Practice2

Dasha Asienga

Due by midnight, Friday, Sept. 30

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook(s), course materials in Moodle, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

Prompt

The data set primatedata.txt contains variables about shoulder blades from 105 primates. There are 5 indices (basically measured distances): AD.BD, AD.CD, EA.CD, Dx.CD, and SH.ACR, and 2 angles recorded (EAD and beta) for each shoulder blade. Unfortunately, I do not have a schematic showing what each index or angle measures. The primate genus (plural: genera) is also recorded in the class variable. Gibbons (Hylobates), orangutangs (Pongo), chimpanzees (Pan), gorillas (Gorilla), and man (Homo) are all represented in the data set.

A research scientist is requesting that you assist them and use this technique they've heard about but don't understand, called PCA, to reduce the dimension of this data set from the 7 measurement variables down to a reasonable dimension to view, in order to look for differences between the primate genus groups. Is it possible to find differences in the PC space? If so, what are they? Your response should include the motivation for why PCA is appropriate for the task, briefly what it does (i.e. explain the method to the researcher), and justification of any choices you have to make in the process of performing it. Additionally, due to the visualization request, you should have an appropriate graphic(s) where you can show differences between the primate groups, if they exist.

Introduction

Purpose of the Analysis

As per the researcher's request, the purpose of this analysis is to understand if there are any differences between the 5 different primate groups in the data set `primatedata`. The 5 primate groups are gibbons (Hylobates), orangutangs (Pongo), chimpanzees (Pan), gorillas (Gorilla), and man (Homo). To explore any differences, we will consider 7 different quantitative variables in the data set. There are 5 indices (measured distances): AD.BD, AD.CD, EA.CD, Dx.CD, and SH.ACR, – and 2 angles recorded for each shoulder blade: EAD and beta. Using these variables, we aim to show differences between the primate groups, if they exist. However, because there are 7 different variables, we will employ Principal Components Analysis (PCA) to reduce the dimensionality of the data and to allow us to be able to visualize and interpret any significant differences.

What is Principal Components Analysis (PCA)?

PCA is a dimension reduction technique designed to help re-express multivariate data in lower dimensions. With 7 quantitative variables in our data set, `primatedata` lives in 7 dimensions. This is difficult to interpret and visualize. Instead, we can employ PCA to project the data into lower dimensions while capturing maximum variability in the data. This is a useful technique if the original variables are somewhat correlated. PCA then creates linear combinations of the original variables known as principal components (PCs). The new variables – the principal components – are uncorrelated, and the first few account for as much variation in the data as possible. Ultimately, this technique allows us to visualize and analyze multivariate data in lower dimensions that are easier to interpret.

Preliminary Analysis

```
primatedata <-
  read.table("https://awagaman.people.amherst.edu/stat240/primatedata.txt",
             header = TRUE)
```

Exploring the Data Set and Subsetting

As seen below, all the variables are numeric with the exception of `class`. Hence, PCA is feasible, but we first need to determine if there are any correlations between the variables before stating that it is appropriate. We can also begin to see that the variables are on different scales, and so the correlation matrix will be more appropriate compared to the covariance matrix.

```
glimpse(primatedata)
```

```
## Rows: 105
## Columns: 8
## $ AD.BD  <dbl> 65.56, 50.91, 46.15, 70.29, 63.16, 50.72, 58.99, 55.38, 64.29, ~
## $ AD.CD  <dbl> 166.0, 93.9, 80.8, 220.5, 144.0, 134.6, 164.0, 144.0, 138.5, 16~
## $ EA.CD  <dbl> 50.55, 61.82, 64.10, 50.00, 57.89, 56.23, 54.96, 52.31, 57.14, ~
## $ Dx.CD  <dbl> 12.80, 13.09, 11.80, 12.75, 12.98, 11.88, 12.46, 11.92, 12.50, ~
## $ SH.ACR <dbl> 70.3, 75.0, 70.0, 61.1, 64.9, 52.6, 58.6, 65.3, 60.0, 55.3, 64.~
## $ EAD    <int> 115, 121, 120, 113, 115, 136, 109, 131, 115, 117, 121, 124, 119~
## $ beta   <int> 14, 20, 25, 12, 14, 14, 11, 16, 16, 20, 16, 12, 17, 19, 11, 11,~
## $ class  <chr> "Hylobates", "Hylobates", "Hylobates", "Hylobates", "Hylobates"~
```

Because we can only use numeric variables for PCA, we will subset the data set to exclude `class`.

```
primatedata2 <- select(primatedata, -class)
```

Is PCA Appropriate?

