WESTERN MICHIGAN UNIVERSITY

STAT 6850

Applied Data Mining

Fall 2014

Instructor: Dr. J.C. Wang

Case Study #1

Exploratory data analysis on the
Vertebral Column Data Set from UCI

Antonio Giraldi

Saurabh Rajratn Kulkarni

Shoruk Mansour

Joan Martinez

Milton Soto Ferrari

Mustafa Yildiz

Description Summary

The objective of this report is to perform basic exploratory data analysis (descriptive nature) on the Vertebral Column Data Set. The background of the given data is as follows:
Variable class2 is for 2-level patient classification and variable class3 is for 3-level patient classification. Each patient is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (in this order): pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis. The following convention is used for the class labels: DH (Disk Hernia), Spondylolisthesis (SL), Normal (NO) and Abnormal (AB).

Initial Analysis

The initial examination of data is a valuable stage of most statistical investigations, not just for summarizing data, but also it may be all that is necessary or desirable. The main idea of this report is beginning an analysis with an exploratory look at the given data from UCI Machine Learning Repository, in order to get a feel for this data, to clarify the general structure of the data, obtain simple descriptive summaries and perhaps getting ideas for more sophisticated analysis if needed. We calculated some descriptive statistics for the six attributes, such as means, medians, maximum and minimum, 1st and 3rd Quartile, class 2 (AB/NO) and class 3 (DH/NO/SL). See Table 1.

We start assessing the structure and quality of the data for a possible presence of errors, and outliers. R program was used again to check for consistency. We were also visually able to spot some suspect values in the data array and used 3*IQR rule to identify outlier on Table 2.

After assessing and removing the outliers, we repeated the analysis without the suspected observation [116] and continue our investigation to find out more about this data. New summary table is created without outliers, see Table 3.

Correlation Matrix

Summary statistics for each variable will be supplemented by correlations as appropriate.
Slope and pincidence shows the strongest correlation and these also are positively multi-correlated with three other attributes (angle, grade, and ptilt); however, pradius shows almost no correlation with other variables except plope and pindicende were some level of negative correlation can be seen. See Fig1.

Graphical Procedures

Appropriate graphical procedures were selected such as Boxplots and Histograms to compare the six groups of attributes (pincidence, ptilt, angle, slope, pradius, and grade). See Fig2 and Fig3.
These figures show the variation within the data set after eliminating the outlier. By looking at all 1st and 3rd Quartile in Fig2, most of our visual focus went to the grade attribute due to its high dispersion. Also grade is the most right skewed attribute from all others as seen in Fig3, which indicates further investigation is needed to see the cause of this variation. Barplot is used

to compare class3 frequency on the vertebral data set (DH, NO, SL). See Fig4. It shows that Spondylolisthesis in class3 has the highest frequency comparing to Normal and Disk Hernia.

Analysis by Class

We wanted to analyze the behavior of these attributes separated by each class (Normal/Abnormal). See Fig5. Looking at this figure we see effect the class factor has on all six attributes. The main findings became clear without need for statistical inferences as some attributes as grade are significantly different between abnormal and normal subjects. We also noted that abnormal subjects show higher values on all attributes except on pradius, we expected this behavior due to the positive and negative correlations between these.

Analysis by Sub-Class

Abnormal classes are subdivided into DH (Disk Hernia) and Spondylolisthesis (SL). We wanted to see how the attributes behave when using sub-classes and comparting to Normal. By constructing a histogram (See Fig6) and fitting a normal curve on each attribute by sub-class we see that attributes under SL appears to follow a mesokurtic behavior, while DH and NO show more clustering around the central tendency value. “Grade” attribute once again is significantly different under de sub-class SL. In general DH and NO behave similarly.

Looking at Fig7, we see the benefits of analyzing data by sub-classes as SL and DH show differences on each attribute. DH and NO show similar characteristics in terms of dispersion and range. On the other side, SL yields higher values on all attributes except on pradius (as expected). We see the dispersion of SL is significantly higher than other classes. In Fig8, we did parallel plots for each sub-class to show again that DH and NO behave similar on all attributes, the shape of the plot are a quite similar with a common trend; sub-class SL shows too much variability in all attributes making it hard to spot a trend or shape.

Despite the differences we already discussed, it is interesting to look at both Scatter plots (Fig9 and Fig10). We again see the positive correlation some variables have while pradius shows low level of correlation. While looking at the “grade”, we again see Spondylolisthesis yield significantly higher values than Disk Hernia and Normal subjects. DH and NO also appear to cluster in the same region on all attributes; we can see this more evidently in “grade”.

