STAT430 Assignment 4: Unsupervised Learning

Ben Murarotto

May 2025

Contents

1 Spinner dolphins dataset.

Spinner dolphins are small delphinids that reside in subtropical and tropical waters. Data were collected on spinner dolphins in Fiji to explore the characteristics of their whistles. The type of whistle (e.g., concave, constant, convex, downsweep, sine, and upsweep) along with a variety of other acoustic properties of individual whistles was measured and is summarised in the Spinner dolphin dataset.

Table 1: Whistle Variable Descriptions

Variable	Description
Whistle	Whistle classification (Concave, Constant, Convex, Downsweep, Sine or Upsweep)
Duration	The time span of the whistle (seconds)
Centre Freq	The frequency recorded at the midpoint (centre) of the whistle (kHz)
Low Freq	The lowest frequency recorded during a whistle (kHz)
Delta Freq	The range in frequency recorded during a whistle (kHz)
Max Freq	Maximum frequency (kHz) recorded during a whistle
Range 50	Centre frequency minus minimum frequency
Range 100	Maximum frequency minus centre frequency
Inflections	Number of points of change in whistle curvature

1.1 Exploratory analysis.

```
df <- read.csv("Spinner_Dolphin.csv", header = T)
names(df)
```

```
## [1] "Whistle"     "Duration"    "Center_Freq" "Low_Freq"    "Delta_Freq"
## [6] "Max_Freq"    "Range_50"    "Range_100"   "Inflections"
```

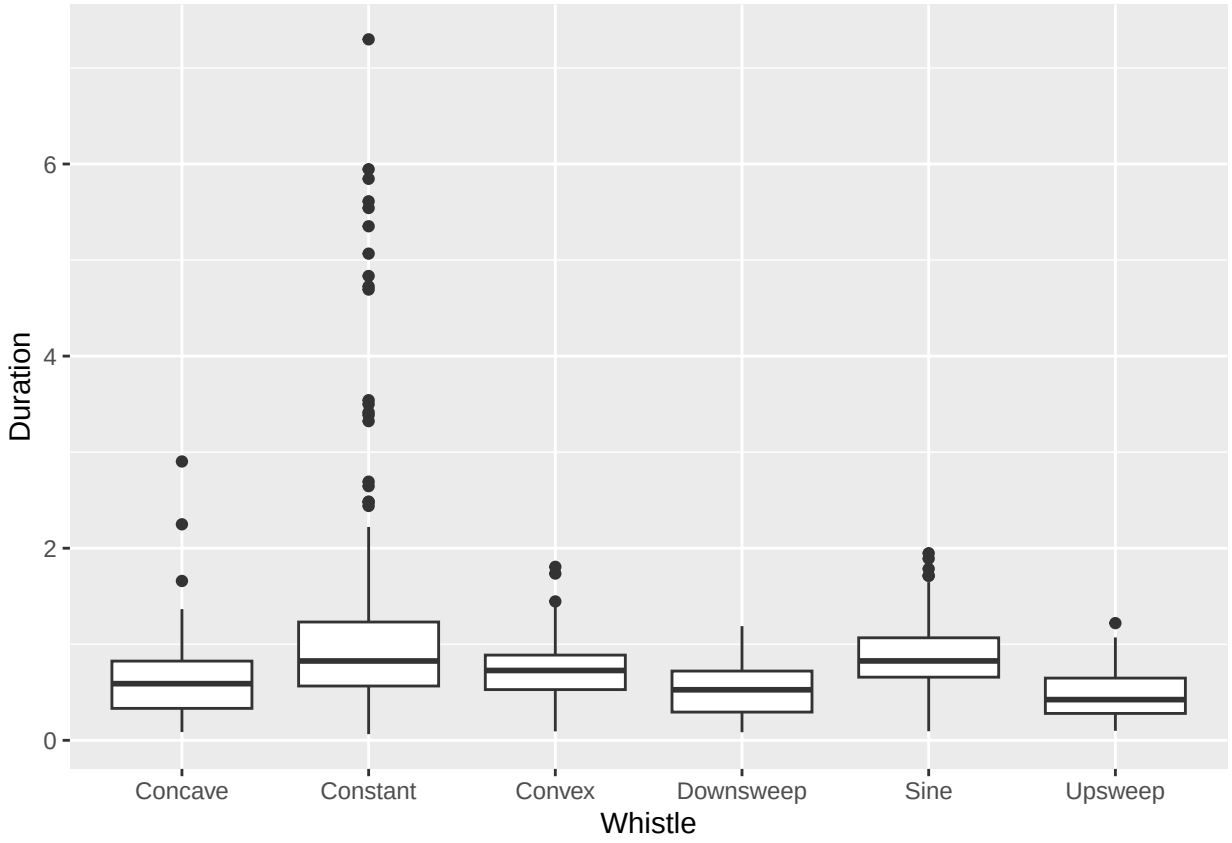
```
head(df)
```

```
##   Whistle Duration Center_Freq Low_Freq Delta_Freq Max_Freq Range_50 Range_100
## 1 Concave    0.761     11197.3   7289.3    14760.7   9302.3   3908.0   -1895.0
## 2 Concave    1.074      9302.3   3150.0    18900.0   3273.0   6152.3   -6029.3
## 3 Concave    1.006      8096.5   1457.9    20592.1   3445.3   6638.6   -4651.2
## 4 Concave    1.004     12403.1   7836.0    14214.0  11197.3   4567.1   -1205.8
## 5 Concave    0.851     11197.3   5573.1    16476.9  11369.5   5624.2     172.2
## 6 Concave    1.077     10335.9   4627.8    17422.2   4651.2   5708.1   -5684.7
##   Inflections
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