As seen in the correlation matrix below, there are some moderate to strong correlations between the variables, suggesting that PCA is appropriate. We will proceed with PCA as our primary technique for this analysis.

```
cor(primatedata2)
```

```
##            AD.BD      AD.CD      EA.CD      Dx.CD    SH.ACR        EAD
## AD.BD   1.000000  0.7100438 -0.7986565  0.3217468 -0.282293 -0.6212922
## AD.CD   0.710044  1.0000000 -0.8142364 -0.0838734 -0.143759 -0.1556655
## EA.CD  -0.798656 -0.8142364  1.0000000 -0.0187430  0.193423  0.0993157
## Dx.CD   0.321747 -0.0838734 -0.0187430  1.0000000 -0.174719 -0.4697561
## SH.ACR -0.282293 -0.1437595  0.1934230 -0.1747188  1.000000  0.1240774
## EAD    -0.621292 -0.1556655  0.0993157 -0.4697561  0.124077  1.0000000
## beta   -0.640000 -0.8613548  0.7410626  0.1328385  0.347479  0.0195885
##             beta
## AD.BD  -0.6400004
## AD.CD  -0.8613548
## EA.CD   0.7410626
## Dx.CD   0.1328385
## SH.ACR  0.3474792
## EAD     0.0195885
## beta    1.0000000
```

Correlation vs Covariance Matrix

The covariance matrix below reveals that some of the variables have larger variances than others. Because PCA is not scale invariant, then variables with large variances would dominate. As such, the correlation matrix will be more appropriate because it is scaled.

```
cov(primatedata2)
```

```
##              AD.BD      AD.CD       EA.CD     Dx.CD   SH.ACR        EAD
## AD.BD   263.26339  484.60719 -142.727866  3.350807 -39.2785 -107.26809
## AD.CD   484.60719 1769.37129 -377.236344 -2.264507 -51.8569  -69.67567
## EA.CD  -142.72787 -377.23634  121.312911 -0.132505  18.2693   11.63994
## Dx.CD     3.35081   -2.26451   -0.132505  0.411984  -0.9617   -3.20842
## SH.ACR  -39.27850  -51.85685   18.269297 -0.961700  73.5394   11.32223
## EAD    -107.26809  -69.67567   11.639940 -3.208425  11.3222  113.22930
## beta   -127.99658 -446.59558  100.607712  1.050962  36.7292    2.56923
##              beta
## AD.BD  -127.99658
## AD.CD  -446.59558
## EA.CD   100.60771
## Dx.CD     1.05096
## SH.ACR   36.72923
## EAD       2.56923
## beta    151.93077
```

