1. Section 1

1.1 a. Means, Standard Deviations, and Medians

The vertebral column data was first read from the ARFF file, then split into classes for processing.

```
library(foreign)
vert <- read.arff("column_2C_weka.arff")
vert_split <- split(vert, vert[,"class"])

sapply(vert_split$Abnormal[0:6], mean)
sapply(vert_split$Abnormal[0:6], median)
sapply(vert_split$Abnormal[0:6], sd)
sapply(vert_split$Normal[0:6], mean)
sapply(vert_split$Normal[0:6], median)
sapply(vert_split$Normal[0:6], sd)
```

1.1.1 Abnormal Data

mean

pelvic_incidence	pelvic_tilt	lumbar_lordosis_angle
64.69256	19.79111	55.92537
sacral_slope	pelvic_radius	degree_spondylolisthesis
44.90145	115.07771	37.77771

standard deviation

pelvic_incidence	pelvic_tilt	lumbar_lordosis_angle
65.27489	18.79890	56.15000
sacral_slope	pelvic_radius	degree_spondylolisthesis
44.63960	115.65032	31.94652

median

pelvic_incidence	pelvic_tilt	lumbar_lordosis_angle
17.66213	10.51587	19.66947
sacral_slope	pelvic_radius	degree_spondylolisthesis
14.51556	14.09060	40.69674

1.1.2 Normal Data

mean

pelvic_incidence	pelvic_tilt	lumbar_lordosis_angle
51.685244	12.821414	43.542605
sacral_slope	pelvic_radius	degree_spondylolisthesis
38.863830	123.890834	2.186572

standard deviation

pelvic_incidence	pelvic_tilt	lumbar_lordosis_angle
50.12312	13.48243	42.63892
sacral_slope	pelvic_radius	degree_spondylolisthesis
37.05969	123.87433	1.15271

median

pelvic_incidence	pelvic_tilt	lumbar_lordosis_angle
12.368161	6.778503	12.361388
sacral_slope	pelvic_radius	degree_spondylolisthesis
9.624004	9.014246	6.307483

1.2 b. Scatter Plots

```
library(foreign)
vert <- read.arff("column_2C_weka.arff")
pairs(vert[0:6], pch = 21, bg = c('green', 'blue')[unclass(vert$class)])
```

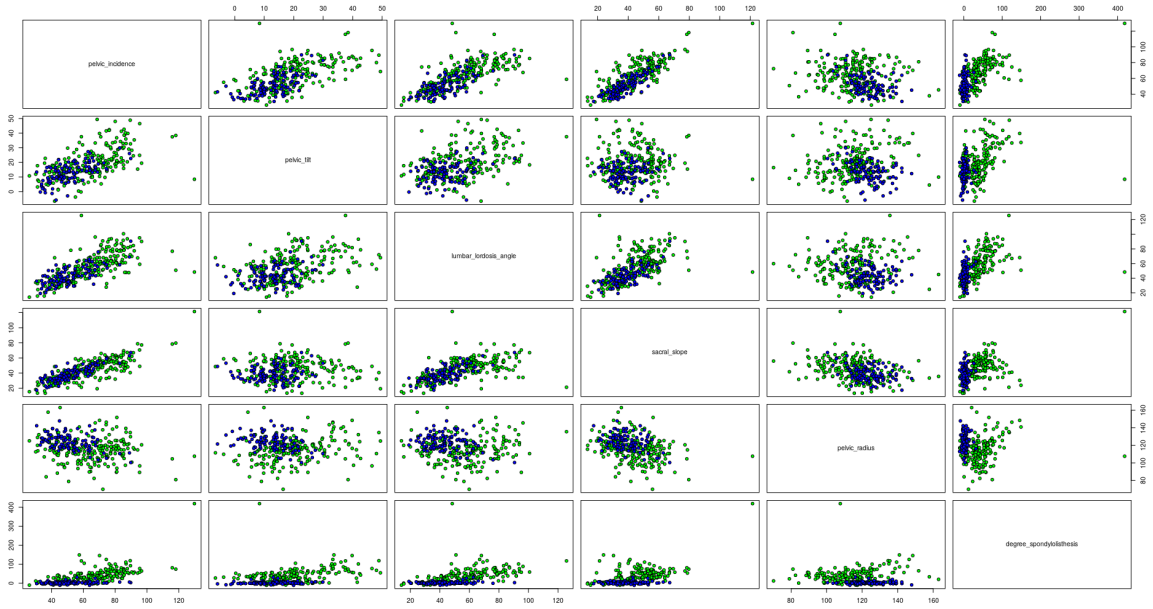


Figure 1.1: Feature Scatter Plot

1.3 c. Opinion about data

Given the values from section a and the scatter plot from section b we can see that the two classes are seperatated well when comparing certain values such as pelvic_radius and degree_spondylolithesis. If we compare the values using the scatter plot from Figure 1.1 we can see that abnormal classes have a larger value with respect of degree_spondylolisthesis then the normal class. This shows that given certain values there is some what of a well defined area of seperation.

