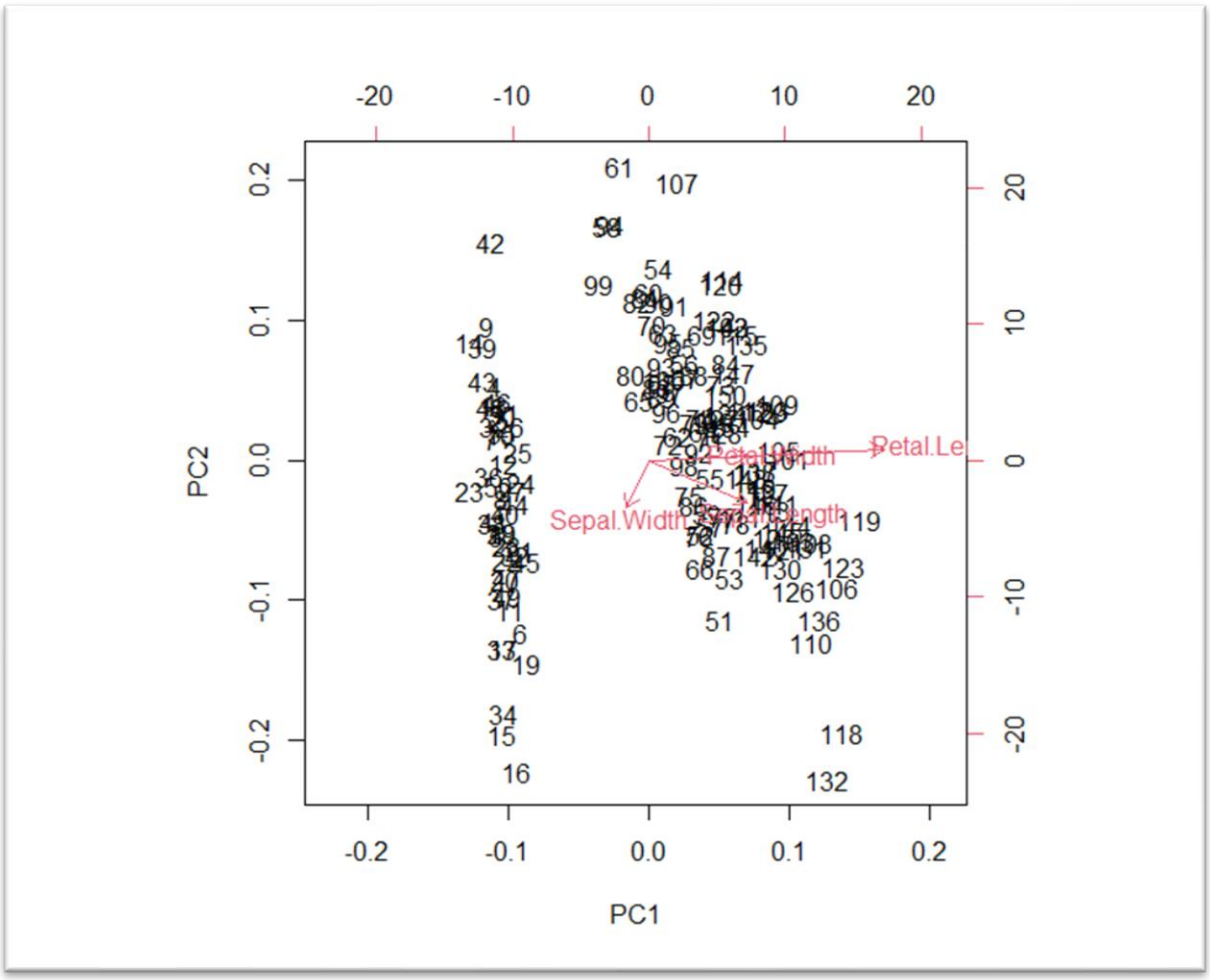

iris_PCA_using_prin
Variances
0
1
2
3
4
Comp.1
Comp.2
Comp.3
Comp.4

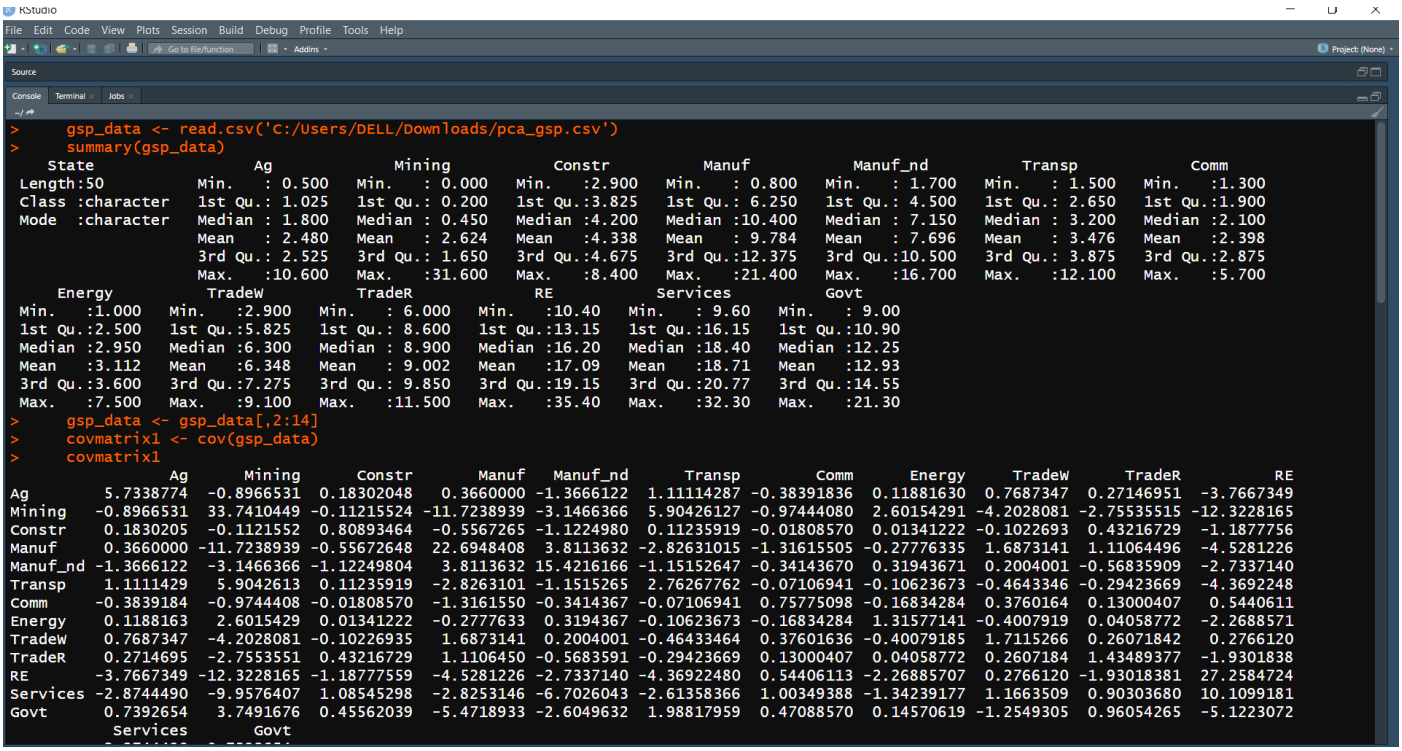

iris_PCA_using_prin
Variances
0
1
2
3
4
Comp.1
Comp.2
Comp.3
Comp.4


Sepal.Width
Petal.Le
Comp.1
Comp.2

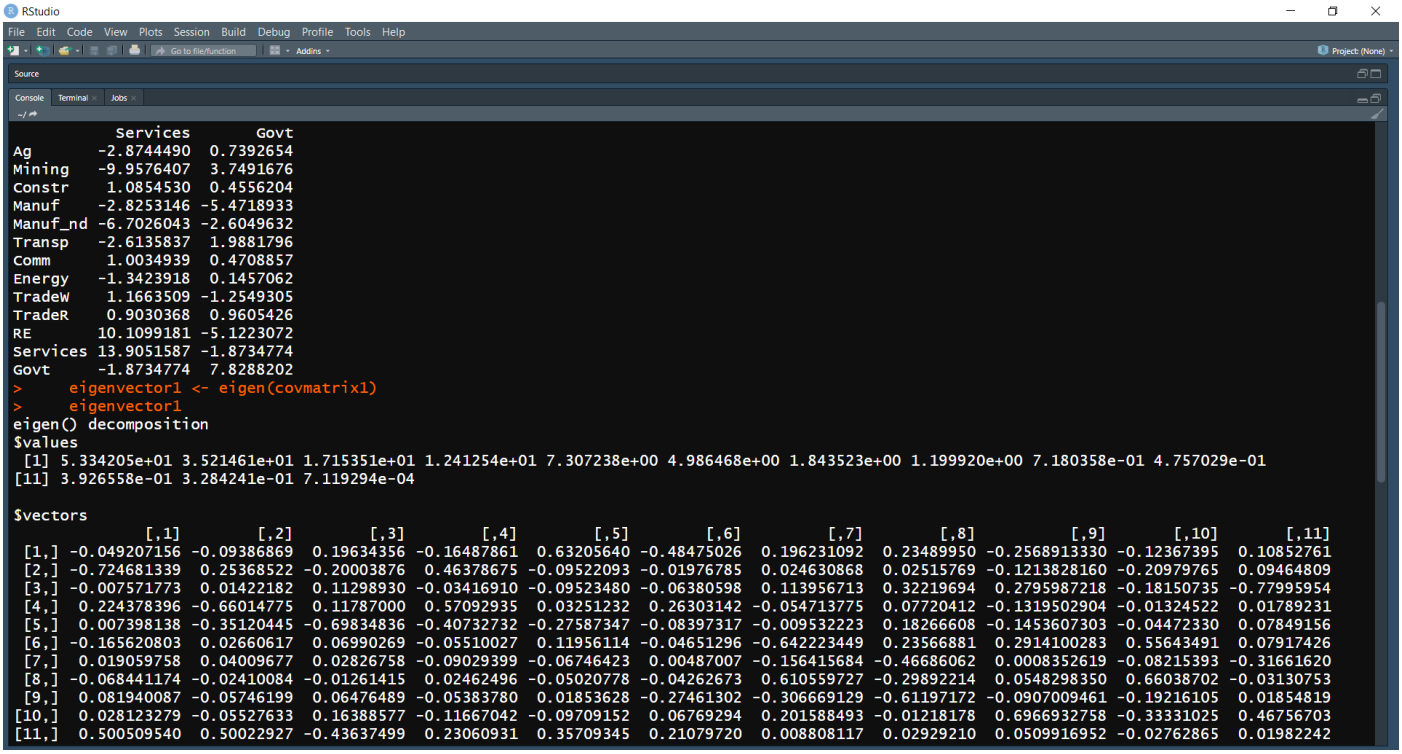

iris_PCA_using_pr
Variances


iris_PCA_using_pr
Variances
0
1
2
3
4
1
2
3
4


-20
-10
0
10
20
PC2
0.2
0.1
0.0
-0.1
-0.2
20
10
0
-10
-20
Petal.Le
Sepal.Width
-0.2
-0.1
0.0
0.1
0.2
PC1



```
> gsp_data <- read.csv('C:/Users/DELL/Downloads/pca_gsp.csv')
> summary(gsp_data)
    State                Ag              Mining            Constr          Manuf           Manuf_nd         Transp           Comm      
 Length:50          Min.   : 0.500   Min.   : 0.000   Min.   :2.900   Min.   : 0.800   Min.   : 1.700   Min.   : 1.500   Min.   :1.300  
 Class :character   1st Qu.: 1.025   1st Qu.: 0.200   1st Qu.:3.825   1st Qu.: 6.250   1st Qu.: 4.500   1st Qu.: 2.650   1st Qu.:1.900  
 Mode  :character   Median : 1.800   Median : 0.450   Median :4.200   Median :10.400   Median : 7.150   Median : 3.200   Median :2.100  
                    Mean   : 2.480   Mean   : 2.624   Mean   :4.338   Mean   : 9.784   Mean   : 7.696   Mean   : 3.476   Mean   :2.398  
                    3rd Qu.: 2.525   3rd Qu.: 1.650   3rd Qu.:4.675   3rd Qu.:12.375   3rd Qu.:10.500   3rd Qu.: 3.875   3rd Qu.:2.875  
                    Max.   :10.600   Max.   :31.600   Max.   :8.400   Max.   :21.400   Max.   :16.700   Max.   :12.100   Max.   :5.700  
     Energy          TradeW          TradeR            RE           Services         Govt      
 Min.   :1.000   Min.   :2.900   Min.   : 6.000   Min.   :10.40   Min.   : 9.60   Min.   : 9.00  
 1st Qu.:2.500   1st Qu.:5.825   1st Qu.: 8.600   1st Qu.:13.15   1st Qu.:16.15   1st Qu.:10.90  
 Median :2.950   Median :6.300   Median : 8.900   Median :16.20   Median :18.40   Median :12.25  
 Mean   :3.112   Mean   :6.348   Mean   : 9.002   Mean   :17.09   Mean   :18.71   Mean   :12.93  
 3rd Qu.:3.600   3rd Qu.:7.275   3rd Qu.: 9.850   3rd Qu.:19.15   3rd Qu.:20.77   3rd Qu.:14.55  
 Max.   :7.500   Max.   :9.100   Max.   :11.500   Max.   :35.40   Max.   :32.30   Max.   :21.30  
> gsp_data <- gsp_data[,2:14]