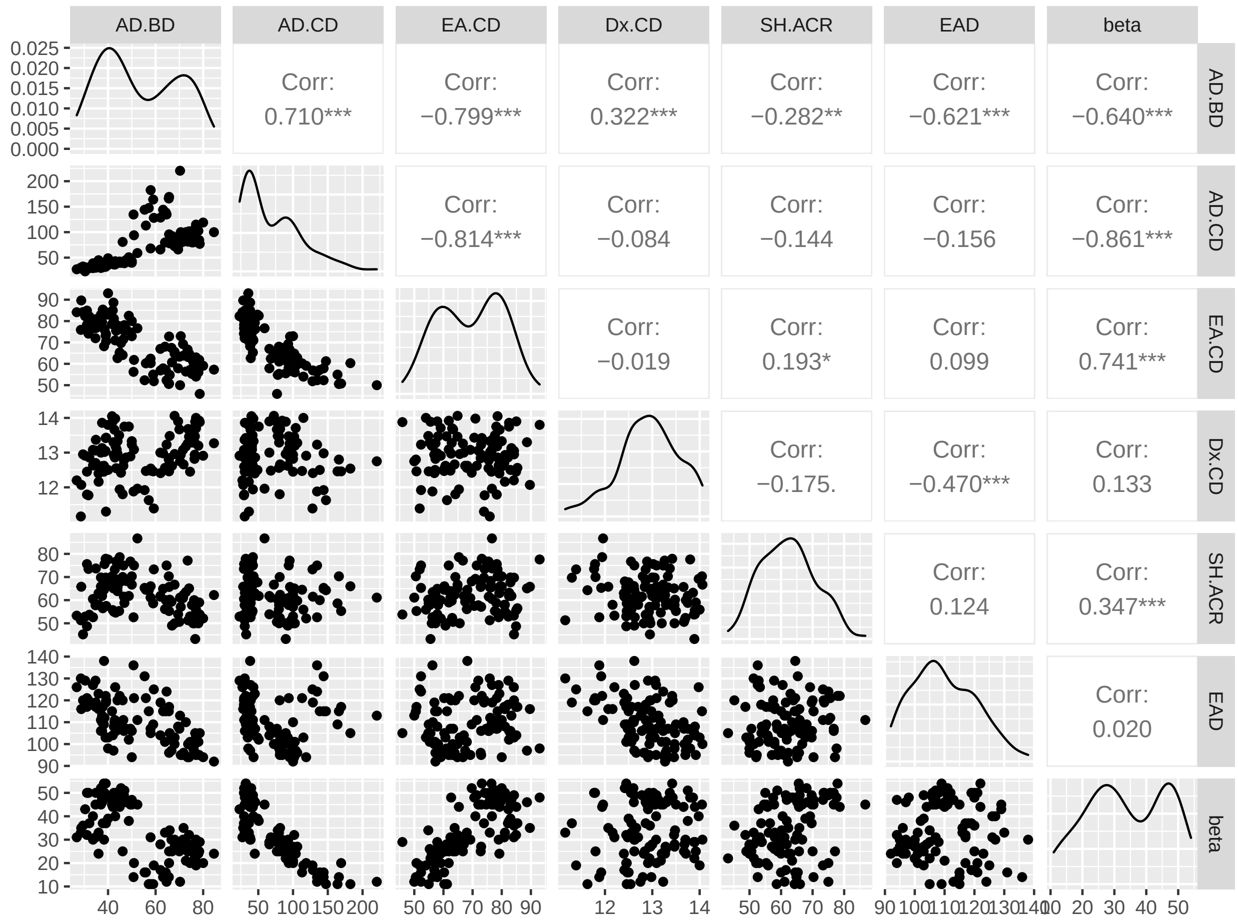
Univariate and Bivariate Analysis

The matrix below allows us to see much more clearly that there are evident relationships between some of the variables in our data set, with some correlations being very strong. In particular, `AD.BD` seems to have a relationship with all the variables in the data set, and `beta` seems to have a strong relationship with a number of the variables. This is important because it suggests that we can reduce the dimensionality of this data set and PCA could prove especially useful.

Some of our variables are not normally distributed, but that is not a major concern because PCA is an exploratory technique.

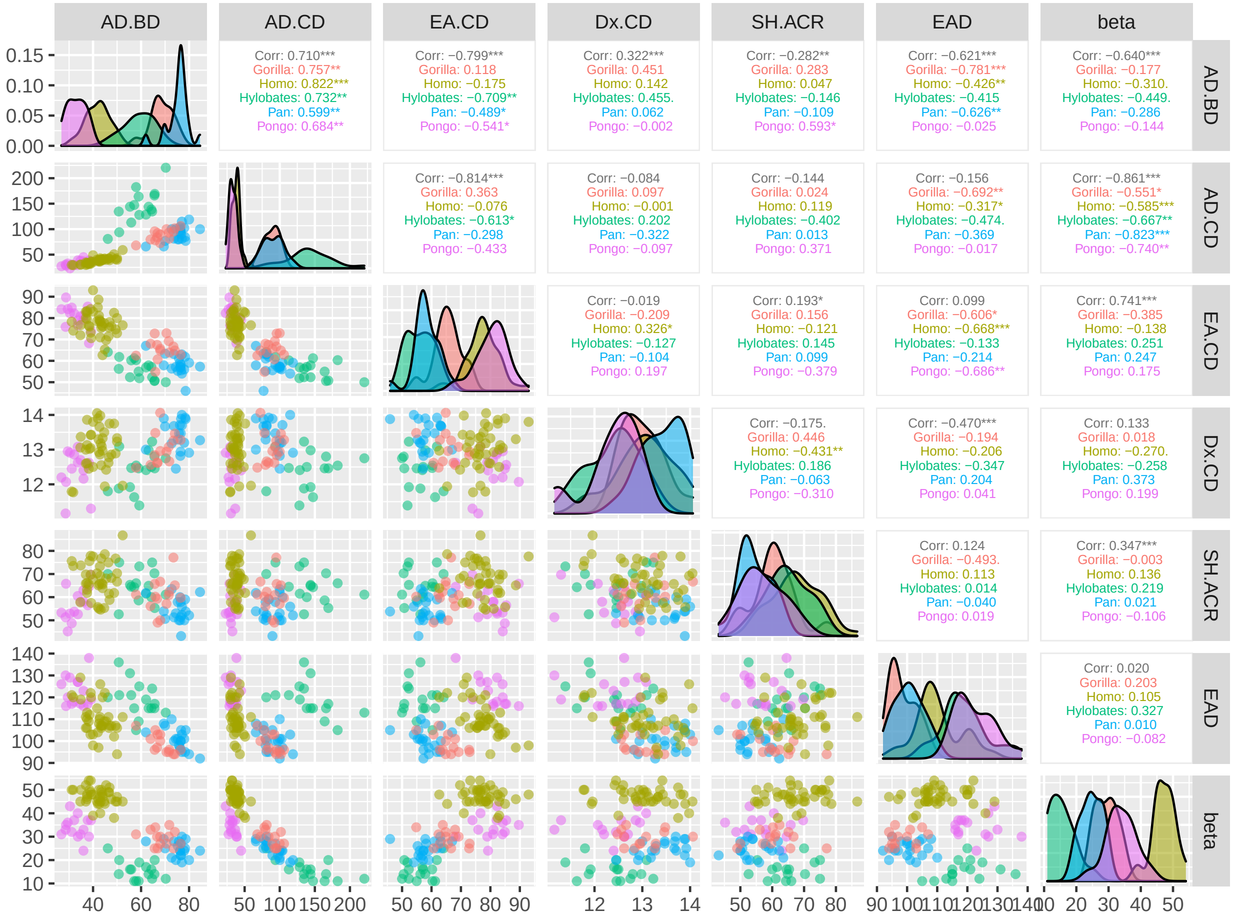
Finally, it appears that there a few outliers in the data set, but they don't seem to be very influential. We will proceed with the data set as it is.

```
ggpairs(primatedata2)
```



Multivariate Analysis and Motivation for PCA

```
ggpairs(primatedata, columns = c(1:7),
        mapping = aes(color = class, alpha = 0.7),
        upper = list(continuous = wrap("cor", size = 2)))
```



As seen in the matrix above, different primate groups have different values for the 7 quantitative variables in our data set. However, it is difficult to explain that difference because we have many variables. Given that `class` is clearly of importance, we want to explain that difference in a simple way while accounting for all the variables in our data set. Let's proceed to perform PCA on our data set.