2. Section 2

2.1 a. Generate 100 3-dimensional vectors from a normal distribution

Generating 100 3-dimensional vectors from a normal disribution with a mean vector as [1 2 1] and a 3x3 covariance matrix as [4 0.8 -0.3; 0.8 2 0.6; -0.3 0.6 5]

```
mean <- c(1,2,1)
sigma <- matrix(c(4, 0.8, -0.3, 0.8, 2, 0.6, -0.3, 0.6, 5), 3,3)
mvnd <- MASS::mvrnorm(n = 100, mean, sigma)
```

$$mean <- \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$$

$$sigma <- \begin{bmatrix} 4 & 0.8 & -0.3 \\ 0.8 & 2 & 0.6 \\ -0.3 & 0.6 & 5 \end{bmatrix}$$

2.2 b. Scatter Plots and Explained Relationships

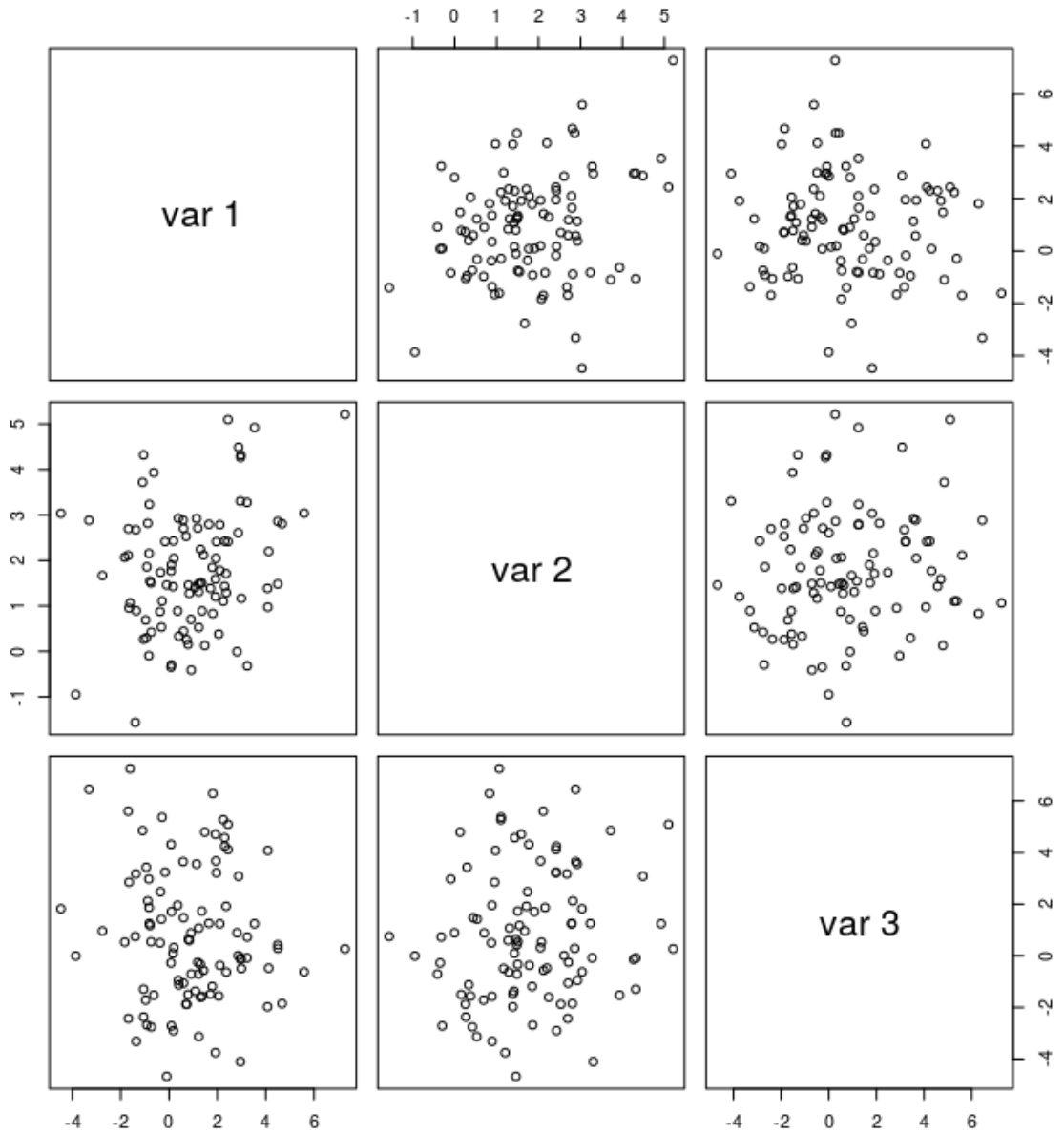


Figure 2.1: Multivariate normal distribution

As you can see from Figure 2.1 it seems that there is only a slight relationship between the variables. Looking at the graph i can see a slight upper right correlation between variable 2 and 3 and a slight upper left hand correlation between variable 1 and 3