> covmatrix1 <- cov(gsp_data)
> covmatrix1
                Ag      Mining      Constr       Manuf   Manuf_nd      Transp        Comm      Energy      TradeW      TradeR          RE
Ag        5.7338774  -0.8966531  0.18302048   0.3660000 -1.3666122  1.11114287 -0.38391836  0.11881630  0.7687347  0.27146951  -3.7667349
Mining   -0.8966531  33.7410449 -0.11215524 -11.7238939 -3.1466366  5.90426127 -0.97444080  2.60154291 -4.2028081 -2.75535515 -12.3228165
Constr    0.1830205  -0.1121552  0.80893464  -0.5567265 -1.1224980  0.11235919 -0.01808570  0.01341222 -0.1022693  0.43216729  -1.1877756
Manuf     0.3660000 -11.7238939 -0.55672648  22.6948408  3.8113632 -2.82631015 -1.31615505 -0.27776335  1.6873141  1.11064496  -4.5281226
Manuf_nd -1.3666122  -3.1466366 -1.12249804   3.8113632 15.4216166 -1.15152647 -0.34143670  0.31943671  0.2004001 -0.56835909  -2.7337140
Transp    1.1111429   5.9042613  0.11235919  -2.8263101 -1.1515265  2.76267762 -0.07106941 -0.10623673 -0.4643346 -0.29423669  -4.3692248
Comm     -0.3839184  -0.9744408 -0.01808570  -1.3161550 -0.3414367 -0.07106941  0.75775098 -0.16834284  0.3760164  0.13000407   0.5440611
Energy    0.1188163   2.6015429  0.01341222  -0.2777633  0.3194367 -0.10623673 -0.16834284  1.31577141 -0.4007919  0.04058772  -2.2688571
TradeW    0.7687347  -4.2028081 -0.10226935   1.6873141  0.2004001 -0.46433464  0.37601636 -0.40079185  1.7115266  0.26071842   0.2766120
TradeR    0.2714695  -2.7553551  0.43216729   1.1106450 -0.5683591 -0.29423669  0.13000407  0.04058772  0.2607184  1.43489377  -1.9301838
RE       -3.7667349 -12.3228165 -1.18777559  -4.5281226 -2.7337140 -4.36922480  0.54406113 -2.26885707  0.2766120 -1.93018381  27.2584724
Services -2.8744490  -9.9576407  1.08545298  -2.8253146 -6.7026043 -2.61358366  1.00349388 -1.34239177  1.1663509  0.90303680  10.1099181
Govt      0.7392654   3.7491676  0.45562039  -5.4718933 -2.6049632  1.98817959  0.47088570  0.14570619 -1.2549305  0.96054265  -5.1223072
            Services       Govt
```



```
            Services       Govt
Ag        -2.8744490  0.7392654
Mining    -9.9576407  3.7491676
Constr     1.0854530  0.4556204
Manuf     -2.8253146 -5.4718933
Manuf_nd  -6.7026043 -2.6049632