Methods

Principal Components Analysis (PCA)

As explained in the introduction, PCA is a dimension reduction technique designed to help re-express multivariate data in lower dimensions. Often, we want to visualize multivariate data, but that is not possible when our data set lives in a high dimension. With 7 quantitative variables in our data set, `primatedata` lives in 7 dimensions. This is difficult to interpret and visualize.

Instead, we will employ PCA to project the data into lower dimensions while capturing maximum variability in the data. As seen in our preliminary analysis, our variables have moderate to strong correlations, making PCA suitable. We also saw that some variables have a large variance, and so because PCA is not scale invariant, we will perform the PCA on a correlation matrix.

The PCA technique will create linear combinations of the original variables known as principal components (PCs). The new variables – the principal components – are uncorrelated and ordered, meaning that the first few account for most of the variation in our original data set.

Choosing the Number of PCs

There are many informal methods to choose how many PCs to proceed with. We hope to explain as much of the variation in our data set as possible. Therefore, for the purpose of this analysis, we will pre-specify that we aim to account for at least 80% of the original variation in order to make sure that our PCs are representative. However, we will also look into some of the other informal methods as well in order to come to a consensus on the number of PCs to proceed with.

Visualizing and Interpreting the Results

Ultimately, PCA allows us to visualize and analyze multivariate data in lower dimensions that are easier to interpret. After choosing the number of PCs to proceed with, we will plot them in their respective PC space using biplot visualizations. These will allow us to visually assess the loadings or weights of the different variables in our data set for each of the PCs we choose to proceed with.

Finally, we will examine if there are any visible clusters and whether `class` (the primary genus of the primate) can explain that difference, hence answering the researcher's question.

Results

Performing PCA

Based on the PCA output below, PC_1 accounts for 49.76% of the variance, PC_2 accounts for 23.91% of the variance, and PC_3 accounts for 13.36% of the variance. We will choose how many PCs to retain in the next section, but the first 3 PCs explain 87% of our original variance.

