Homework 5

Manav Mandhani

This homework is due on Mar 3, 2015 in class.

In 1898, Hermon Bumpus, an American biologist working at Brown University, collected data on one of the first examples of natural selection directly observed in nature. Immediately following a bad winter storm, he collected 136 English house sparrows, *Passer domesticus*, and brought them indoors. Of these birds, 64 had died during the storm, but 72 recovered and survived. By comparing measurements of physical traits, Bumpus demonstrated physical differences between the dead and living birds. He interpreted this finding as evidence for natural selection as a result of this storm:

```
bumpus <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/bumpus_full.csv")
head(bumpus)
```

```
##    Sex   Age Survival Length Wingspread Weight Skull_Length Humerus_Length
## 1 Male Adult    Alive    154        241   24.5         31.2           17.4
## 2 Male Adult    Alive    160        252   26.9         30.8           18.7
## 3 Male Adult    Alive    155        243   26.9         30.6           18.6
## 4 Male Adult    Alive    154        245   24.3         31.7           18.8
## 5 Male Adult    Alive    156        247   24.1         31.5           18.2
## 6 Male Adult    Alive    161        253   26.5         31.8           19.8
##   Femur_Length Tarsus_Length Sternum_Length Skull_Width
## 1         17.0          26.0           21.1        14.9
## 2         18.0          30.0           21.4        15.3
## 3         17.9          29.2           21.5        15.3
## 4         17.5          29.1           21.3        14.8
## 5         17.9          28.7           20.9        14.6
## 6         18.9          29.1           22.7        15.4
```

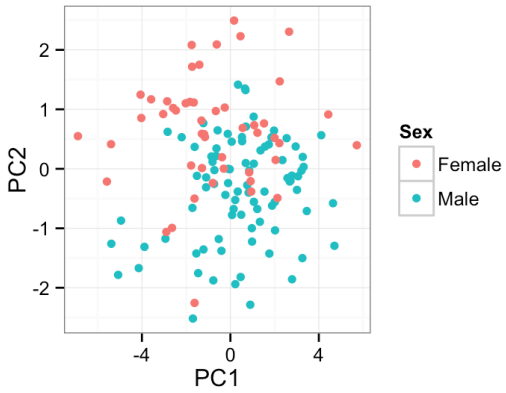
The data set has three categorical variables (`Sex` , with levels `Male` and `Female` , `Age` , with levels `Adult` and `Young` , and `Survival` , with levels `Alive` and `Dead`) and nine numerical variables that hold various aspects of the birds' anatomy, such as wingspread, weight, etc.

Question 1 (3 pts): Perform a PCA on the numerical columns of this data set. Then make three plots potting the data as PC2 vs. PC1, colored by (i) sex, (ii) age, (iii) survival.

```
bumpus %>% select(-Sex, -Age, -Survival) %>% scale() %>% prcomp() -> pca
pca_data <- data.frame(pca$x, Sex = bumpus$Sex)
head(pca_data)
```

```
##          PC1        PC2        PC3        PC4        PC5        PC6
## 1 -3.8856548 -1.3120204 -0.5218476  1.1129461 -0.1905058 -0.2336805
## 2  0.8224474 -0.3998802  0.7775967 -0.8865862 -0.2099604  1.2483988
## 3 -0.6426387  0.2143840  0.1084914  0.1627538 -0.6951568  1.9420813
## 4 -1.2227485  0.7674720  1.0376675  0.8433835 -0.5152572 -0.2262541
## 5 -1.6146988  0.3682892  1.3968497  0.2316305 -0.3493214 -0.6703105
## 6  2.7783623 -0.1149411  1.0431605  0.4265468  0.3594881  0.5671766
##           PC7        PC8         PC9  Sex
## 1  0.96374801  0.1525549 -0.57521792 Male
## 2  0.02392220 -1.0953482 -0.04083636 Male
## 3 -0.01924521 -0.3374622  0.10601678 Male
## 4  0.88126091 -0.7067789  0.74563716 Male
## 5  0.56087560 -0.5365881 -0.44692206 Male
## 6  0.68771676  0.7389925  0.20013441 Male
```