Transp    -2.6135837  1.9881796
Comm       1.0034939  0.4708857
Energy    -1.3423918  0.1457062
TradeW     1.1663509 -1.2549305
TradeR     0.9030368  0.9605426
RE        10.1099181 -5.1223072
Services  13.9051587 -1.8734774
Govt      -1.8734774  7.8288202
> eigenvector1 <- eigen(covmatrix1)
> eigenvector1
eigen() decomposition
$values
 [1] 5.334205e+01 3.521461e+01 1.715351e+01 1.241254e+01 7.307238e+00 4.986468e+00 1.843523e+00 1.199920e+00 7.180358e-01 4.757029e-01
[11] 3.926558e-01 3.284241e-01 7.119294e-04

$vectors
             [,1]         [,2]         [,3]         [,4]         [,5]         [,6]         [,7]         [,8]          [,9]         [,10]        [,11]
 [1,] -0.049207156 -0.09386869  0.19634356 -0.16487861  0.63205640 -0.48475026  0.196231092  0.23489950 -0.2568913330 -0.12367395  0.10852761
 [2,] -0.724681339  0.25368522 -0.20003876  0.46378675 -0.09522093 -0.01976785  0.024630868  0.02515769 -0.1213828160 -0.20979765  0.09464809
 [3,] -0.007571773  0.01422182  0.11298930 -0.03416910 -0.09523480 -0.06380598  0.113956713  0.32219694  0.2795987218 -0.18150735 -0.77995954
 [4,]  0.224378396 -0.66014775  0.11787000  0.57092935  0.03251232  0.26303142 -0.054713775  0.07720412 -0.1319502904 -0.01324522  0.01789231
 [5,]  0.007398138 -0.35120445 -0.69834836 -0.40732732 -0.27587347 -0.08397317 -0.009532223  0.18266608 -0.1453607303 -0.04472330  0.07849156
 [6,] -0.165620803  0.02660617  0.06990269 -0.05510027  0.11956114 -0.04651296 -0.642223449  0.23566881  0.2914100283  0.55643491  0.07917426
 [7,]  0.019059758  0.04009677  0.02826758 -0.09029399 -0.06746423  0.00487007 -0.156415684 -0.46686062  0.0008352619 -0.08215393 -0.31661620
 [8,] -0.068441174 -0.02410084 -0.01261415  0.02462496 -0.05020778 -0.04262673  0.610559727 -0.29892214  0.0548298350  0.66038702 -0.03130753
 [9,]  0.081940087 -0.05746199  0.06476489 -0.05383780  0.01853628 -0.27461302 -0.306669129 -0.61197172 -0.0907009461 -0.19216105  0.01854819
[10,]  0.028123279 -0.05527633  0.16388577 -0.11667042 -0.09709152  0.06769294  0.201588493 -0.01218178  0.6966932758 -0.33331025  0.46756703
[11,]  0.500509540  0.50022927 -0.43637499  0.23060931  0.35709345  0.21079720  0.008808117  0.02929210  0.0509916952 -0.02762865  0.01982242
```