PC_1 and PC_3 have all the variables in our data set with the exception of `Dx.CD`, which carries the largest weight in PC_2 alongside `EAD`. `SH.ACR` has the largest weight for PC_3. It is difficult to interpret the loadings, but this will become easier when we visualize our chosen PCs.

```
myPCA1 <- princomp(primatedata2, cor = TRUE, scores = TRUE)
summary(myPCA1)
```

```
## Importance of components:
##                          Comp.1   Comp.2   Comp.3    Comp.4    Comp.5    Comp.6
## Standard deviation     1.866323 1.293709 0.966999 0.7230887 0.5089244 0.3216263
## Proportion of Variance 0.497594 0.239097 0.133584 0.0746939 0.0370006 0.0147776
## Cumulative Proportion  0.497594 0.736692 0.870275 0.9449693 0.9819699 0.9967476
##                            Comp.7
## Standard deviation     0.15088786
## Proportion of Variance 0.00325245
## Cumulative Proportion  1.00000000
```

```
myPCA1$loadings
```

```
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## AD.BD   0.493  0.229  0.146         0.290  0.297  0.710
## AD.CD   0.479 -0.222  0.163        -0.390 -0.717  0.166
## EA.CD  -0.477  0.160 -0.115  0.282 -0.631         0.502
## Dx.CD          0.643        -0.715 -0.250
## SH.ACR -0.201 -0.136  0.928 -0.176 -0.123  0.186
## EAD    -0.201 -0.602 -0.239 -0.609                0.398
## beta   -0.461  0.281  0.135         0.528 -0.597  0.235
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```

Choosing the Number of PCs to Retain

There are no formal ways of determining how many PCs to retain. We will examine a number of informal methods before coming to our final conclusion how many PCs to ultimately proceed with. The most important criteria is to explain as much of the variation in our original variables as possible. Our pre-specified goal is to explain at least 80% of the variation in the data, but let us examine other methods as well to determine whether they agree with our final conclusion.

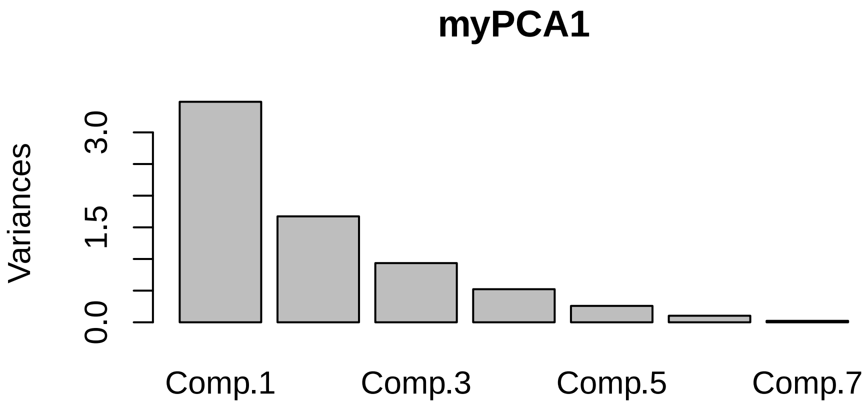
Explaining 70% or 80% of the variation: In order to explain 80% of the original variance, as pre-specified above, we would need to keep the first 3 PCs.

Kaiser's Rule: Kaiser's Rule states that we should retain PCs with a variance($\lambda_i \geq 1$), that is, $sd \geq 1$. Following this rule, we we would only keep the first 2 PCs.

Jolliffe's Rule: On the other hand, Jolliffe's Rule states that we should retain PCs with a variance($\lambda_i \geq 0.70$), that is, $sd \geq 0.8366$. Following this rule, we would keep the first 3 PCs.

Scree Plot: Lastly, let us examine a scree plot. Based on the scree plot below, I would keep the first 3 PCs.

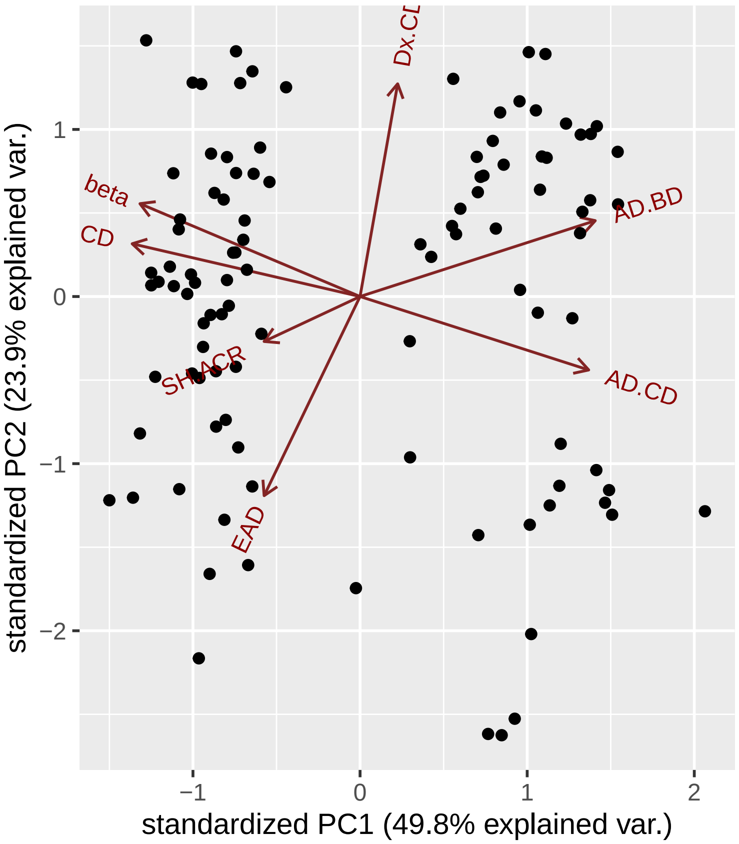
```
plot(myPCA1)
```



Based on examining all the informal methods and in an aim to maximize variance explained, we will proceed with the first 3 PCs, which will explain 87% of the original variance in our data set.

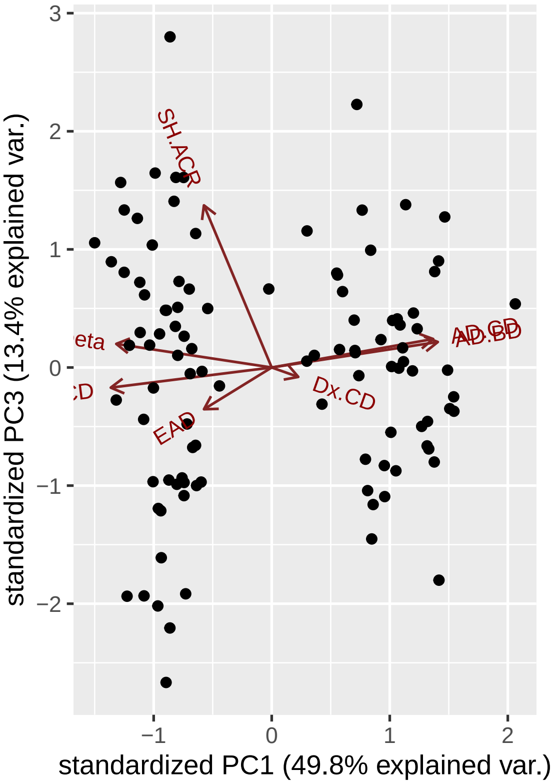
Visualizing and Interpreting the Results