```
summary(df)
```

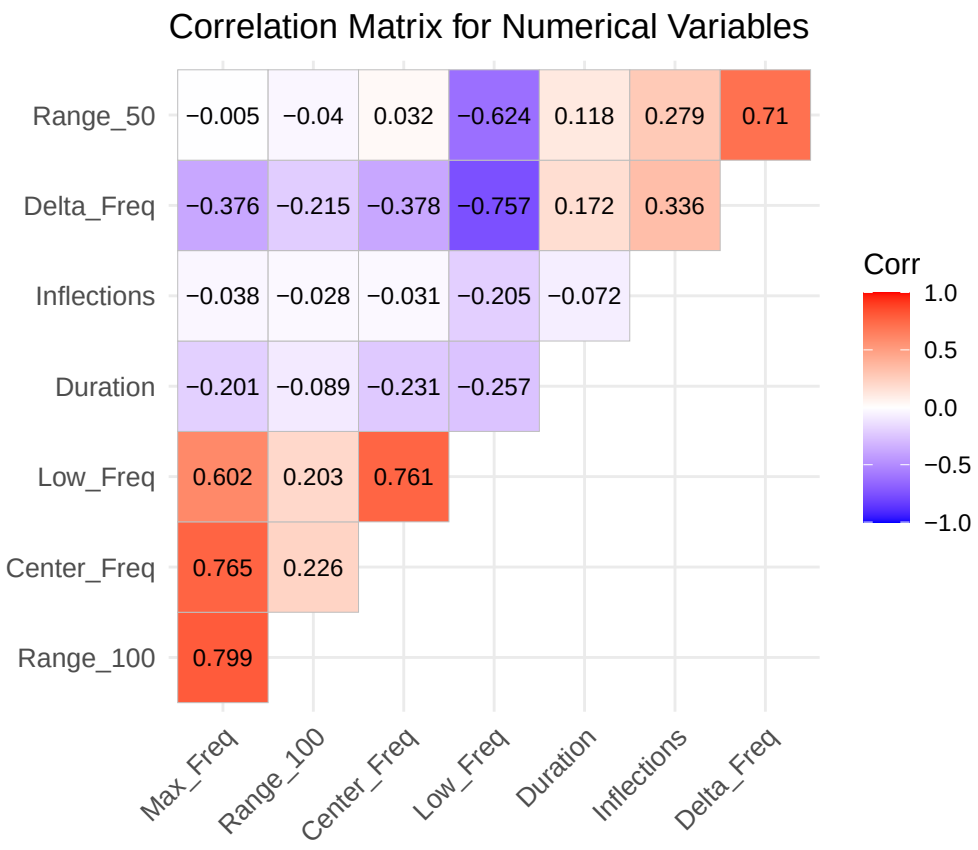
```
##    Whistle             Duration       Center_Freq       Low_Freq
##  Length:1800        Min.   :0.0640   Min.   : 2756   Min.   :    0
##  Class :character   1st Qu.:0.4040   1st Qu.:11025   1st Qu.: 5649
##  Mode  :character   Median :0.6525   Median :12231   Median : 7512
##                     Mean   :0.7092   Mean   :12311   Mean   : 7617
##                     3rd Qu.:0.8790   3rd Qu.:13437   3rd Qu.: 9476
##                     Max.   :7.2980   Max.   :20327   Max.   :16800
##    Delta_Freq       Max_Freq          Range_50       Range_100
##  Min.   : 2734   Min.   :  689.1   Min.   :  383   Min.   :-8785.6
##  1st Qu.: 9840   1st Qu.: 8785.5   1st Qu.: 3466   1st Qu.:-2584.0
##  Median :12392   Median :11369.5   Median : 4536   Median : -689.1
##  Mean   :12333   Mean   :11213.1   Mean   : 4694   Mean   :-1094.4
##  3rd Qu.:14949   3rd Qu.:13436.7   3rd Qu.: 5766   3rd Qu.:  172.3
##  Max.   :22050   Max.   :21533.2   Max.   :11899   Max.   : 7924.2
##   Inflections
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.055
##  3rd Qu.:1.000
##  Max.   :3.000
```