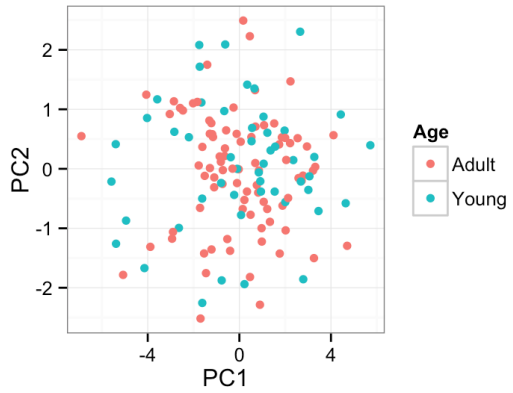
```
ggplot(pca_data, aes(x=PC1, y=PC2, color=Sex)) + geom_point()
```




```
pca_data <- data.frame(pca$x, Age = bumpus$Age)
head(pca_data)
```

```
##          PC1         PC2         PC3         PC4         PC5         PC6
## 1 -3.8856548 -1.3120204 -0.5218476  1.1129461 -0.1905058 -0.2336805
## 2  0.8224474 -0.3998802  0.7775967 -0.8865862 -0.2099604  1.2483988
## 3 -0.6426387  0.2143840  0.1084914  0.1627538 -0.6951568  1.9420813
## 4 -1.2227485  0.7674720  1.0376675  0.8433835 -0.5152572 -0.2262541
## 5 -1.6146988  0.3682892  1.3968497  0.2316305 -0.3493214 -0.6703105
## 6  2.7783623 -0.1149411  1.0431605  0.4265468  0.3594881  0.5671766
##           PC7         PC8          PC9    Age
## 1  0.96374801  0.1525549 -0.57521792 Adult
## 2  0.02392220 -1.0953482 -0.04083636 Adult
## 3 -0.01924521 -0.3374622  0.10601678 Adult
## 4  0.88126091 -0.7067789  0.74563716 Adult
## 5  0.56087560 -0.5365881 -0.44692206 Adult
## 6  0.68771676  0.7389925  0.20013441 Adult
```

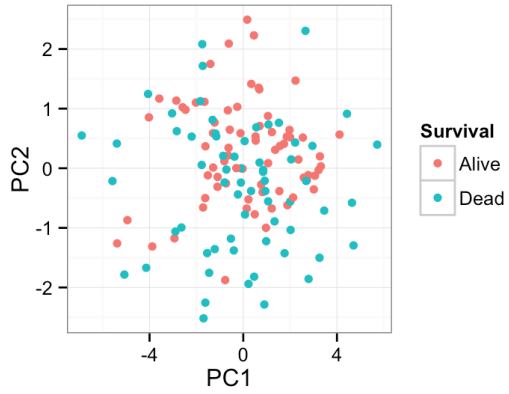
```
ggplot(pca_data, aes(x=PC1, y=PC2, color=Age)) + geom_point()
```




```
pca_data <- data.frame(pca$x, Survival = bumpus$Survival)
head(pca_data)
```

