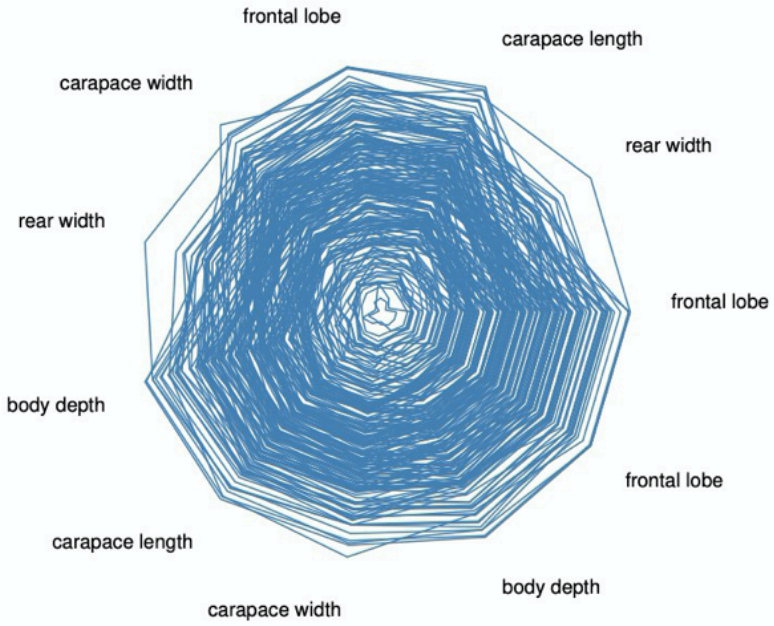
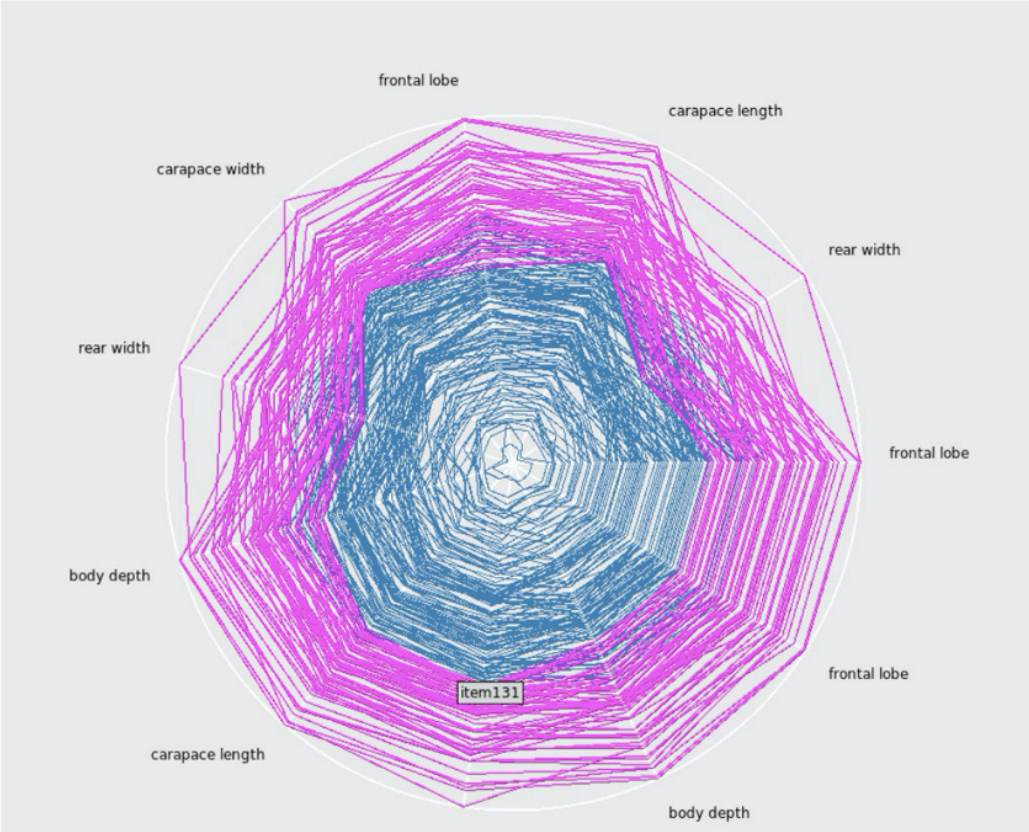
Rock Crab Visualization using radial axis