2.3 c. Euclidean and Mahalanobis distance

```
library(fields)
x1 <- mvnd[1,]
```

```
x2 <- mvnd[2,]
Euclidean <- rdist(x1, x2)
```

$$x1 <- \begin{bmatrix} 0.08159917 & -0.3455406 & -0.2774757 \end{bmatrix}$$

$$x2 <- \begin{bmatrix} 2.096857 & 2.785129 & 1.236984 \end{bmatrix}$$

$$Euclidean <- \begin{bmatrix} 4.019446 \end{bmatrix}$$

```
library(stats)
x <- mvnd[1:5,]
mean<-colMeans(x)
cov<-cov(x)
Mahalanobis<-mahalanobis(x,mean,cov)
```

$$x <- \begin{bmatrix} 0.08159917 & -0.3455406 & -0.277475694 \\ 2.09685726 & 2.7851294 & 1.236983735 \\ 1.64938859 & 2.7939131 & 1.255435381 \\ 2.85857359 & 2.6075639 & 0.005900806 \\ 2.87544416 & 4.4884076 & 3.080050468 \end{bmatrix}$$

$$mean <- \begin{bmatrix} 1.912373 & 2.465895 & 1.060179 \end{bmatrix}$$

$$cov <- \begin{bmatrix} 1.3194324 & 1.800401 & 0.8444832 \\ 1.8004009 & 3.056076 & 1.9542692 \\ 0.8444832 & 1.954269 & 1.7625221 \end{bmatrix}$$

$$Mahalanobis <- \begin{bmatrix} 3.1344661 & 0.0400899 & 2.9093868 & 2.9826712 & 2.9333861 \end{bmatrix}$$

3. Section 3

3.1 Eigenvalues & Eigenvectors

```
records <- read.table("five-dimensional-records.txt")
  mean <- colMeans(records)
cov <- cov(records)
```

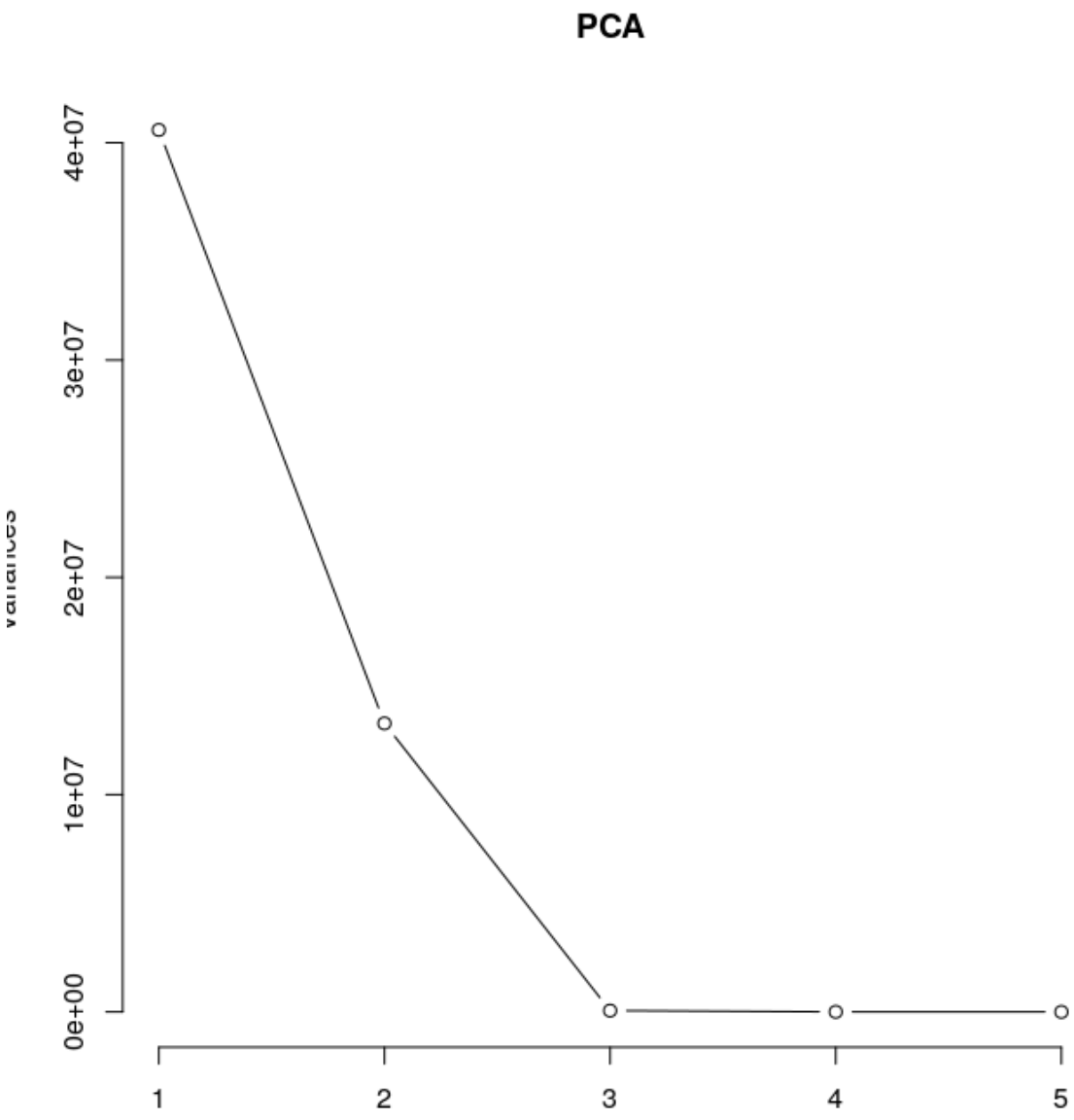
$$mean <- \begin{bmatrix} 6241.66667 & 11.44167 & 2333.33333 & 120.83333 & 17000.00000 \end{bmatrix}$$