RStudio

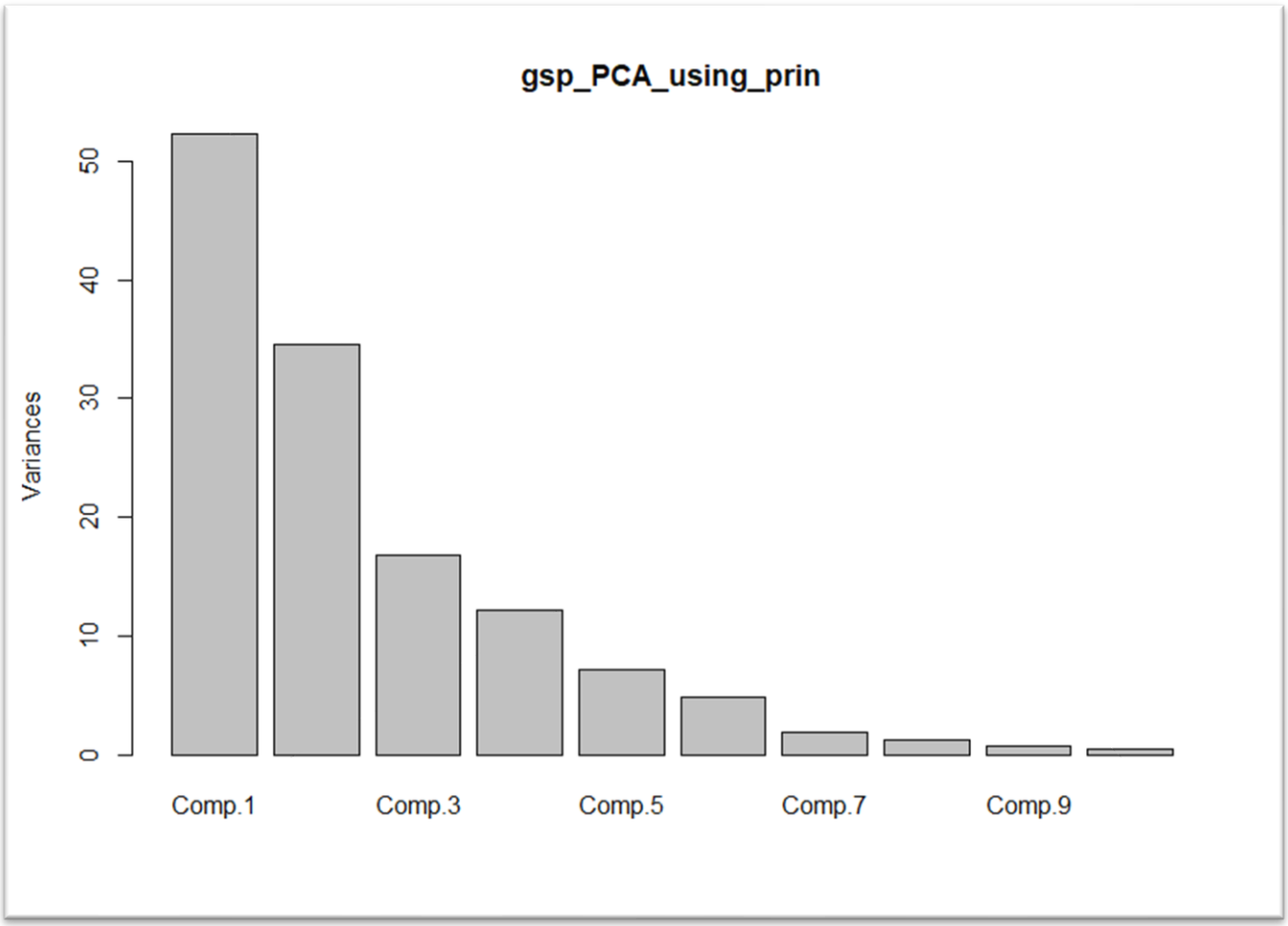
File Edit Code View Plots Session Build Debug Profile Tools Help