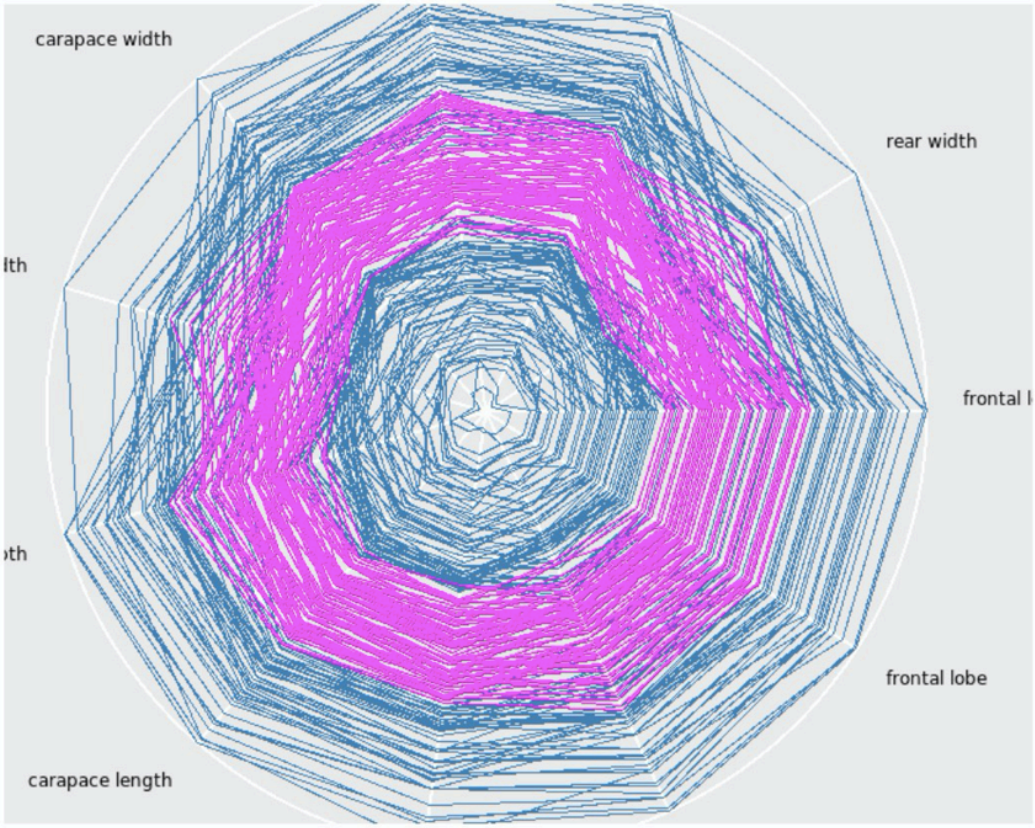
```
label = sub(" \\(mm\\)", "", colnames(lepto))
colnames(lepto) = label
ord <- eulerian(ncol(lepto))
seq <- names(lepto)[ord]
sa1<-l_serialaxes(data = lepto,linkingGroup = "lepto", sequence=seq, showItemLabels = TRUE)
l_export(sa1, filename = "sa1.pdf", height = 500, width = 500)
```