```
ggbiplot::ggbiplot(myPCA1)
```



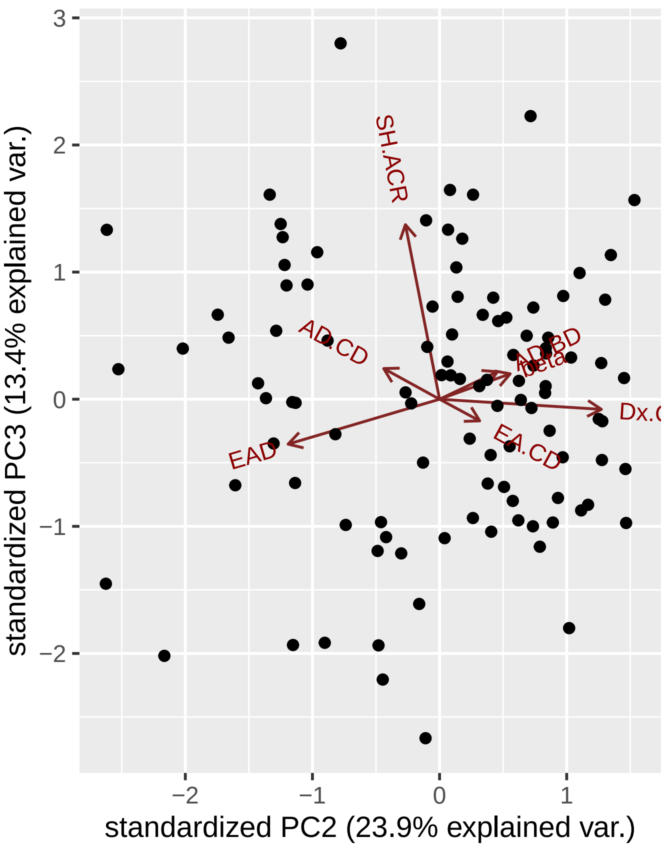
In looking at the first biplot of PC_1 and PC_2, there appears to be some visible clusters: one on the left, one on the top right, and one on the bottom right. In the next section, we will attempt to classify these clusters more formally. We see that the two clusters on the right are explained by higher values of `AD.BD` and `AD.CD` respectively, which will be quite interesting to investigate.

```
ggbiplot::ggbiplot(myPCA1, choices = c(1,3))
```



In looking at the biplot of PC_1 and PC_3, we also see some visible clusters: one on the left and one on the right. In the next section, we will attempt to classify these clusters more formally. However, it seems like the cluster on the right is a combination of the 2 clusters we saw on the right in the first visual, largely explained by higher values of both `AD.CD` and `AD.BD`. In both plots, we see one score on the very far right.

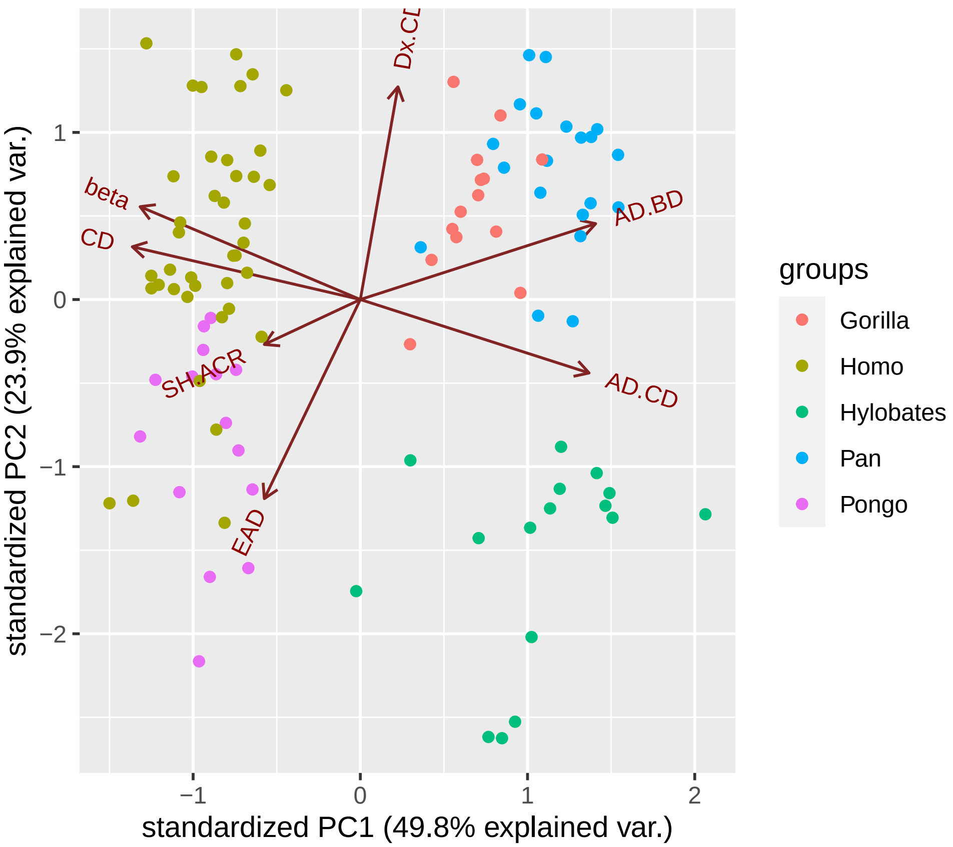
```
ggbiplot::ggbiplot(myPCA1, choices = c(2,3))
```



Finally, this last biplot of PC_2 and PC_3 does not reveal any visible clusters, but it seems that most of the scores are concentrated on the right of the plot. It's hard to interpret and there are no discernible clusters. We see, however, an exceptionally high value recorded for the distance `SH.ACR` at the very top of the plot. There are also some scores on the edges of the plot.

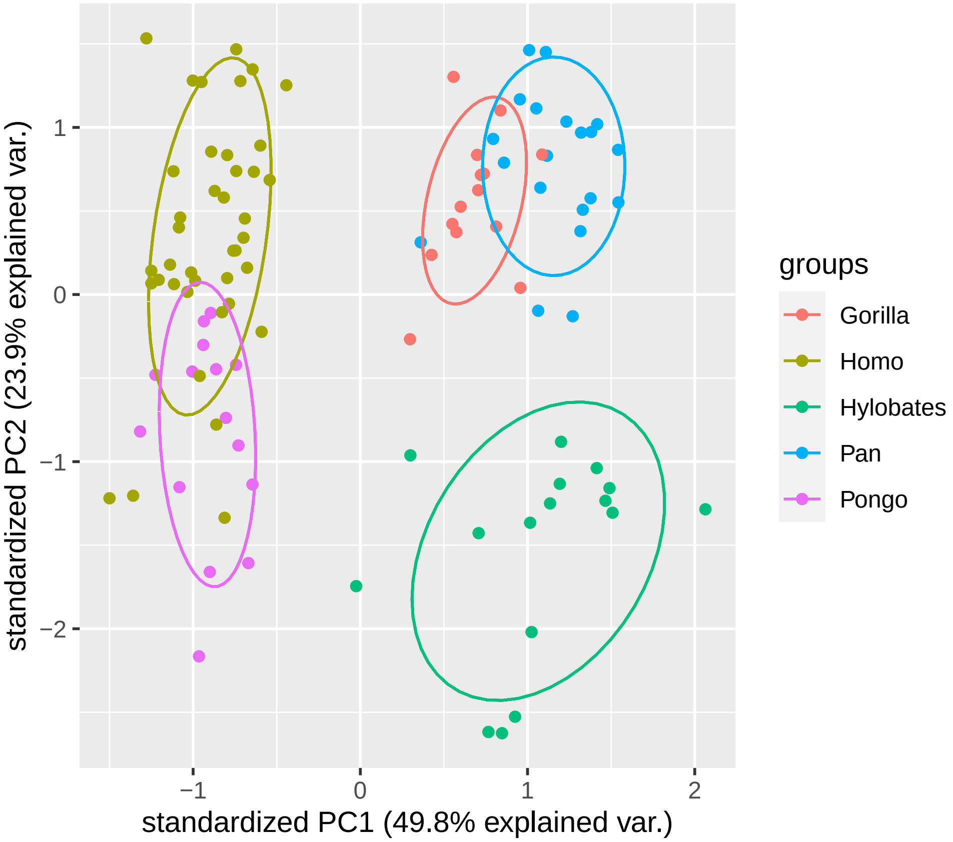
Investigating Difference Across the Different Primate Groups