$$cov <- \begin{bmatrix} 1.183356e+07 & 59.924242 & 4152121.2121 & 173507.5758 & 490909.091 \\ 5.992424e+01 & 3.191742 & 342.1212 & 141.9621 & 9818.182 \\ 4.152121e+06 & 342.121212 & 1540606.0606 & 73424.2424 & 963636.364 \\ 1.735076e+05 & 141.962121 & 73424.2424 & 13208.3333 & 569090.909 \\ 4.909091e+05 & 9818.181818 & 963636.3636 & 569090.9091 & 40545454.545 \end{bmatrix}$$

```
  Eigenvalues <- eigen(cov)$values
  Eigenvectors <- eigen(cov)$vectors
```

$$Eigenvalues <- \begin{bmatrix} 4.058981e+07 & 1.327940e+07 & 6.078551e+04 & 2.835137e+03 & 5.672175e-01 \end{bmatrix}$$

$$Eigenvectors <- \begin{bmatrix} 0.0210326211 & 9.430084e-01 & 0.332053840 & 0.0057119337 & -0.0006728422 \\ 0.0002420299 & -8.480421e-06 & -0.002006136 & -0.0006633593 & -0.9999977384 \\ 0.0269236336 & 3.312548e-01 & -0.942852072 & 0.0239126004 & 0.0018793378 \\ 0.0141541619 & 1.292481e-02 & -0.020405538 & -0.9996077844 & 0.0007073532 \\ 0.9993159400 & -2.895527e-02 & 0.018703144 & 0.0133939822 & 0.0001957042 \end{bmatrix}$$



graph.png

Figure 3.1: Eigenvectors line graph

3.2 reduced representation

```
PCA <- as.data.frame(prcomp(records)$x)[1:2]
```

$$PCA <- \begin{bmatrix} 7989.734 & -685.3013 \\ -7153.694 & -5315.8562 \\ -8091.763 & -2891.1789 \\ 7926.393 & -2743.7012 \\ 7927.907 & -2588.2250 \\ -4949.073 & 2079.0494 \\ -1158.977 & -5367.2371 \\ -2912.664 & 3101.7248 \\ 1105.817 & 3774.9869 \\ 8103.076 & 3358.3627 \\ -4900.497 & 3631.3980 \\ -3886.258 & 3645.9779 \end{bmatrix}$$