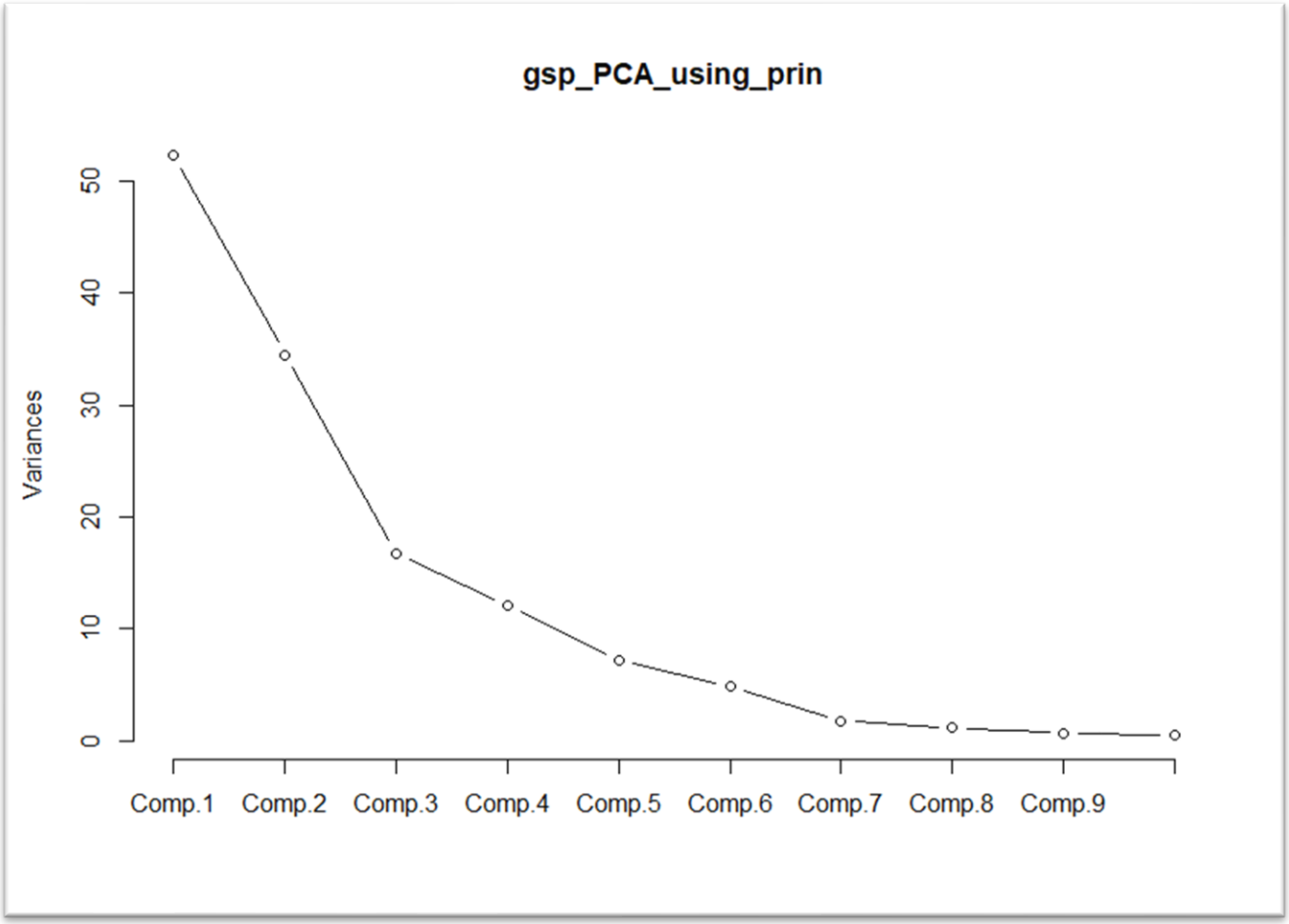
Source

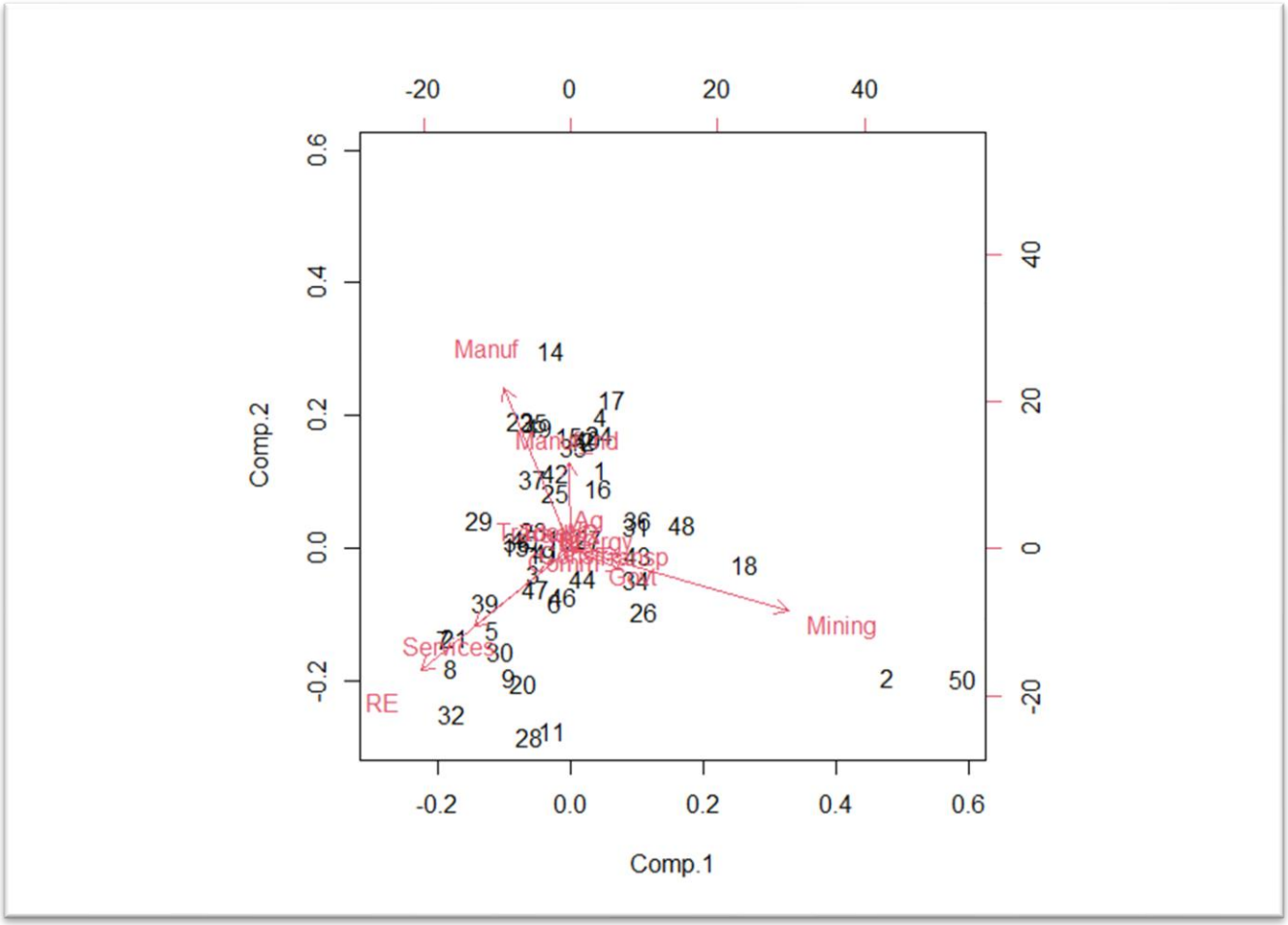
Console Terminal Jobs

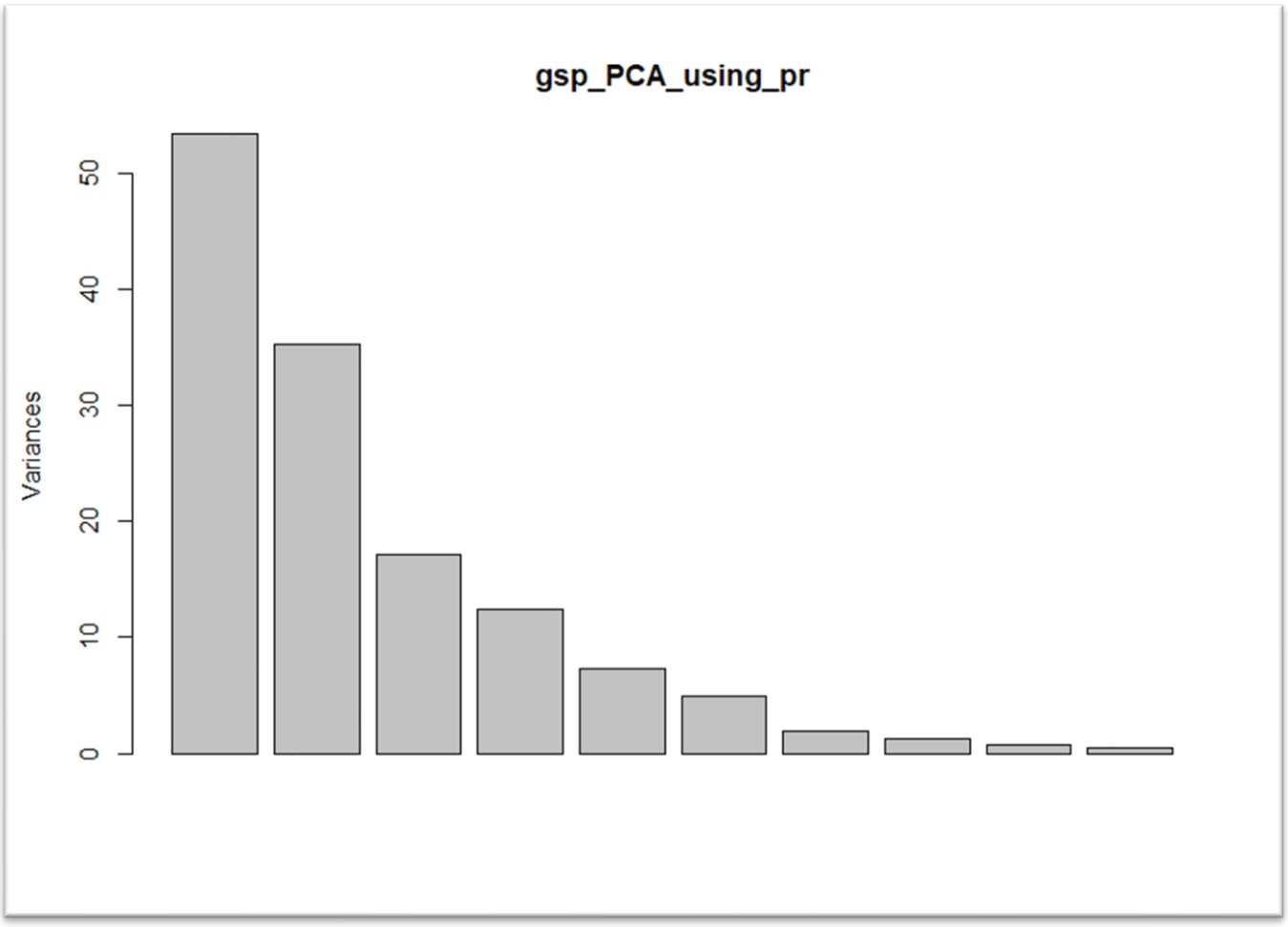
```
             [,12]     [,13]
 [1,]  0.158352658 0.2769961
 [2,] -0.003047295 0.2774047
 [3,] -0.243439909 0.2777556
 [4,]  0.038307651 0.2746544
 [5,]  0.010979928 0.2774309
 [6,]  0.009591358 0.2803394
 [7,]  0.748929273 0.2723715
 [8,] -0.098781447 0.2786372
 [9,] -0.563283666 0.2753083
[10,]  0.063666338 0.2858987
[11,] -0.040233164 0.2781999
[12,]  0.056932145 0.2766557
[13,] -0.130833613 0.2736553

>     gsp_PCA_using_prin <- princomp(gsp_data)
>     summary(gsp_PCA_using_prin)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6     Comp.7      Comp.8      Comp.9     Comp.10
Standard deviation     7.2301594 5.8745480 4.1000537 3.4877344 2.6760219 2.21059680 1.34411782 1.084399330 0.838853405 0.682780255
Proportion of Variance 0.3940306 0.2601256 0.1267107 0.0916898 0.0539776 0.03683437 0.01361786 0.008863652 0.005304035 0.003513954
Cumulative Proportion  0.3940306 0.6541562 0.7808669 0.8725567 0.9265343 0.96336872 0.97698658 0.985850231 0.991154265 0.994668220
                          Comp.11      Comp.12       Comp.13
Standard deviation     0.620324655 0.567323202 2.641384e-02
Proportion of Variance 0.002900496 0.002426025 5.258928e-06
Cumulative Proportion  0.997568716 0.999994741 1.000000e+00
>     gsp_PCA_using_pr <- prcomp(gsp_data)
>     summary(gsp_PCA_using_pr)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8    PC9    PC10   PC11    PC12    PC13
Standard deviation     7.304 5.9342 4.1417 3.52314 2.70319 2.23304 1.35776 1.09541 0.8474 0.68971 0.6266 0.57308 0.02668
Proportion of Variance 0.394 0.2601 0.1267 0.09169 0.05398 0.03683 0.01362 0.00886 0.0053 0.00351 0.0029 0.00243 0.00001
Cumulative Proportion  0.394 0.6542 0.7809 0.87256 0.92653 0.96337 0.97699 0.98585 0.9911 0.99467 0.9976 0.99999 1.00000
>     # Comparing variance values
```