Now we have all attributed shown in a Parallel Coordinates Plot using the standard deviation as the y- scale. This also confirms our finding regarding the differences Spondylolisthesis shows among subjects. It is our suspicious class type while DH and NO are acting alike. See Fig11.
By constructing a 3D graph (See Fig12) using uncorrelated variables for a more formal and detailed analysis, the SL is still separated from the other two sub-classes in our data set.

In conclusion, by doing exploratory data analysis on the Vertebral Column Data Set from UCI we had some insight on the behavior of all measurements and attributes of subjects’ vertebras. Spondylolisthesis shows significant differences in comparison to Normal and Disk Hernia.

Appendix

Table 1. Initial Analysis: Data distributions and Class Frequency

pincidence		ptilt		angle		slope		pradius		grade		class2	
Min. :	26.15	Min. :	-6.55	Min. :	14	Min. :	13.37	Min. :	70.08	Min. :	-11.06	AB:	210
1st Qu.:	46.43	1st Qu.:	10.67	1st Qu.:	37	1st Qu.:	33.35	1st Qu.:	110.71	1st Qu.:	1.6	NO:	100
Median:	58.69	Median :	16.36	Median:	49.56	Median:	42.41	Median:	118.27	Median:	11.77	**class3**	
Mean :	60.5	Mean :	17.54	Mean :	51.93	Mean :	42.95	Mean :	117.92	Mean :	26.3	DH:	60
3rd Qu.:	72.88	3rd Qu.:	22.12	3rd Qu.:	63	3rd Qu.:	52.69	3rd Qu.:	125.47	3rd Qu.:	41.28	NO:	100
Max. :	129.83	Max. :	49.43	Max. :	125.74	Max. :	121.43	Max. :	163.07	Max. :	418.54	SL:	150

Table 2. Identifying Outliers using 3*IQR:

Row	slope	grade
116	121.43	418.54

Table 3. Subset Summary Table:

pincidence		ptilt		angle		slope		pradius		grade		class2	
Min. :	26.15	Min. :	-6.55	Min. :	14	Min. :	13.37	Min. :	70.08	Min. :	-11.06	AB:	209
1st Qu.:	46.43	1st Qu.:	10.69	1st Qu.:	37	1st Qu.:	33.34	1st Qu.:	110.71	1st Qu.:	1.59	NO:	100
Median :	58.6	Median :	16.42	Median :	49.78	Median :	42.37	Median :	118.34	Median :	11.46	class3	
Mean :	60.27	Mean :	17.57	Mean :	51.94	Mean :	42.7	Mean :	117.95	Mean :	25.03	DH:	60
3rd Qu.:	72.64	3rd Qu.:	22.18	3rd Qu.:	63	3rd Qu.:	52.55	3rd Qu.:	125.48	3rd Qu.:	40.88	NO:	100
Max. :	118.14	Max. :	49.43	Max. :	125.74	Max. :	79.7	Max. :	163.07	Max. :	148.75	SL:	149