3.3 Scatter Plot

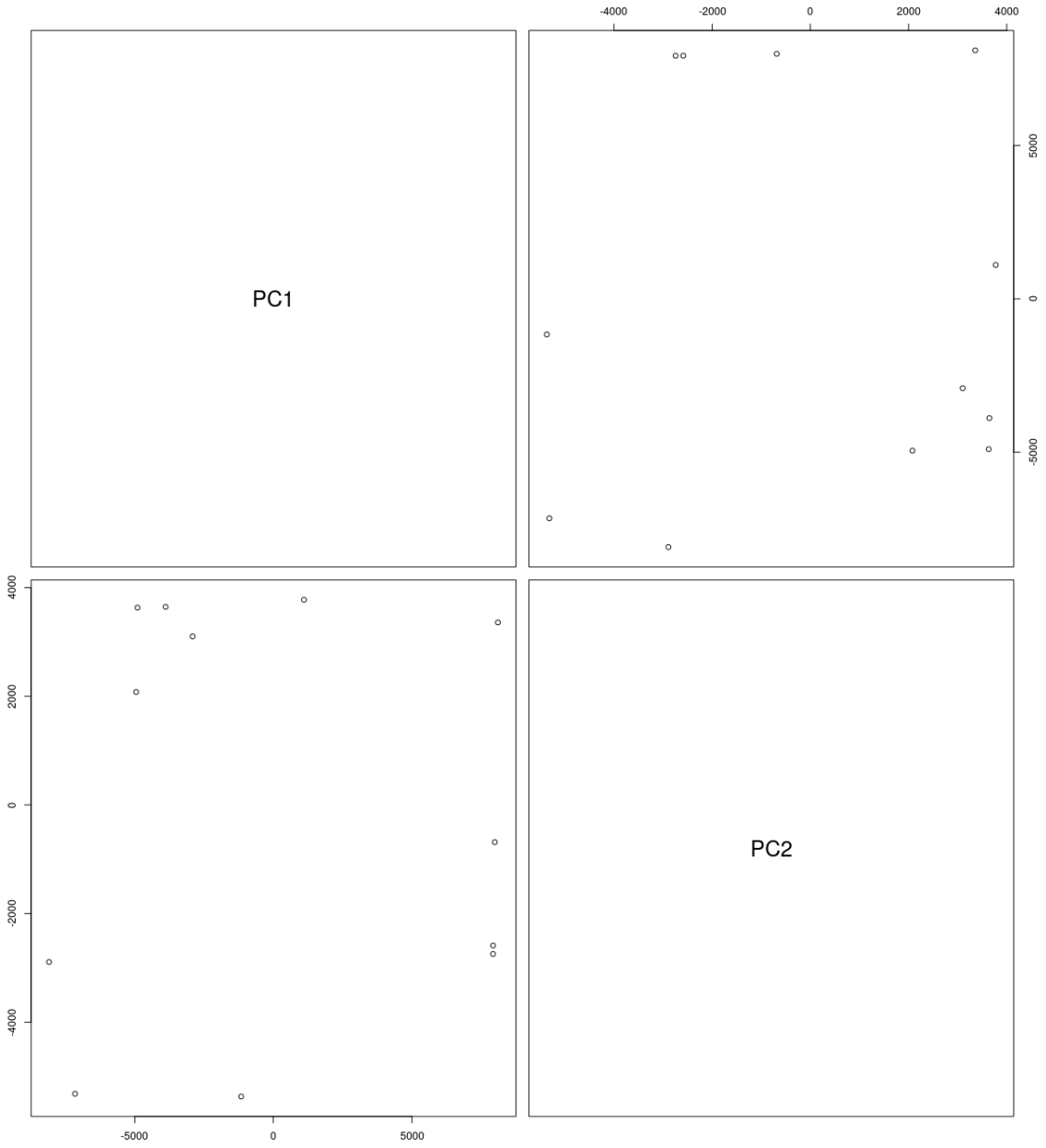


Figure 3.2: PCA Scatter Plot

3.4 Reconstruct Data

```
reconstructed <- PCA$x %*% t(PCA$rotation) + colMeans(records)
sqerr <- mean((records - reconstructed)^2)
```

$$records <- \begin{bmatrix} 5700 & 12.8 & 2500 & 270 & 25000 \\ 1000 & 10.9 & 600 & 10 & 10000 \\ 3400 & 8.8 & 1000 & 10 & 9000 \\ 3800 & 13.6 & 1700 & 140 & 25000 \\ 4000 & 12.8 & 1600 & 140 & 25000 \\ 8200 & 8.3 & 2600 & 60 & 12000 \\ 1200 & 11.4 & 400 & 10 & 16000 \\ 9100 & 11.5 & 3300 & 60 & 14000 \\ 9900 & 12.5 & 3400 & 180 & 18000 \\ 9600 & 13.7 & 3600 & 390 & 25000 \\ 9600 & 9.6 & 3300 & 80 & 12000 \\ 9400 & 11.4 & 4000 & 100 & 13000 \end{bmatrix}$$

$$reconstructed <- \begin{bmatrix} 5700.0000 & 2334.6917 & 17166.6667 & 160.60833 & 8120.833 \\ -5230.2250 & 120.2917 & 4508.3333 & 2222.50000 & 10000.000 \\ -508.3333 & 16997.3583 & -1321.8917 & 10.00000 & -1758.333 \\ -2320.8333 & 6243.8250 & 1700.0000 & 17019.16667 & 8011.442 \\ 14758.3333 & 12.8000 & -612.5000 & 6260.83333 & 10333.333 \\ 8200.0000 & 2330.1917 & 17266.6667 & -49.39167 & -4879.167 \\ -5030.2250 & 120.7917 & 4308.3333 & 2222.50000 & 16000.000 \\ 5191.6667 & 17000.0583 & 978.1083 & 60.00000 & 3241.667 \\ 3779.1667 & 6242.7250 & 3400.0000 & 17059.16667 & 1011.442 \\ 20358.3333 & 13.7000 & 1387.5000 & 6510.83333 & 10333.333 \\ 9600.0000 & 2331.4917 & 17966.6667 & -29.39167 & -4879.167 \\ 3169.7750 & 120.7917 & 7908.3333 & 2312.50000 & 13000.000 \end{bmatrix}$$