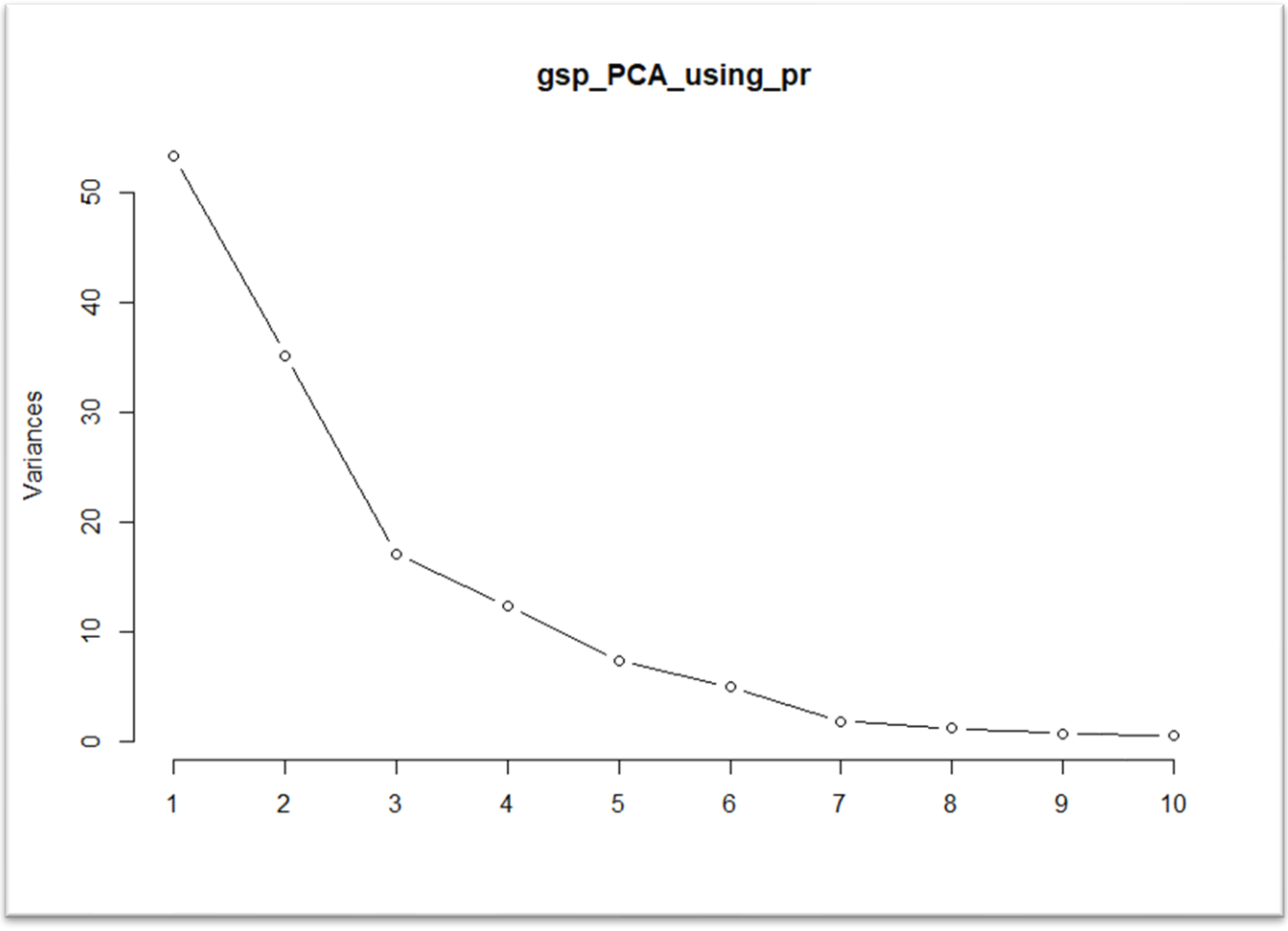
```
>     # Comparing variance values
>     eigenvector1$values
 [1] 5.334205e+01 3.521461e+01 1.715351e+01 1.241254e+01 7.307238e+00 4.986468e+00 1.843523e+00 1.199920e+00 7.180358e-01 4.757029e-01
[11] 3.926558e-01 3.284241e-01 7.119294e-04
>     gsp_PCA_using_prin$sdev^2
     Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6       Comp.7       Comp.8       Comp.9      Comp.10
5.227521e+01 3.451031e+01 1.681044e+01 1.216429e+01 7.161093e+00 4.886738e+00 1.806653e+00 1.175922e+00 7.036750e-01 4.661889e-01
     Comp.11      Comp.12      Comp.13
3.848027e-01 3.218556e-01 6.976908e-04
>     gsp_PCA_using_pr$sdev^2
 [1] 5.334205e+01 3.521461e+01 1.715351e+01 1.241254e+01 7.307238e+00 4.986468e+00 1.843523e+00 1.199920e+00 7.180358e-01 4.757029e-01
[11] 3.926558e-01 3.284241e-01 7.119294e-04
>
```