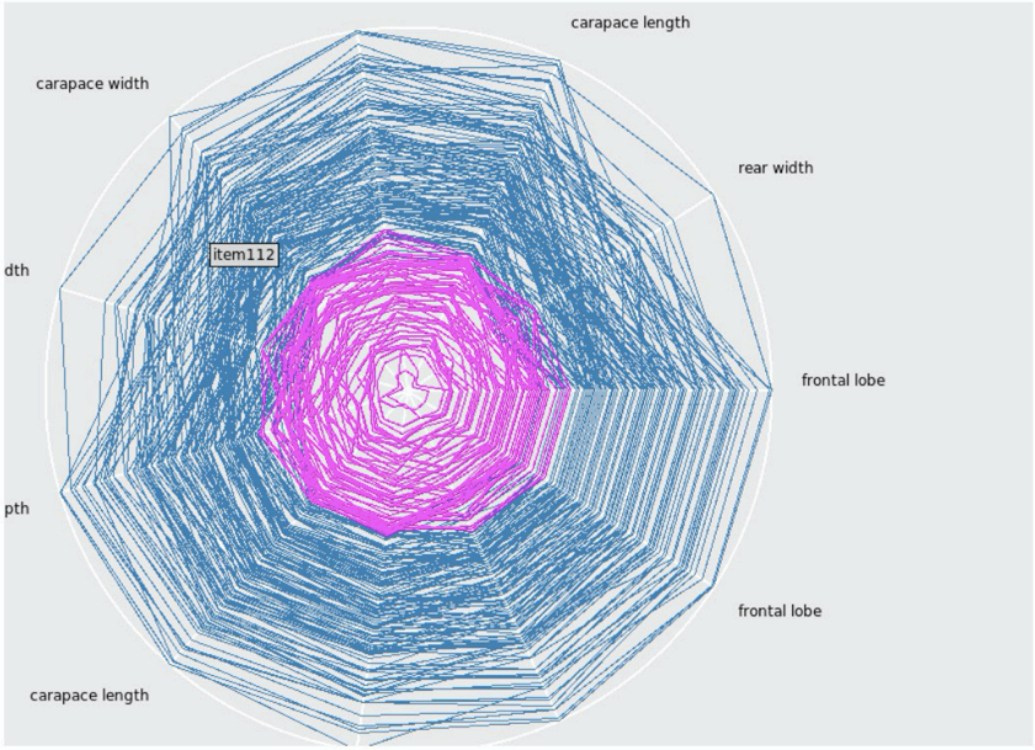


b. i.

The outermost glyph are fairly even sided. From there to outer 1/3rd of the data, the graph appears to change into a different shape. Moving closer to the center, the graph appears to be more even-sided again.


frontal lobe
carapace length
rear width
frontal lobe
frontal lobe
body depth
carapace length
carapace width
rear width
body depth
item131


carapace width
rear width
lth
frontal l
oth
frontal lobe
carapace length



ii.

Yes, they are different. Certain lines that are more on the outside between two variates could be more on the inside between other two variates as the relationship between variates are different for different crabs. So by choosing different pair of axes, if the relationship between the axes is very different from the first pair, then the order for it being chosen will be earlier or later.

iii.

The shape inner 1/3rd of the shape seem to be indistinguishable and even-sided glyph, moving outward, one can see concentrations of some irregular glyphs that are not even-sided and outermost graph shows some even-sided glyphs.

iv.

All pairs of axies seem to be positvely correlated, so a pentagon could possibly summarize the positions of the points in the five-dimensional space of the measurements.

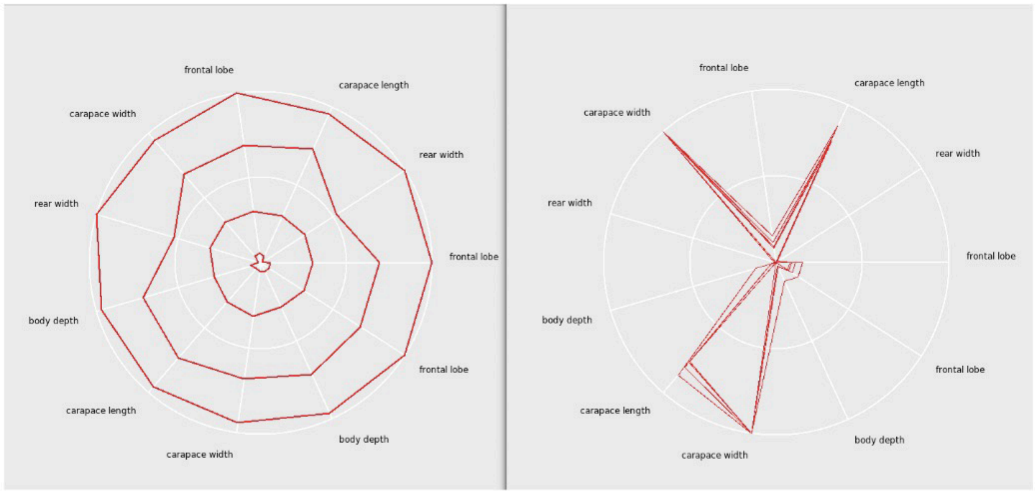
v.

The smaller the size, the harder it is to distinguish one from another. Size of body characteristics are not very obvious, hence the center one-third of the data doesn't have an irregular shape, apepars fairly even. Growing larger with larger body characteristics, their features are more distinguishable.

c. Construct a second radial axis display **sa2** identical to **sa1** and have them both appear side by side on your screen.

```
sa1<-l_serialaxes(data = lepto,linkingGroup = "lepto", sequence=seq, showItemLabels = TRUE)
sa2<-l_serialaxes(data = lepto,linkingGroup = "lepto", sequence=seq, showItemLabels = TRUE)
```

i.



ii.