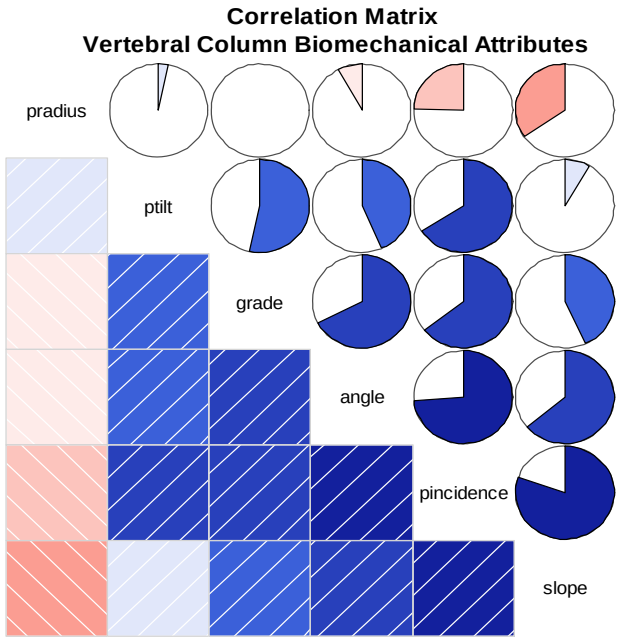

Fig1

Boxplots, Histograms and Frequency Bars

a) Analysis by Attribute

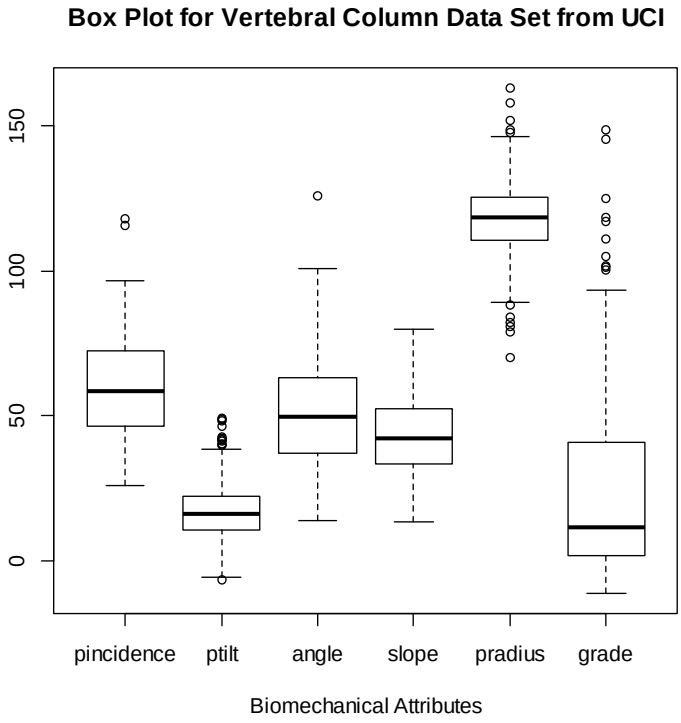

Fig2

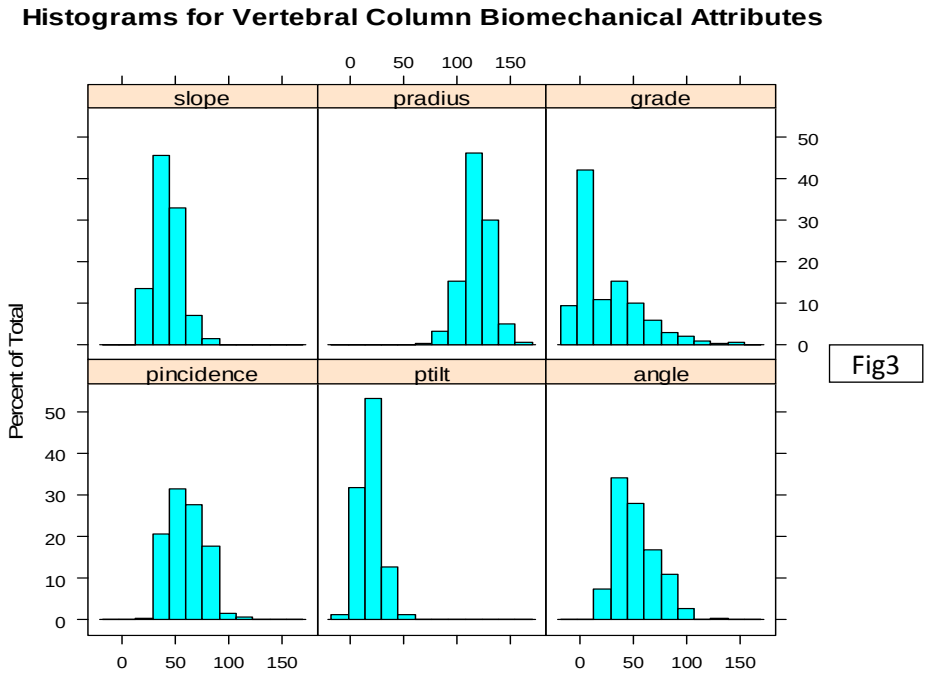

Fig3

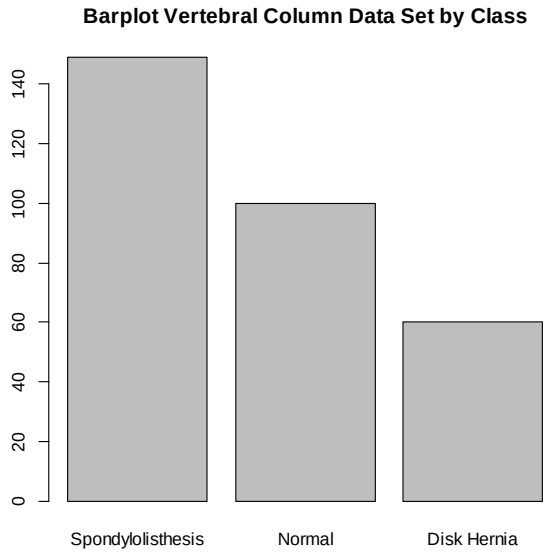

Fig4

b) Analysis by Class

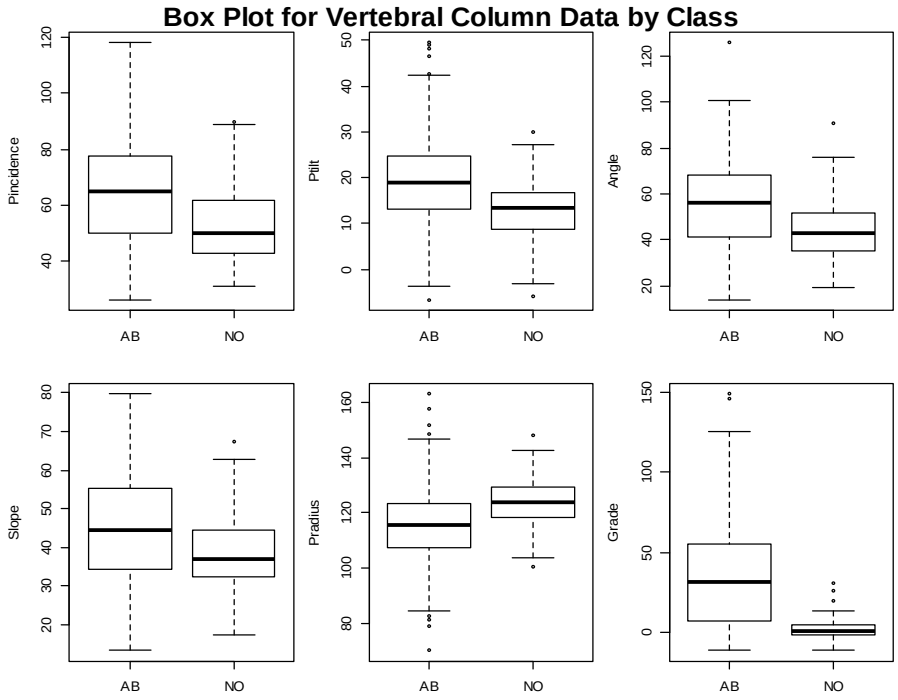

Fig5

c) Analysis by Sub-Class

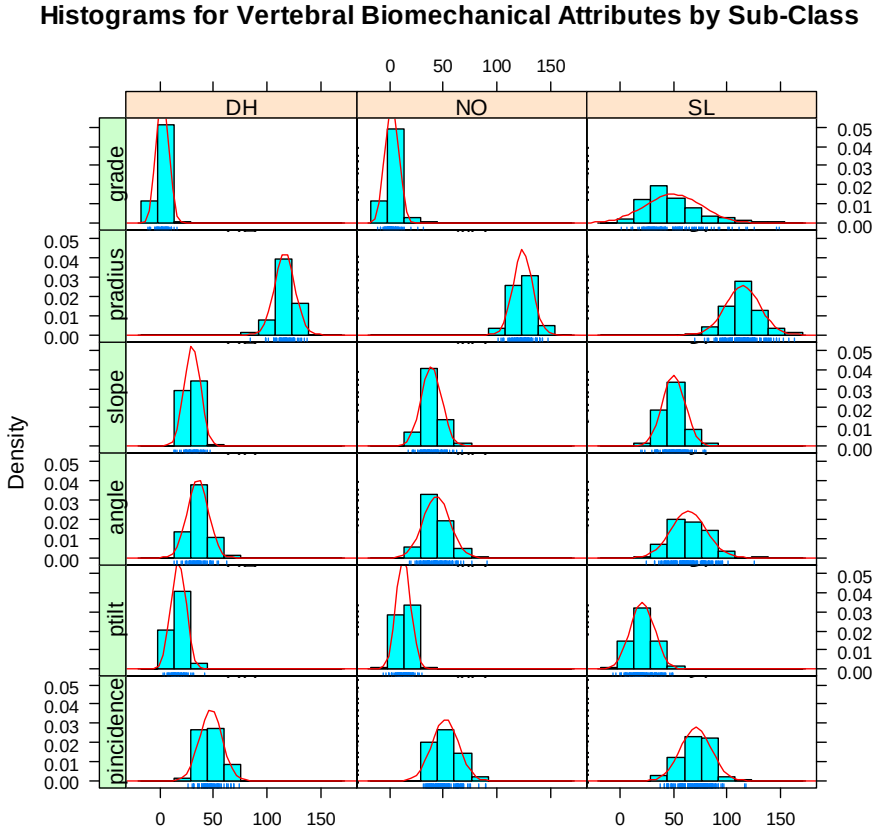

Fig6

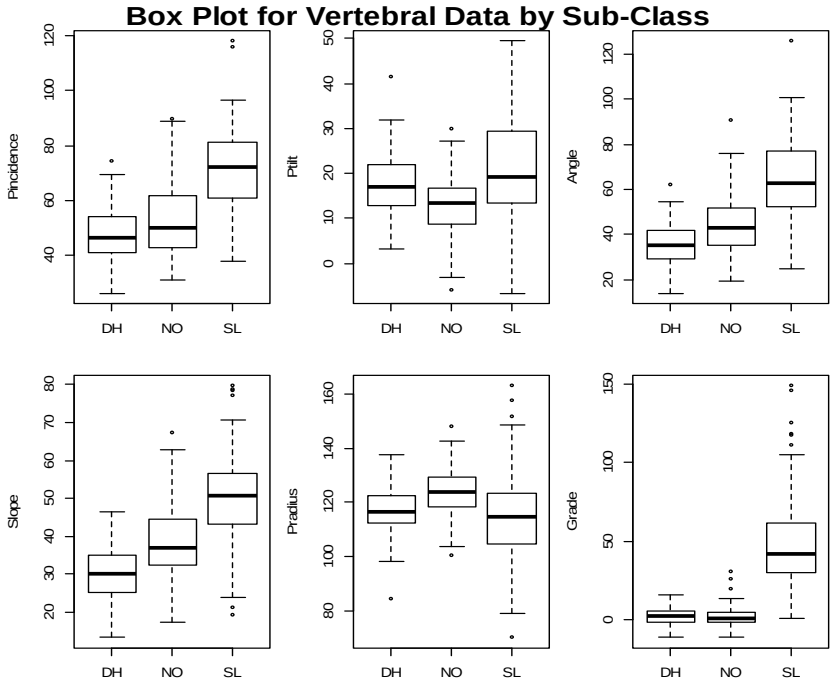

Fig7

Parallel Coordinates Plot

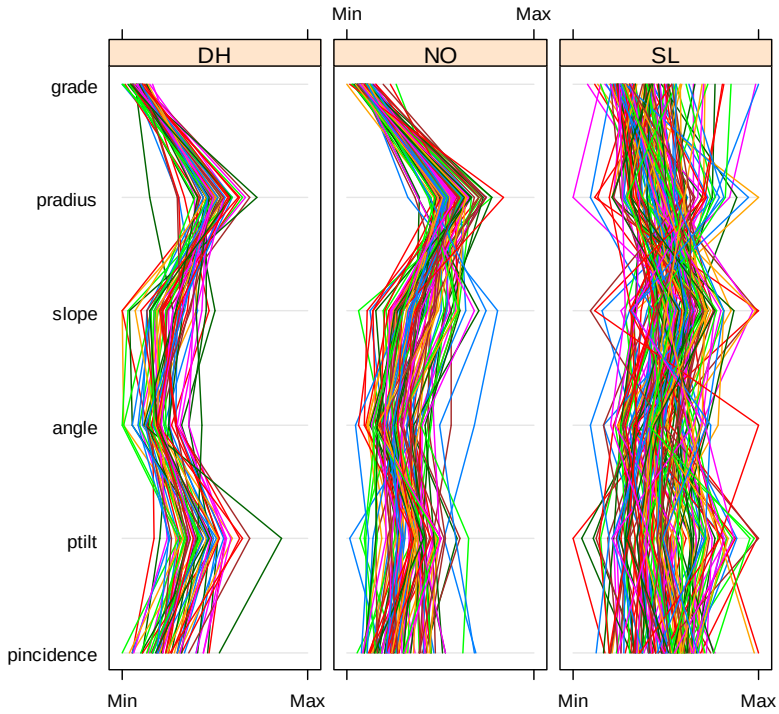

Fig8

Scatterplots Matrix on Vertebral Column Biomechanical Attributes

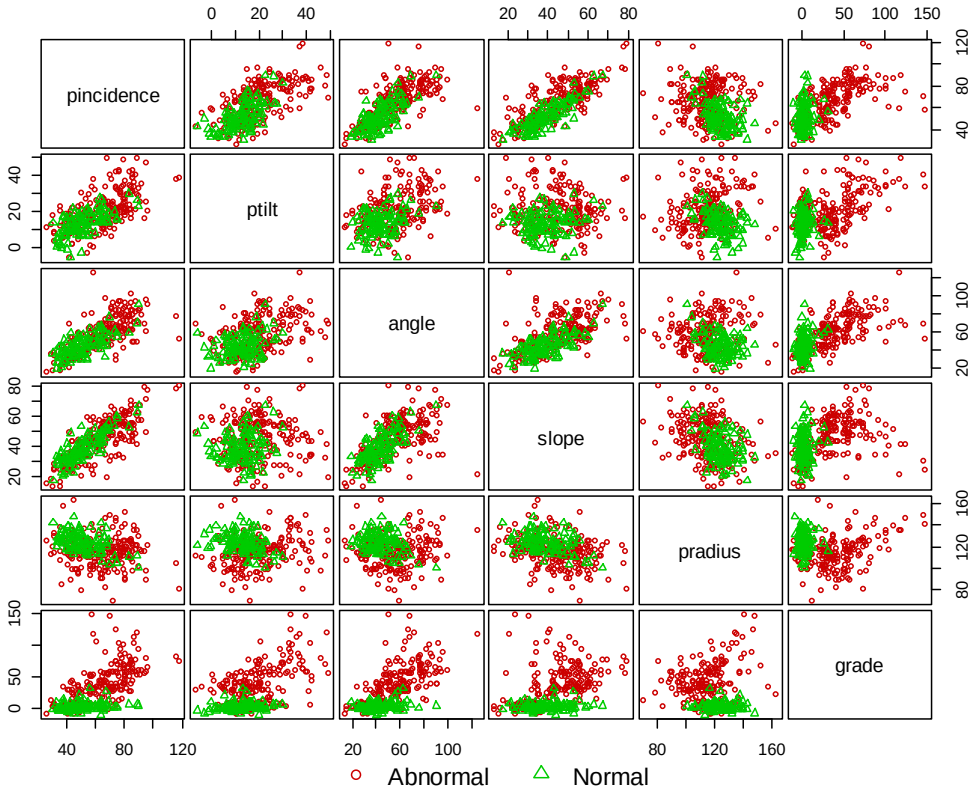

Fig9

Scatterplots Matrix on Vertebral Column Biomechanical Attributes

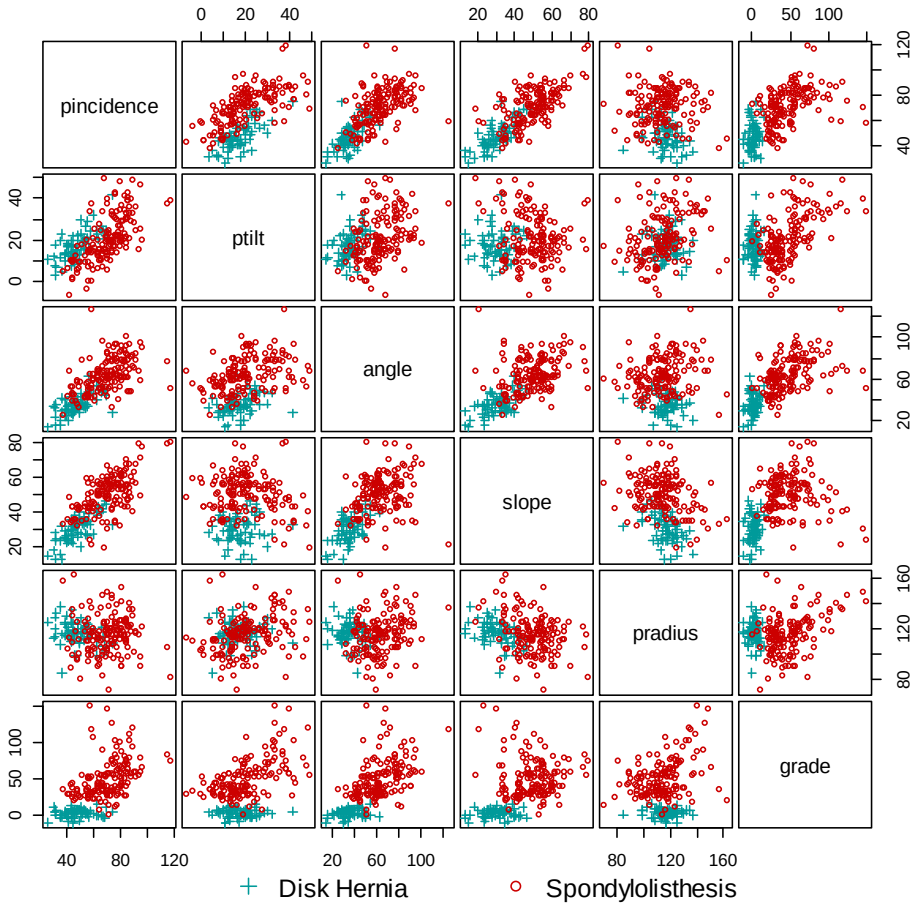

Fig10

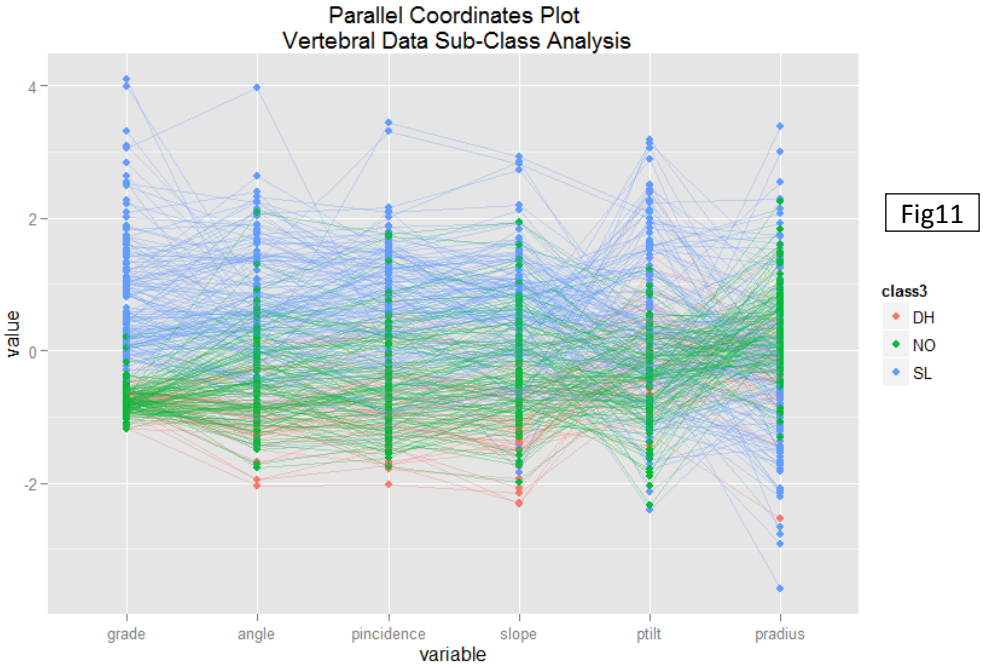

Fig11

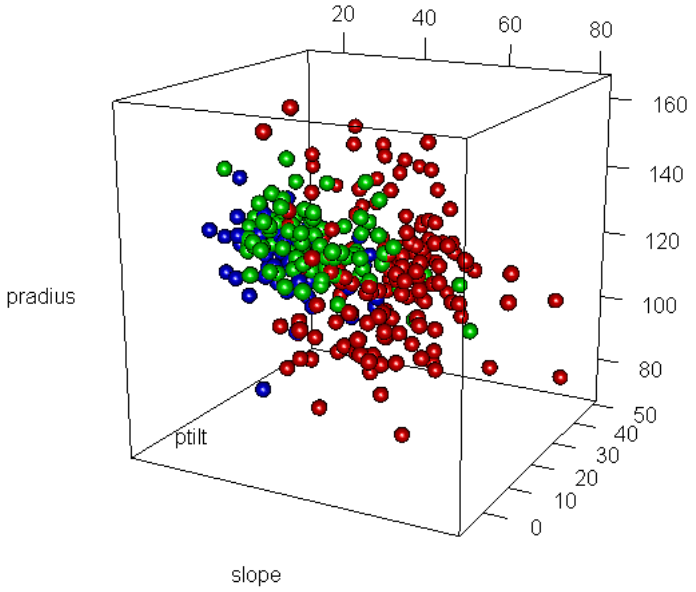

Fig12