```
ggbiplot::ggbiplot(myPCA1, groups = primatedata[,"class"])
```



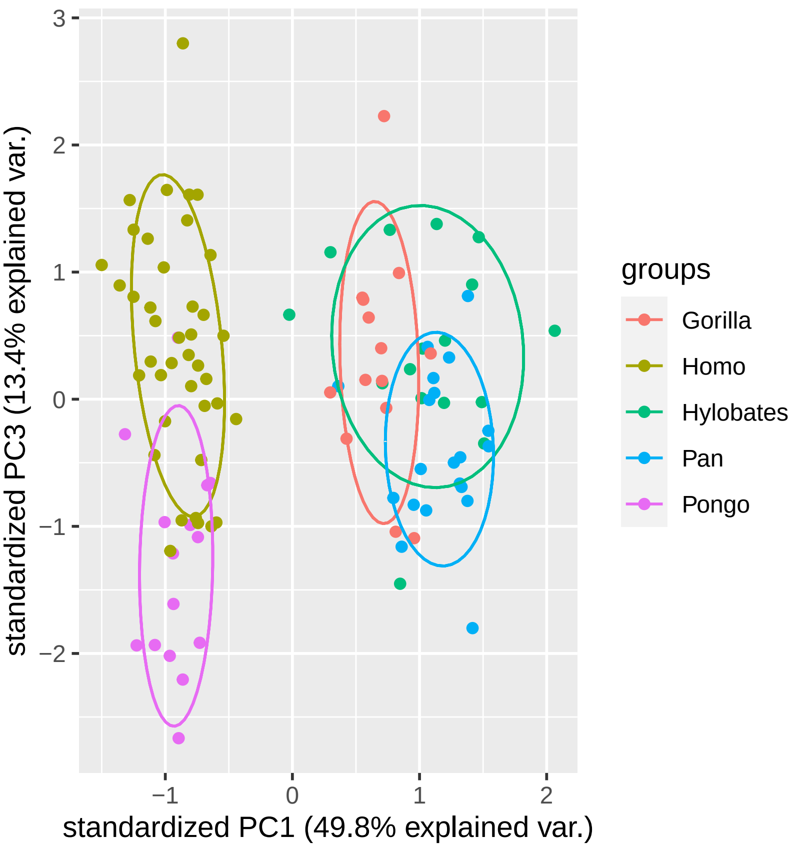
When we look at the biplot above of PC_1 and PC_2, we indeed see some clusters based on the primary genus group of the primates. The loadings help us visualize key variables distinguishing these groups. For example, Hylobates have really high positive weight for `AD.CD` and a high negative weight for the angle `EAD`.