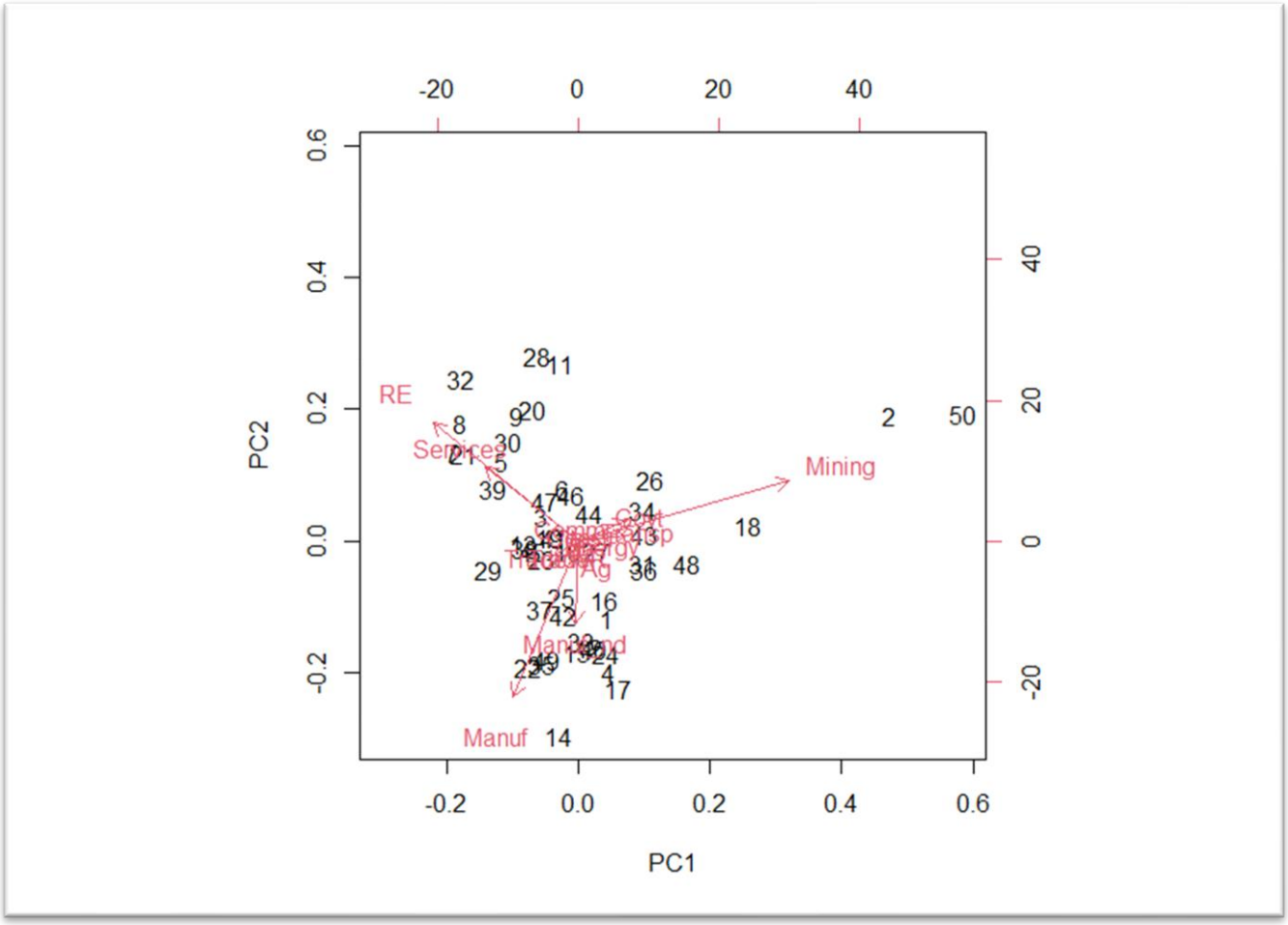



gsp_PCA_using_prin
Variances
0
10
20
30
40
50
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7
Comp.8
Comp.9


-20
0
20
40
0.6
0.4
0.2
0.0
-0.2
Comp.2
Comp.1
-0.2
0.0
0.2
0.4
0.6
40
20
0
-20
Manuf
Mining
Services
RE
Ag
14
17
4
1
16
29
48
18
26
39
5
30
8
9
20
2
50
32
28
11


gsp_PCA_using_pr
Variances
0
10
20
30
40
50


gsp_PCA_using_pr
Variances
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10


-20
0
20
40
0.6
0.4
0.2
0.0
-0.2
PC2
40
20
0
-20
RE
Services
Mining
Manuf
28
11
32
20
8
30
21
5
39
26
47
46
44
34
18
29
48
25
16
37
42
1
17
14
-0.2
0.0
0.2
0.4
0.6
PC1