```
library(ggplot2)
df$Whistle <- as.factor(df$Whistle)
ggplot(df, aes(x = Whistle, y = Duration)) + geom_boxplot()
```

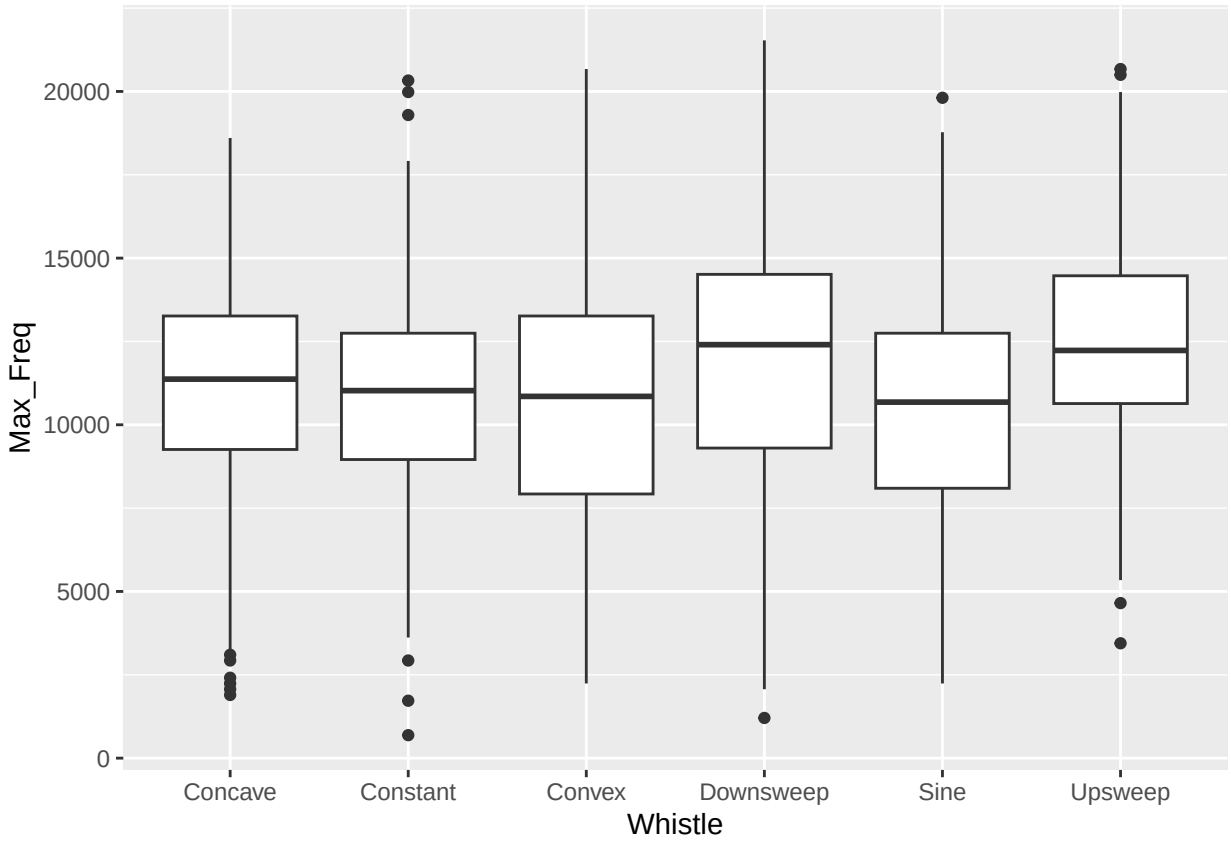