Red – **SL** Blue – **DH** Green - **NO**

R-Code:

```
## Data Frame creation from tables
    vertebral <- read.table("column_2C.dat", header=F)
    colnames(vertebral) <- c('pincidence','ptilt','angle','slope',
                             'pradius','grade','class2')
    vertebral$class3 <- read.table("column_3C.dat", header=F)[[7]]

    head(vertebral)

    summary(vertebral)

##Identifying Outliers
    require(outliers)
    which(abs(scores(vertebral[1:6], type="iqr")) > 3, arr.ind =T)
    vertebral[116,c(4,6)]
    vertebral2 <- vertebral[-116,]

##Summary Stats
    summary(vertebral2)

##Correlation Matrix
    (cormatrix <- cor(vertebral2[1:6],vertebral2[1:6]))
       require(corrgram)
       corrgram(vertebral2, order=TRUE, lower.panel=panel.shade,
       upper.panel=panel.pie, text.panel=panel.txt)
       title("Correlation Matrix \nVertebral Column Biomechanical Attributes", line = 2)

##Boxplot
    boxplot(vertebral2[1:6], main = "Box Plot for Vertebral Column Data Set from UCI",
          xlab = "Biomechanical Attributes")
    #Class2
    oldpar <- par(mfrow=c(2,3), mar=c(4.1,4.1,0.5,0.5), oma=c(0,0,2,0))
    boxplot(pincidence ~ class2, data=vertebral2, ylab="Pincidence")
    boxplot(ptilt ~ class2, data=vertebral2, ylab="Ptilt")
    boxplot(angle ~ class2, data=vertebral2, ylab="Angle")
    boxplot(slope ~ class2, data=vertebral2, ylab="Slope")
    boxplot(pradius ~ class2, data=vertebral2, ylab="Pradius")
    boxplot(grade ~ class2, data=vertebral2, ylab="Grade")
    title(main="Box Plot for Vertebral Data by Class", outer=T, line=0, cex.main=2)

    #Class3
    boxplot(pincidence ~ class3, data=vertebral2, ylab="Pincidence")
    boxplot(ptilt ~ class3, data=vertebral2, ylab="Ptilt")