$$sqerr <- \begin{bmatrix} 76488567 \end{bmatrix}$$

4. Section 4

```
PCA <- prcomp(vert[0:6])
Eigenvalues <- PCA$sdev^2
Eigenvectors <- PCA$rotation
reduced <- as.data.frame(PCA$x)[1:2]
pairs(reduced, pch = 21, bg = c('green', 'blue')[unclass(vert$class)])
```

$$Eigenvalues <- \begin{bmatrix} 1.780994e+03 & 3.453271e+02 & 1.887770e+02 & 1.060179e+02 & 8.861407e+01 & 7.207841e-18 \end{bmatrix}$$

$$Eigenvectors <- \begin{bmatrix} -0.32364565 & 0.47663485 & -0.001544813 & 0.37367725 & -0.44170387 & -5.773503e-01 \\ -0.11319229 & 0.09856328 & -0.264657410 & 0.75411376 & 0.07354147 & 5.773503e-01 \\ -0.30367474 & 0.53278398 & -0.496541893 & -0.33941176 & 0.51202411 & 1.089295e-11 \\ -0.21045336 & 0.37807157 & 0.263112598 & -0.38043651 & -0.51524534 & 5.773503e-01 \\ 0.02995983 & -0.32180920 & -0.774612852 & -0.17510604 & -0.51463973 & 3.590517e-12 \\ -0.86315378 & -0.48243804 & 0.118940778 & -0.03291431 & 0.08359925 & -3.067324e-12 \end{bmatrix}$$