```
num.df <- subset(df, select = -1)
```

```
library(ggcorrplot)
corr_matrix <- cor(num.df)
ggcorrplot(corr_matrix,
           method = "square",
           lab = TRUE,
           lab_size = 3,
           lab_col = "black",
           hc.order = TRUE,
           type = "upper",
           tl.cex = 10,
           digits = 3,
           title = "Correlation Matrix for Numerical Variables",
           show.legend = TRUE)
```




```
ggplot(df, aes(x = Whistle, y = Max_Freq)) + geom_boxplot()
```

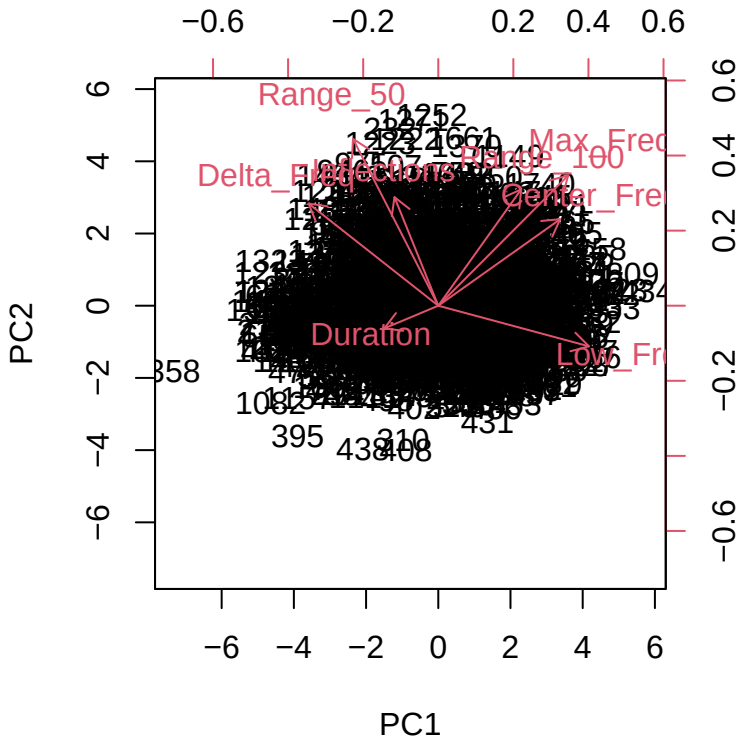


```r
set.seed(8)
s <- sample(1:nrow(df), 20)
num.df <- subset(df, select = -1)
subset.df <- num.df[s, ]
```