    boxplot(angle ~ class3, data=vertebral2, ylab="Angle")
    boxplot(slope ~ class3, data=vertebral2, ylab="Slope")
    boxplot(pradius ~ class3, data=vertebral2, ylab="Pradius")
    boxplot(grade ~ class3, data=vertebral2, ylab="Grade")
    title(main="Box Plot for Vertebral Data by Sub-Class", outer=T, line=0, cex.main=2)

##Barplots
    cl3names <- c("Spondylolisthesis", "Normal", "Disk Hernia")
    barplot(sort(table(vertebral2$class3),dec=T), names.arg = cl3names,
          main = "Barplot Vertebral Column Data Set by Class")

##Histogram
    require(lattice)
    histogram(~pincidence+ptilt+angle+slope+pradius+grade,data=vertebral2,
       main = "Histograms for Vertebral Column Biomechanical Attributes", xlab="")
    useOuterStrips(histogram(~pincidence+ptilt+angle+slope+pradius+grade|class3,data=vertebral2,
       type="density", panel = function(x, ...) {
            panel.histogram(x, ...)
            panel.rug(x, ...)
            panel.mathdensity(dmath = dnorm, col = "red",
                              args = list(mean=mean(x),sd=sd(x)))
         },
       main = "Histograms for Vertebral Biomechanical Attributes by Sub-Class", xlab=""))
```

```
##Scatter Plots Class2
    cols <- c(rgb(.8,0,0),rgb(0,.8,0))
    pchs <- c(1,2)
    #x11()
    pairs(vertebral2[1:6],col=cols[vertebral2$class2],pch=pchs[vertebral2$class2],gap=.3,cex=.8,
        main = "Scatterplots Matrix on Vertebral Column Biomechanical Attributes")
    legend(.34,.05, legend=c("Abnormal","Normal"),pch=pchs,col=cols,bty="n",xpd=NA, horiz=T)

##Scatter Plots Class3 for Abnormal
    abnorm <- subset(vertebral2, class3!="NO")
    abnorm$class3 <- droplevels(abnorm$class3)
    cols <- c(rgb(0,0.6,0.6),rgb(0.8,0,0))
    pchs <- c(3,1)
    #x11()
    pairs(abnorm[1:6],col=cols[abnorm$class3],pch=pchs[abnorm$class3],gap=.3,cex=.8,
        main = "Scatterplots Matrix on Vertebral Column Biomechanical Attributes")
    legend(.25,.05, legend=c("Disk Hernia","Spondylolisthesis"),
        pch=pchs,col=cols,bty="n",xpd=NA, horiz=T)

##Parallel Coordinates Plot
    require(GGally)
    ggparcoord(data = vertebral2, columns = 1:6, groupColumn = 8,
               order = "anyClass", showPoints = TRUE,
               title = "Parallel Coordinates Plot\n Vertebral Data Sub-Class Analysis",
               alphaLines = 0.3)
    parallelplot(~vertebral2[1:6] | class3, data=vertebral2, layout=c(3,1),
             main="Parallel Coordinates Plot")

##3D Visualization
    require(rgl)
    cols3d <- c(rgb(0,0,.8,0.5),rgb(0,.8,0,0.5),rgb(.8,0,0,0.5))
    plot3d(vertebral2[c(4,2,5)], size=1.5, type='s', col=cols3d[vertebral2[[8]]])
```