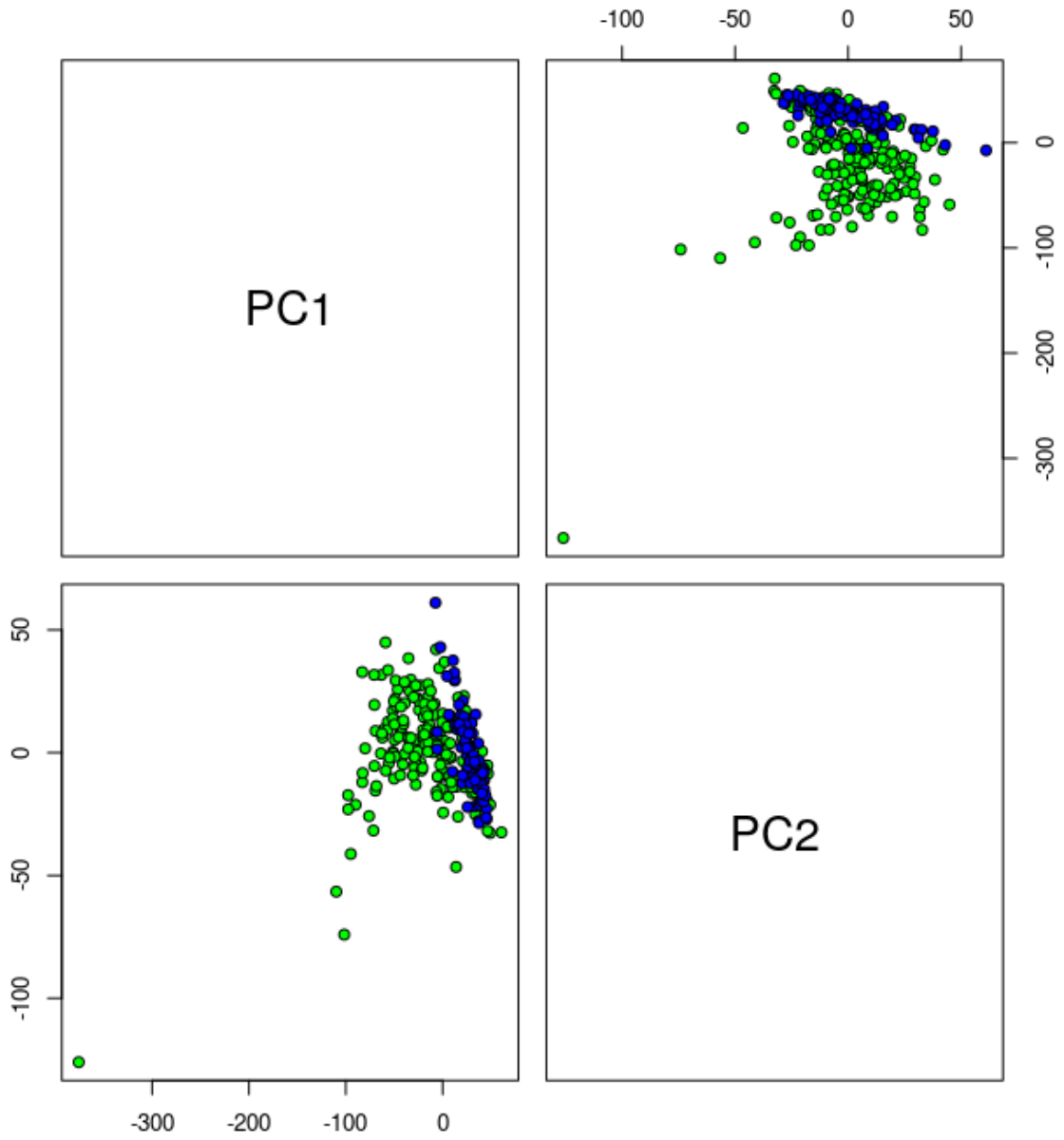


Figure 4.1: PCA Vertebral Column Data Set

5. Section 5

```
library('tsne')
records <- read.table('five-dimensional-records.txt')
tsne_10 <- tsne(records, perplexity=10)
tsne_50 <- tsne(records, perplexity=50)
```

$$tsne\_10 <- \begin{bmatrix} -352.32073 & 278.07448 \\ -16.54169 & -481.36909 \\ 400.59149 & -135.67055 \\ -398.91724 & -11.26791 \\ -175.19482 & 58.76914 \\ 250.22489 & -336.66253 \\ 362.52881 & 127.94042 \\ 82.24818 & 108.50328 \\ 145.93197 & 354.96502 \\ -125.27177 & 462.93079 \\ 74.42643 & -146.08627 \\ -247.70552 & -280.12678 \end{bmatrix}$$

$$tsne\_50 <- \begin{bmatrix} -157.92121 & -259.202370 \\ -84.49455 & -66.797062 \\ -224.85017 & 89.352029 \\ 63.06430 & 296.897437 \\ 303.43655 & -7.161843 \\ -145.51641 & 266.365026 \\ -288.65359 & -93.832651 \\ 35.04427 & -239.403781 \\ 225.58846 & -203.064369 \\ 100.09517 & -39.776473 \\ -15.59996 & 106.572968 \\ 189.80716 & 150.051089 \end{bmatrix}$$