```r
pr.out <- prcomp(num.df, scale = TRUE)
pr.out$rotation[, 1:4]
```

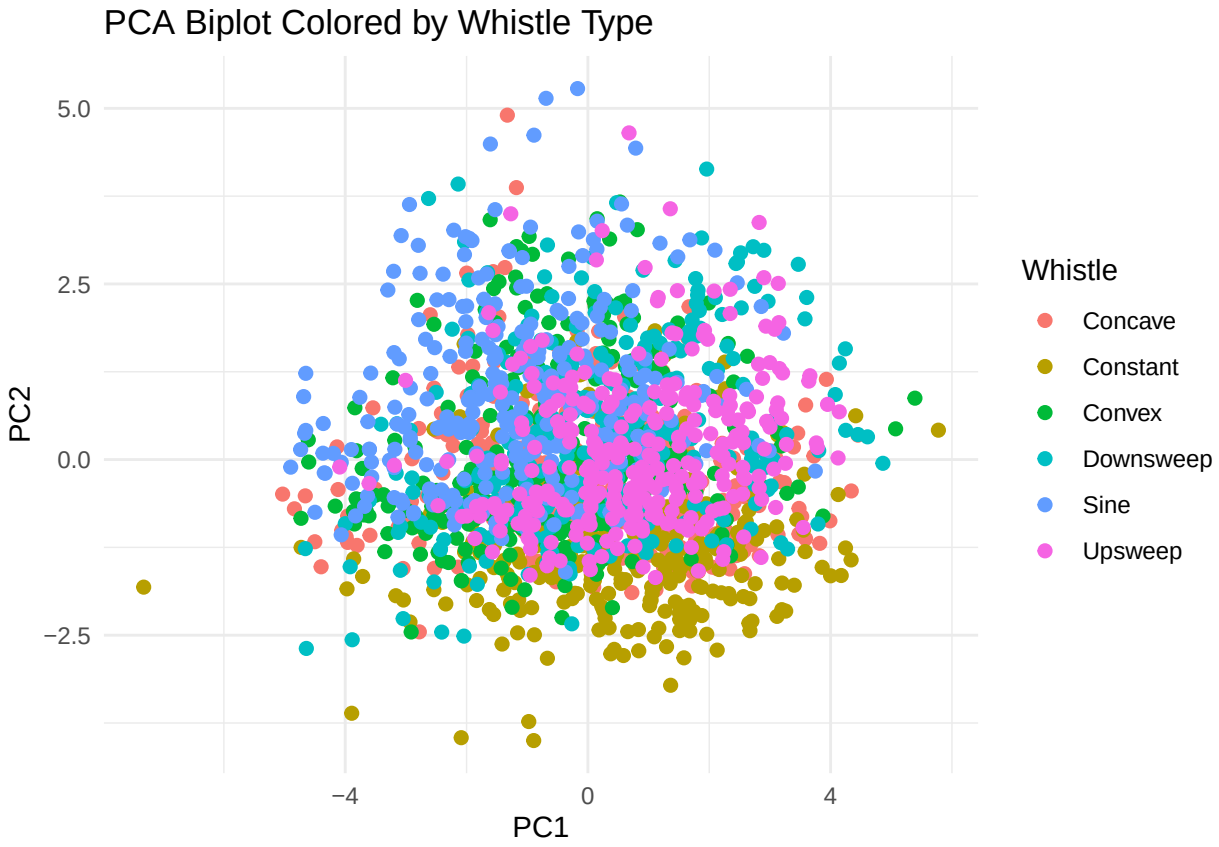
```
##                    PC1         PC2         PC3         PC4
## Duration    -0.1796201 -0.07542078  0.67391834  0.46922617
## Center_Freq  0.4059887  0.28806469 -0.19045114  0.54797470
## Low_Freq     0.5016894 -0.13514599 -0.16469104  0.23131309
## Delta_Freq  -0.4315627  0.33994868 -0.02631764  0.06299618
## Max_Freq     0.4323704  0.43372889  0.18365969 -0.02593695
## Range_50    -0.2838887  0.55555277  0.02427076  0.30401354
## Range_100    0.2759778  0.38740221  0.45527732 -0.55188993
## Inflections -0.1458443  0.36147033 -0.49005029 -0.15619477
```