The shape appear to be somewhat different in sa1. I.e, the inner most star and the third start from in to out seem to have more irregular shape while the second and largest star are fairly even sided. In sa2, the stars appear to have similar shape.

iii.

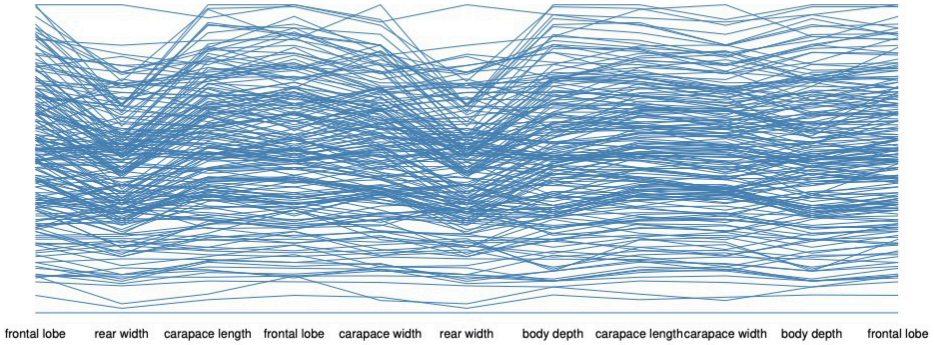
Carapace width is the dominant feature in sa2.

iv.

It can be used to see the shape of the crab and ignore their sizes. For example, for the crab, it was difficult to see dominant feature in sa1, but this is captured in sa2.

Visualizing Rock Crabs using Parallel Axis

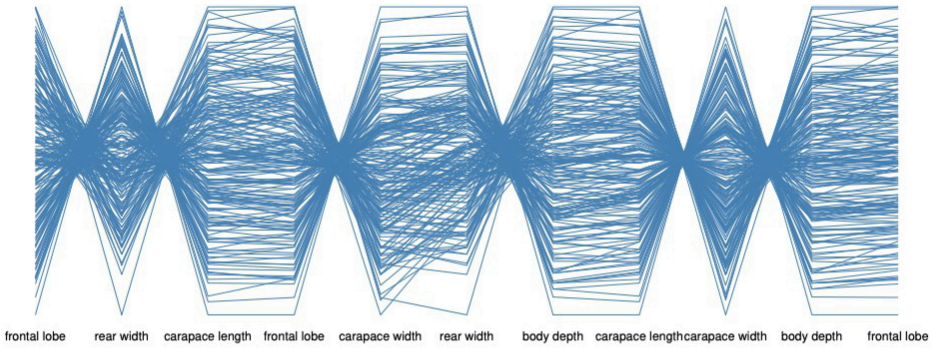
```
label = sub(" \\(mm\\)", "", colnames(lepto))
colnames(lepto) = label
ord <- eulerian(ncol(lepto))
seq <- names(lepto)[ord]
sa1<-l_serialaxes(data = lepto,linkingGroup = "lepto", sequence=seq, showItemLabels = TRUE,axesLayout = "p
l_export(sa1, filename = "sa1parallel.pdf", height = 400, width = 1000)
```



b. i.

```
flipAxes <- function(data) {
                        dataCopy <- data
                        for (i in 1:ncol(data)) {
                             if (i%%2 == 0) dataCopy[,i] <- (-1) * dataCopy[,i]
                             }
                        dataCopy}

data = flipAxes(lepto)
ord2 <- eulerian(ncol(data))
seq2 <- names(data)[ord2]
sa3<-l_serialaxes(data = data, sequence=seq2, linkingGroup = "lepto", showItemLabels = TRUE,axesLayout = "
l_export(sa3, filename = "sa3parallel.pdf", height = 400, width = 1000)
```



ii.

Between rear width and carapace length, there are 2 points of convergence that are not parallel. Between rear width and body depth there are also two points of convergence that are not paralle. Between carapace width and rear width one can see two groups are positively correlated within themselves but negatively correlated with each other. carapace length and carapace width shows criss-crossed lines with one point of convergence. There apears to be two groups between two parallel variables except between carapace length and width. This could indicate structure similar to a plane in 5d space.

iii.