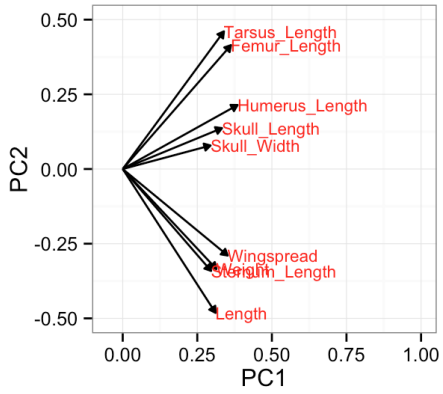
```
##          PC1        PC2        PC3        PC4        PC5        PC6
## 1 -3.8856548 -1.3120204 -0.5218476  1.1129461 -0.1905058 -0.2336805
## 2  0.8224474 -0.3998802  0.7775967 -0.8865862 -0.2099604  1.2483988
## 3 -0.6426387  0.2143840  0.1084914  0.1627538 -0.6951568  1.9420813
## 4 -1.2227485  0.7674720  1.0376675  0.8433835 -0.5152572 -0.2262541
## 5 -1.6146988  0.3682892  1.3968497  0.2316305 -0.3493214 -0.6703105
## 6  2.7783623 -0.1149411  1.0431605  0.4265468  0.3594881  0.5671766
##           PC7        PC8         PC9 Survival
## 1  0.96374801  0.1525549 -0.57521792    Alive
## 2  0.02392220 -1.0953482 -0.04083636    Alive
## 3 -0.01924521 -0.3374622  0.10601678    Alive
## 4  0.88126091 -0.7067789  0.74563716    Alive
## 5  0.56087560 -0.5365881 -0.44692206    Alive
## 6  0.68771676  0.7389925  0.20013441    Alive
```

```
ggplot(pca_data, aes(x=PC1, y=PC2, color=Survival)) + geom_point()
```



Question 2 (1 pt): Now visualize the rotation matrix of the PCA obtained under Question 1.

```
rotation_data <- data.frame(pca$rotation, variable=row.names(pca$rotation))
# define a pleasing arrow style
arrow_style <- arrow(length = unit(0.05, "inches"),
                     type = "closed")
# now plot, using geom_segment() for arrows and geom_text for labels
ggplot(rotation_data) + geom_segment(aes(xend=PC1, yend=PC2), x=0, y=0, arrow=arrow_style) + geom_text(aes(x=PC1, y=PC2, label=
variable), hjust=0, size=3, color='red') + xlim(-.05,1) + ylim(-.5,.5) + coord_fixed() # fix aspect ratio to 1:1
```



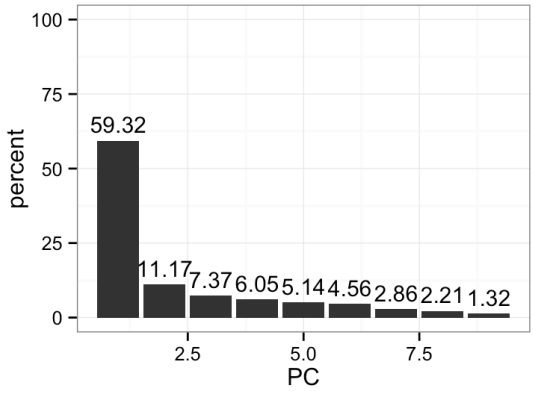
Question 3 (4 pts): Given the four plots from Questions 1 and 2, how do you interpret PC1 and PC2? What does PC1 tell you about a data point? What does PC2 tell you about a data point?

Neither PC1 nor PC2 tell us anything about Age or Survival, but from the first plot it is evident that the Sex can be separated by PC2.

PC1 is positively correlated with the body size of the birds but there is no particular correlation with PC2 as it is negatively correlated with some attributes and positively correlated with others.

Question 4 (1 pt): What percentage of the variation in the data does PC1 explain?

```
percent <- 100*pca$sdev^2/sum(pca$sdev^2)
perc_data <- data.frame(percent=percent, PC=1:length(percent))
ggplot(perc_data, aes(x=PC, y=percent)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=round(percent, 2)), size=4, vjust=-.5) + ylim(0,100)
```



PC1 explains 59.32% of the variation in the data

Question 5 (1 pt): Does the PCA suggest any specific physical characteristics for birds that survived? Consider only PC1 and PC2 for your answer.

No, neither PC1 nor PC2 suggest any physical characteristics for birds that survived