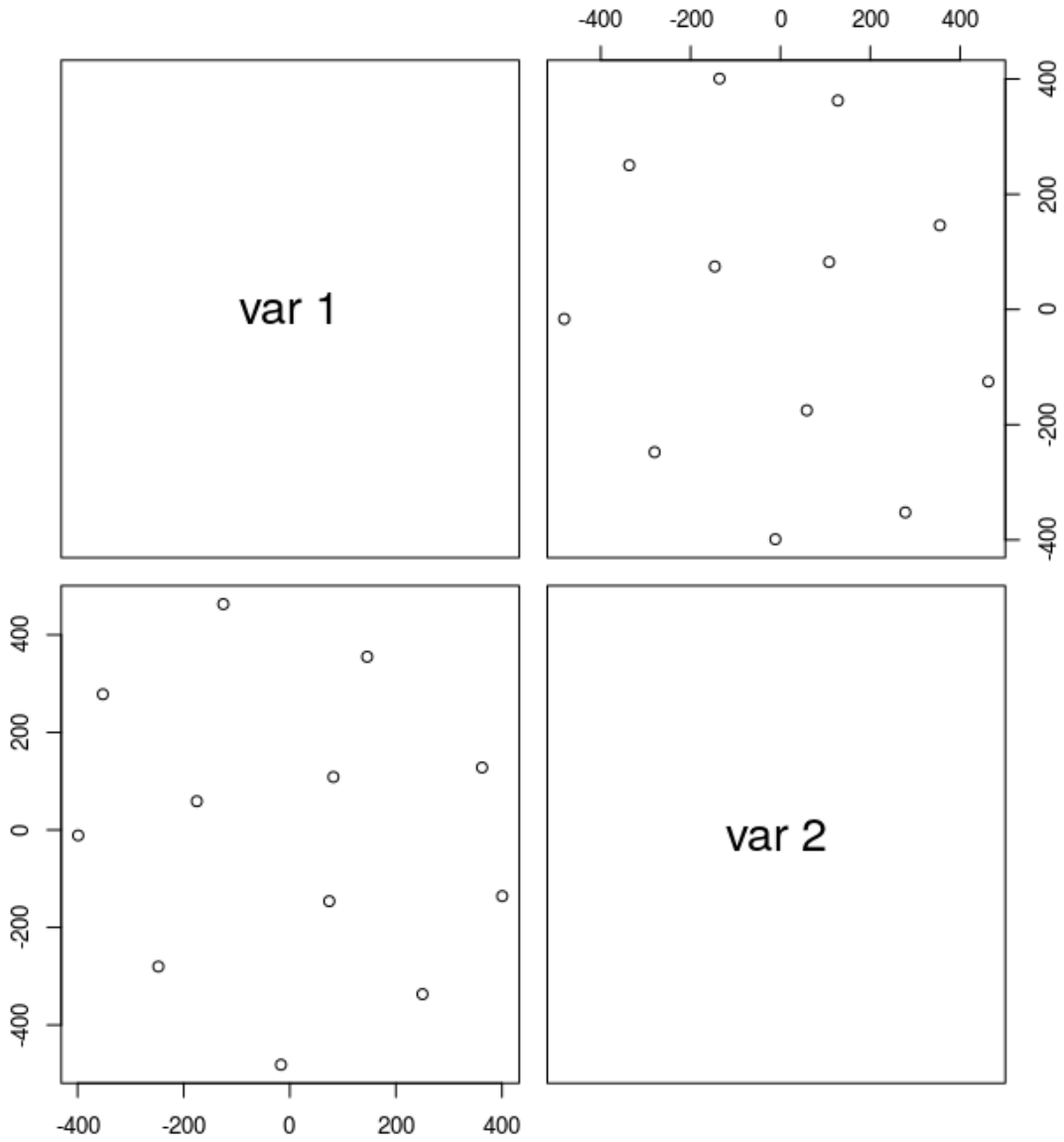


Figure 5.1: tsne perplexity 10

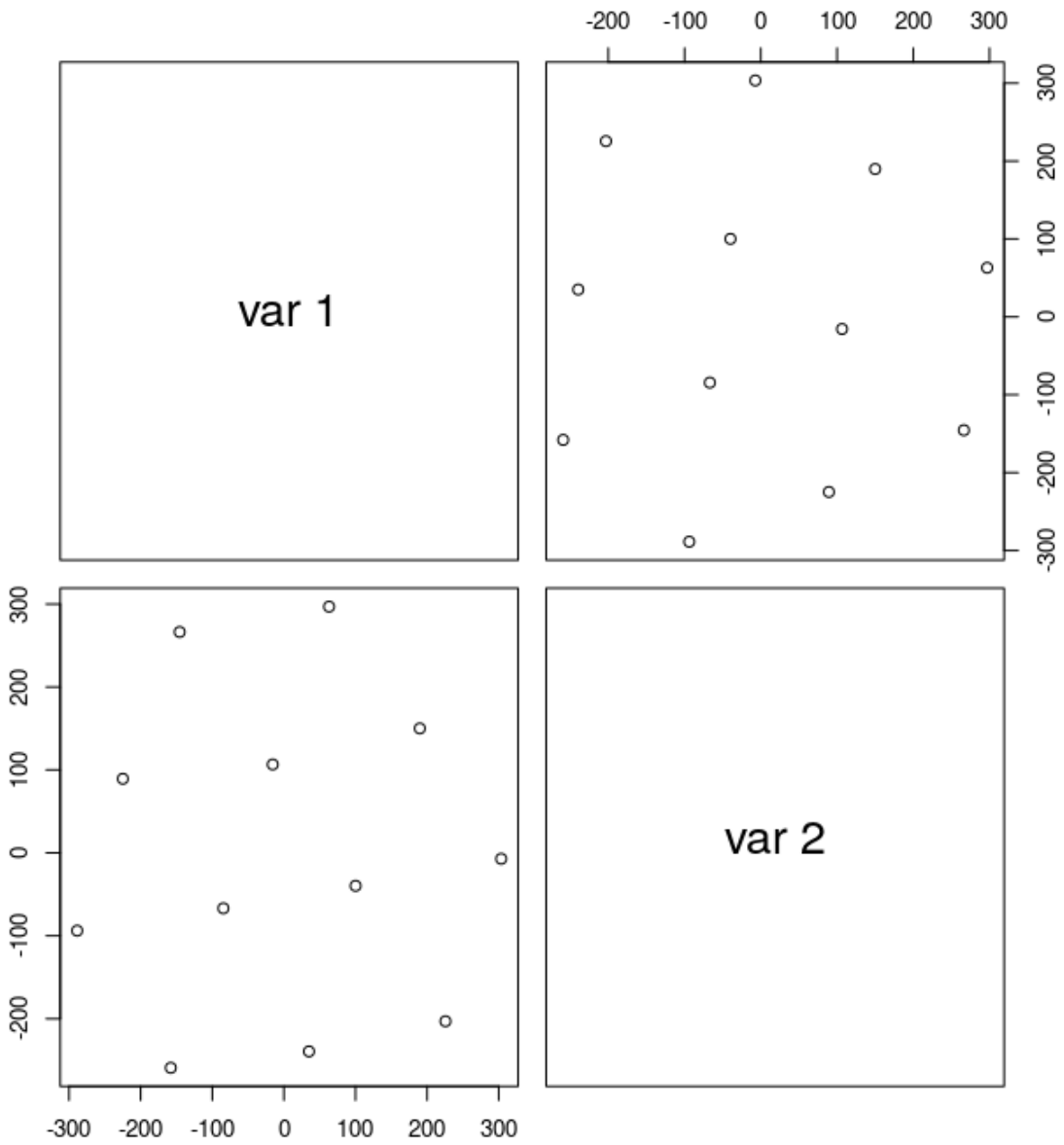


Figure 5.2: tsne perplexity 50

The data seems to be more correlated and well structed and not as spread out as the other data in question 3