The lines appear to criss-cross at two points of convergence, one is in the upper left while the other is in the lower right relatively speaking. The point of convergence aren't exactly directly above or beneath each other, so the two lines are probably not parallel to each other. The sign is negative as the criss cross shape show a negative correlation.

iv.

```
unique(sa3['color'])
group1 <- as.numeric(rownames(data[sa3['color']=="#46468282B4B4",]))
group2<- as.numeric(rownames(data[sa3['color']=="#3333A0A02C2C",]))
length(group1)
length(group2)
```

Group1 Rows:

[1] 1 2 4 5 6 7 8 9 10 11 12 13 15 17 18 20 22 23 25 27 31 37 38 40 41

[26] 42 43 44 45 48 49 50 52 53 54 56 57 60 62 64 66 67 70 75 77 79 82 85 86 88

[51] 93 96 97 98 100 101 102 105 108 109 110 111 112 113 114 115 116 117 120 126 128 129 132 133 135

[76] 137 138 144 145 147 152 153 156 157 158 161 162 163 165 167 169 172 173 175 177 178 180 181 182 183

[101] 186 189 192 195 197 198 199 200

Length of Group 1: 108

Group2 Rows:

[1] 3 14 16 19 21 24 26 28 29 30 32 33 34 35 36 39 46 47 51 55 58 59 61 63 65

[26] 68 69 71 72 73 74 76 78 80 81 83 84 87 89 90 91 92 94 95 99 103 104 106 107 118

[51] 119 121 122 123 124 125 127 130 131 134 136 139 140 141 142 143 146 148 149 150 151 154 155 159 160

[76] 164 166 168 170 171 174 176 179 184 185 187 188 190 191 193 194 196

Length of Group2: 92

v.

The lines look roughly parallel as the two points of convergence seem to align roughly on the same vertical line.

```
unique(sa3['color'])
subgroup1a<- as.numeric(rownames(data[sa3['color']=="#3333A0A02C2C",]))
subgroup1b<- as.numeric(rownames(data[sa3['color']=="#E3E31A1A1C1C",]))
length(subgroup1a)
length(subgroup1b)
```

subgroup1a rows:

[1] 3 14 16 21 24 26 28 29 30 32 33 34 35 36 39 55 58 59 61 65 68 71 72 74 76

[26] 78 80 81 83 84 89 90 91 92 94 95 99 103 106 118 119 121 122 123 125 130 136 139 141 142

[51] 143 150 151 154 159 160 166 168 170 184 188 191 193 194 196

subgroup1a length: 65

subgroup1b rows:

[1] 19 46 47 51 63 69 73 87 104 107 124 127 131 134 140 146 148 149 155 164 171 174 176 179 185

[26] 187 190

subgroup1b length: 27

vi.

The lines look roughly parallel as the two points of convergence seem to align roughly on the same vertical line.

```
unique(sa3['color'])
subgroup2a<- as.numeric(rownames(data[sa3['color']=="#6A6A3D3D9A9A",]))
subgroup2b<- as.numeric(rownames(data[sa3['color']=="#46468282B4B4",]))
length(subgroup2a)
length(subgroup2b)
```

subgroup2a rows:

[1] 1 2 5 6 8 10 11 12 15 18 20 23 25 27 37 41 43 44 48 49 53 56 60 62 66

[26] 67 70 79 82 85 86 88 96 101 105 110 112 114 115 116 117 126 129 132 133 135 144 145 153 156

[51] 162 169 177 178 183 186 192 195 197 198 200

subgroup2a length: 61

subgroup2b rows

[1] 4 7 9 13 17 22 31 38 40 42 45 50 52 54 57 64 75 77 93 97 98 100 102 108 109

[26] 111 113 120 128 137 138 147 152 157 158 161 163 165 167 172 173 175 180 181 182 189 199

subgroup2b length: 47

vii.

```
load("type.Rda")
summary(type[subgroup1a,])
```

```
##    Species       Sex
##  blue  :17   female:65
##  orange:48   male  : 0
```

```
summary(type[subgroup1b,])
```

```
##    Species       Sex
##  blue  :27   female:27
##  orange: 0   male  : 0
```

```
summary(type[subgroup2a,])
```

```
##    Species       Sex
##  blue  :49   female: 8
##  orange:12   male  :53
```

