Lab Worksheet 5

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:

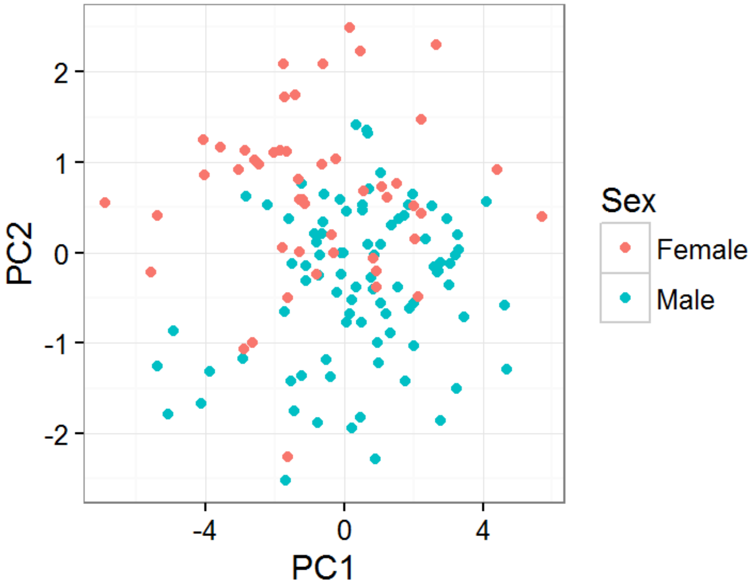
```
bumpus <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/bumpus_full.csv")
head(bumpus)
```

```
##    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
```

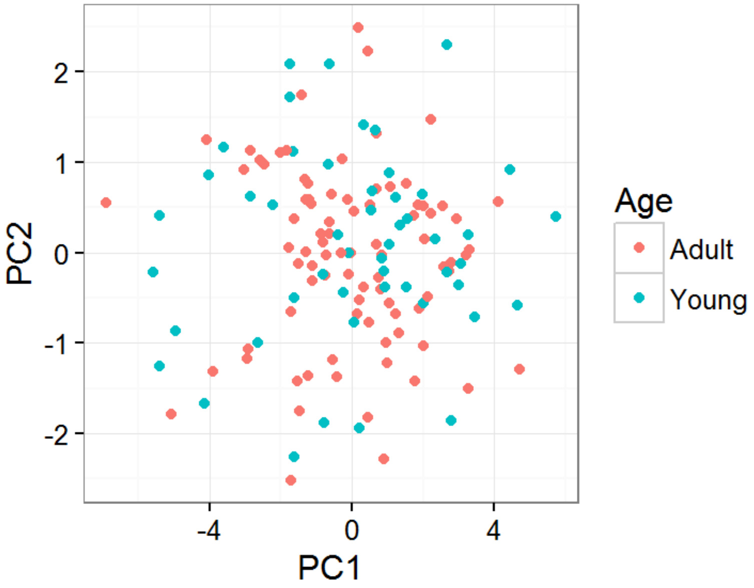
The data set has three categorical variables (`Sex`, with levels `Male` and `Female`, `Age`, with levels `Adult` and `Young`, and `Survival`, with levels `Alive` and `Dead`) and nine numerical variables that hold various aspects of the birds' anatomy, such as wingspread, weight, etc.

Question 1: 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.

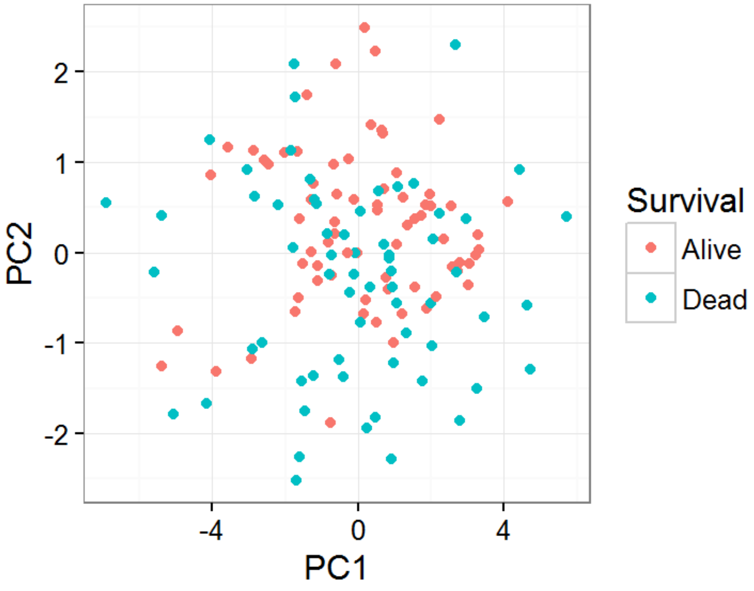
```
bumpus %>% select(-Sex, -Age, -Survival) %>%
  scale() %>%
  prcomp() ->
  pca

bumpus.pca <- data.frame(bumpus, pca$x)
ggplot(bumpus.pca, aes(x=PC1, y=PC2, color=Sex)) + geom_point()
```



```
ggplot(bumpus.pca, aes(x=PC1, y=PC2, color=Age)) + geom_point()
```

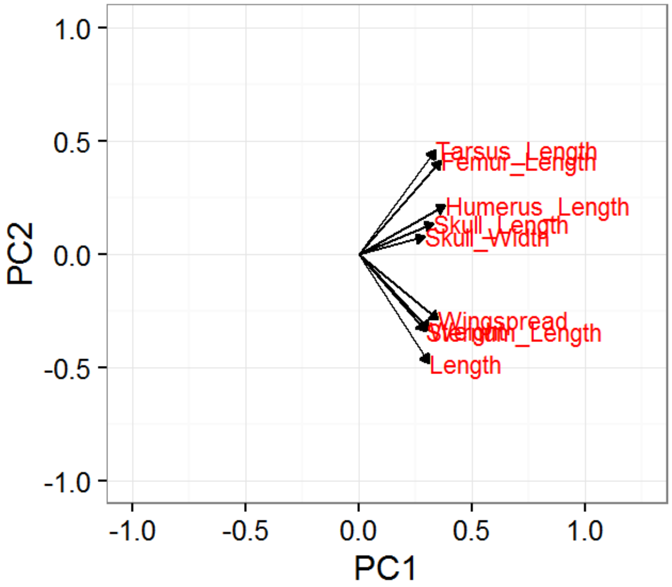


```
ggplot(bumpus.pca, aes(x=PC1, y=PC2, color=Survival)) + geom_point()
```



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

```
rotation_data <- data.frame(pca$rotation, variable=row.names(pca$rotation))

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(-1.,1.25) +
  ylim(-1.,1.) +
  coord_fixed() # fix aspect ratio to 1:1
```



Question 3: 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?

PC1 measures the overall body size of the birds. All the variables contribute to PC1, hence the larger the animal, the larger the PC1 value.

PC2 measures difference between male and female birds. Female Birds score positvely, and male birds score negatively on PC2.

Question 4: What percentage of the variation in the data does PC1 explain?

```
100*pca$sdev^2/sum(pca$sdev^2)
```

```
## [1] 59.322242 11.171593  7.369513  6.048066  5.141103  4.555629  2.861469
## [8]  2.209724  1.320662
```

PC1 explains 59.32 % of the variation *Question 5: Does the PCA suggest any specific physical characteristics for birds that survived? Consider only PC1 and PC2 for your answer.*

Not really, dead and alive birds seem are in multiple places onthe PC2 vs PC1 plot. At most there is a minor tendency for dead birds to score more negative on PC2.