```
ggbiplot::ggbiplot(myPCA1, var.axes = FALSE , ellipse = TRUE,
                   groups = primatedata[,"class"])
```



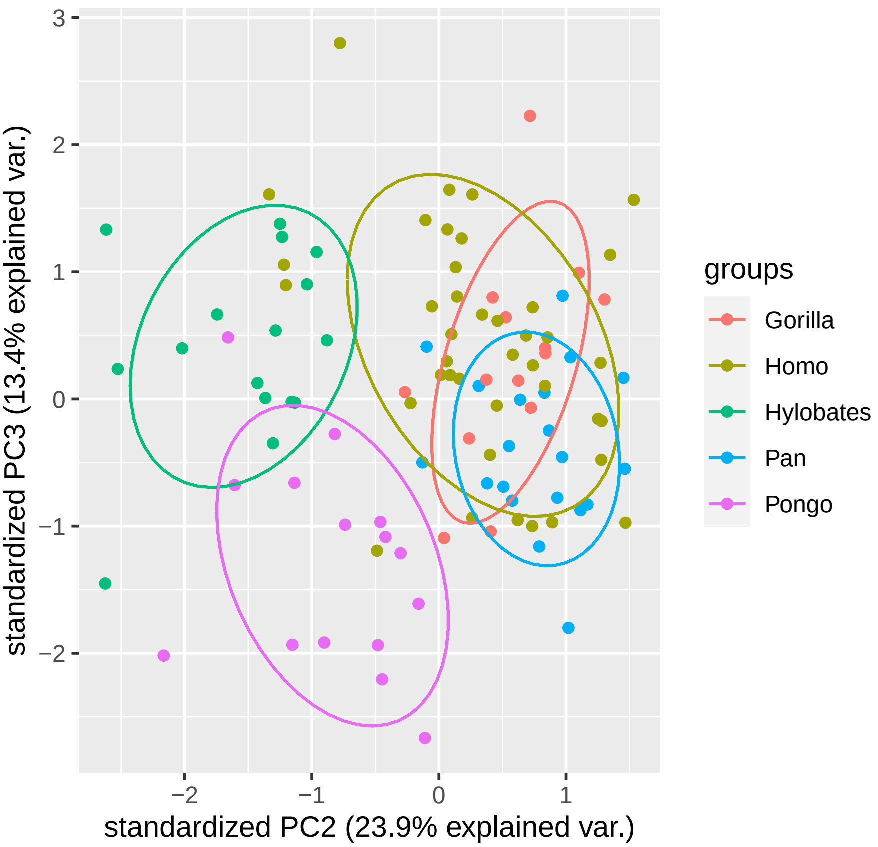
Still looking at PC_1 and PC_2, we can see that Hylobates have a distinct defining characteristic, and there is some overlap between Gorilla and Pan, as well as between Homo and Pongo.

```
ggbiplot::ggbiplot(myPCA1, choices = c(1,3), var.axes = FALSE , ellipse = TRUE,
                   groups = primatedata[,"class"])
```



The PC_1 and PC_3 space reveals 2 key groups of primates. There is a lot of overlap between Gorilla, Hylobates, and Pan, but these 3 are very different from the observed overlap between Homo and Pongo. This PC space reveals that there is a key difference in shoulder blades between 2 key groups of primates, the first group consisting of Gorilla, Hylobates, and Pan, which seem to have a lot of similarities, and the second group consisting of Homo and Pongo, with similarities as well.

```
ggbiplot::ggbiplot(myPCA1, choices = c(2,3), var.axes = FALSE , ellipse = TRUE,
                   groups = primatedata[,"class"])
```



Finally, the PC_2 and PC_3 space does not reveal much, as expected from the previous visualization.

Conclusion

The aim of this analysis was to answer the researcher's question: understanding if there are any differences between the 5 different primate groups in the data set `primatedata`. The 5 primate groups are gibbons (Hylobates), orangutangs (Pongo), chimpanzees (Pan), gorillas (Gorilla), and man (Homo). Because our data lives in 7 dimensions, we employed PCA to reduce the dimensionality for easier visualization and interpretation.

We ran PCA on a correlation matrix and using a number of informal methods, we decided to proceed with 3 PCs, which accounted for 87% of the variation in the original data set. We then visualized our data in the PC space and examined for any clusters across the primate groups.

Our PC_1 and PC_2 space revealed that there were indeed some visible differences across the primate groups. Hylobates was distinctly different, and there was some overalap between Gorilla and Pan, and between Homo and Pongo. When we looked at PC_1 and PC_3, these clusters broke down into 2 major groups. The group on the left was a mix of Homo and Pongo, while the group on the right was a mix of Gorilla, Hylobates, and Pan. This revealed that perhaps there were some strong similarities within the groups but key differences across them. Our PC_2 and PC_3 space did not reveal much.

Ultimately, we were able to answer the researcher's question and examine for any differences across the primate groups. We found that there are indeed distinct differences across all 5 groups. However, we also had a key finding as we discovered that some primate groups are more similar than others, and this can motivate further exploratory and/ or inferential analysis.