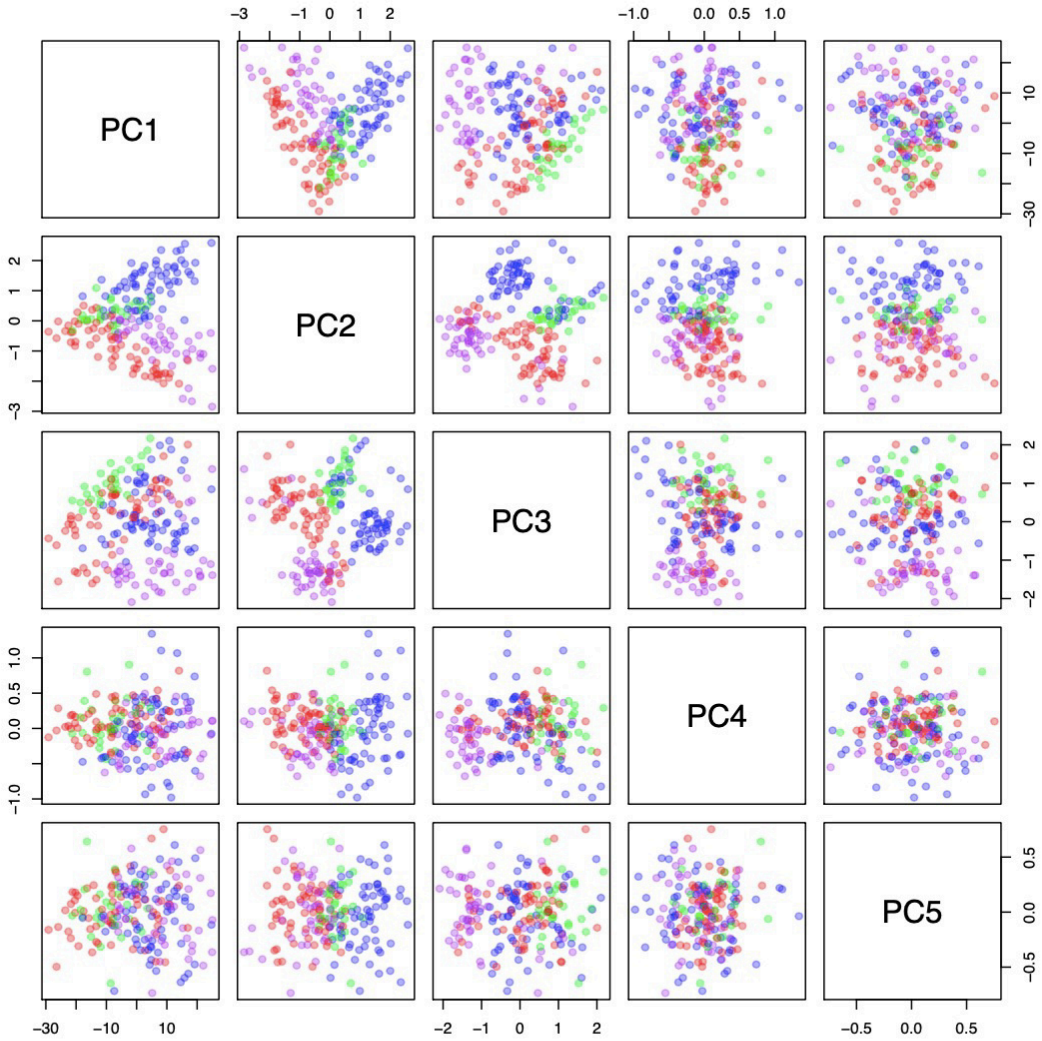
```
summary(type[subgroup2b,])
```

```
##    Species       Sex
##  blue  : 7   female: 0
##  orange:40   male  :47
```

The summary data appears to show that the first group are mainly females with subgroup1a orange females, subgroup1b blue females. The second group are mainly males, with subgroup2a mainly blue males and subgroup2b mainly orange males. The summary findings are consistent with the conjecture.

Rock Crab Visualization using PCA

```
res<-prcomp(lepto)
data <- data.frame(res$x)
data$color <- "blue"
data[subgroup1b,"color"] <- "green"
data[subgroup2a,"color"] <- "red"
data[subgroup2b,"color"] <- "purple"
pairs(data[,1:5], pch=19, col=adjustcolor(data$color, alpha=0.3 ))
```



ii.

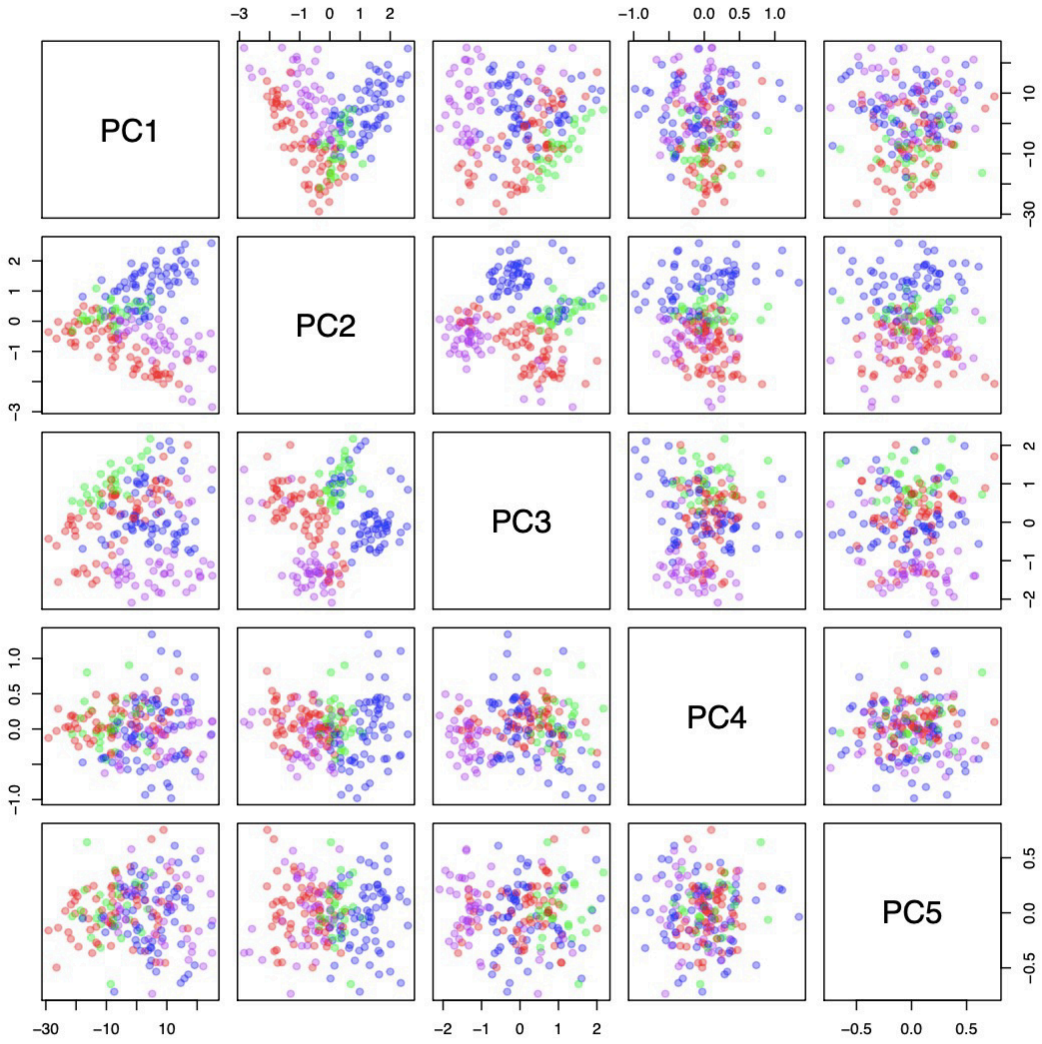
```
res<-prcomp(lepto)
data2 <- data.frame(res$x)

# orange female
data2$color <- "blue"

# blue female
bluefem <- as.numeric(rownames(type[(type$Species=="blue") & (type$Sex == "female"),]))
data2[bluefem,"color"] <- "green"
# blue male
bluem <- as.numeric(rownames(type[(type$Species=="blue") & (type$Sex == "male"),]))
data2[bluem,"color"] <- "red"
```

```
# orange male
orangem <- as.numeric(rownames(type[(type$Species=="orange") & (type$Sex == "male"),]))
data2[orangem,"color"] <- "purple"

pairs(data2[,1:5], pch=19, col=adjustcolor(data$color, alpha=0.3 ))
```



iii.

PC2 and PC3 seem to best separate the four groups.My grouping fared pretty well in separating the same groups on those two principals. My grouping looks similar to the one achieved by the second grouping based on species and sex.

g. Principal components continued.

```
library(loon); library(PairViz)
x <- prcomp(lepto)$x
```

```
nav <- l_navgraph(x[, 1:3])
p <- nav$plot
g <- l_glyph_add_serialaxes(p, data = lepto[, eseq(ncol(lepto))],
                            showAxes = TRUE, showArea = TRUE, label = "serialaxes")
p["glyph"] <- g
```

h.

```
km <- kmeans(lepto, centers = 4)
p["color"] <- km$cluster
```

```
km <- kmeans(lepto, centers = 4)
classes <- paste(type[,1], type[,2], sep = ":")
table(classes, km$cluster)
```

```
##
## classes          1  2  3  4
##   blue:female    3 16 19 12
##   blue:male     10 19 14  7
##   orange:female 18 21  9  2
##   orange:male   14 14 17  5
```

The clustering by "kmeans" doesn't seem very accurate. The scientist answer in "type" is not well identified based on the kmeans clustering method as shown in this above table.

ii.

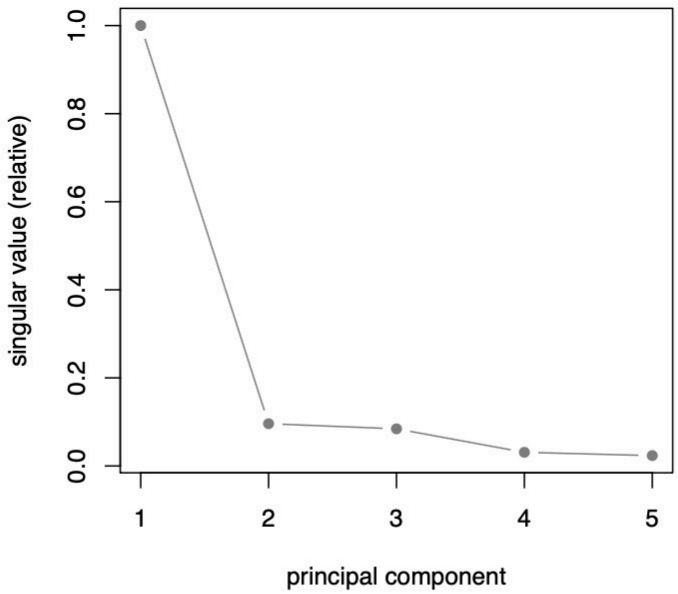
```
p["color"] <- "grey50"
table(classes, p["color"])
```

Using the navigation graph, I identified 4 groups in the PC2:PC3 plane. In the PC2:PC3 plane, it was easy to see clusters of data grouped around 4 centers, which I used to identify the 4 different groups with the idea that similar data are grouped together. Using this method, I was able to more accurately identify the 4 classes as shown in the table below. The four groups I have colored in the graph closely relate to one of the four classes identified in the scientist answer.

classes	#1F1F7878B4B4	#3333A0A02C2C	#7F7F7F7F7F7F	#E3E31A1A1C1C
blue:female	1	1	47	1
blue:male	0	44	3	3
orange:female	47	0	0	3
orange:male	0	0	0	50

iii.

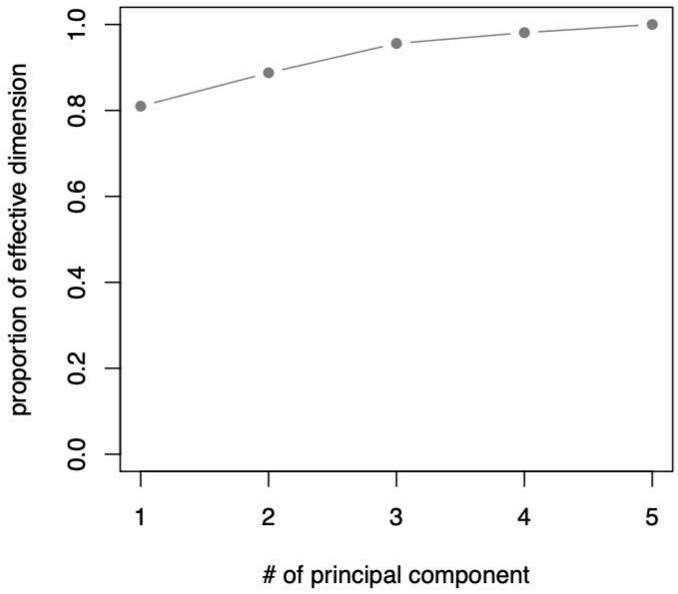
```
sdev <- prcomp(lepto)$sdev
plot(sdev/max(sdev), xlab="principal component",
ylab="singular value (relative)", col="grey50", pch=16,type="b")
```



According to the scree plot, one principal component would be efficient, as there is a huge drop from the first component to the second component. Component 1 has much bigger singular value relative to components 2,3,4,5.

Looking at the following plot that measures the effective (fractional) dimensionality of the data, we see that around 80% of the dimensionality is retained by 1 principal component, which means that 1 component is probably close enough to identify the groups.

```
plot(cumsum(sdev)/sum(sdev), xlab="# of principal component", ylab="proportion of effective dimension",
     ylim=0:1, col="grey50", pch=16,type="b")
```



But the scree plot is not enough, because need PC2 and PC3 to separate the data.