```r
biplot(pr.out, scale = 0)
```



Looking at the arrows in our biplot we can observe the representations of our variables. The first PC places most weight on Low_Freq. Low_Freq and and Duration variables are aligned almost exclusively to PC1. The remaining variables are at diagonals to the principal components meaning they share influence across both PCs and reflects their moderate correlation. The frequency related variables (Max_Freq, Center_Freq, Range_100) are clustered suggesting that these variables are positively correlated and hold similar contributions to this low dimensional representation.

```
pca.df <- as.data.frame(pr.out$x)
pca.df$Whistle <- df$Whistle
ggplot(pca.df, aes(x = PC1, y = PC2, color = Whistle)) +
  geom_point(size = 2) +
  labs(title = "PCA Biplot Colored by Whistle Type",
       x = "PC1",
       y = "PC2") +
  theme_minimal()
```



Constant whistles appear slightly more concentrated along the lower edge of the PC2 axis, suggesting they may have lower variation in the features captured by PC2 which we determined is the opposite direction of the frequency variables and closer to the direction of Duration. This suggests that constant whistles may be characterised by longer durations but lower variation in frequency, distinguishing them from other whistle types. Upsweep and Sine are centralised in this visualisation and grow towards the right and left along the PC1 axis respectively and upwards along the PC2 axis. We could possibly determine that our Upsweep whistles are associated with higher frequencies that peak towards the midpoint of the whistle. Sine whistles on the other hand are more associated with longer duration and greater number of inflexions. Concave, Convex, and Downsweep whistles show moderate dispersion across both PC1 and PC2, with no clear separation, suggesting overlapping feature characteristics in the reduced space.

2 Snails dataset.

In this segment, I am using the Snails dataset. This dataset contains 10 variables listed below

Table 2: Variable Descriptions

Variable	Description
Location	Location of site
Aperture width (AW)	Width of the opening into the shell
Aperture length (AL)	Length of the opening into the shell
Circularity	AW / AL
N.Telochonc.Whorls	Number of whorls in the telochonch
Whorl.First.Primary	Whorl number in which first primary groove (PG) appears
N.Primary.Whorls	Number of primary whorls
Whorl.Axial.Striae	Whorl in which axial striae (AS) first appear
Whorl.Tertiary.Groove	Whorl in which tertiary grooves (TG) first appear
Whorl.Spiral.Striae	Whorl in which spiral striae (SS) first appear

```
snail.df <- read.csv("Snails_Dataset.csv", header = T)
head(snail.df)
```

```
##   Location ApertureWidth ApertureLength Circularity N.Telochonc.Whorls
## 1     Cuba           7.7            9.8        0.79                  7
## 2     Cuba           8.6           10.8        0.80                  7
## 3     Cuba           7.1            9.0        0.79                  7
## 4     Cuba           7.9           10.9        0.72                  6
## 5     Cuba           7.2            9.8        0.73                  7
## 6     Cuba           6.8            8.7        0.78                  7
##   Whorl.First.Primary N.Primary.Whorls Whorl.Axial.Striae Whorl.Tertiary.Groove
## 1                   3                8                  2                     6
## 2                   3                5                  3                     0
## 3                   3                4                  3                     7
## 4                   2                4                  2                     6
## 5                   3                3                  2                     0
## 6                   3                5                  3                     0
##   Whorl.Spiral.Striae
## 1                   6
## 2                   0
## 3                   0
## 4                   0
## 5                   5
## 6                   0
```

```
unique(snail.df$Location)
```

```
##  [1] "Cuba"      "Brazil"    "Bahamas"   "Nicaragua" "Haiti"     "Angola"
##  [7] "Congo"     "Bermuda"   "Belize"    "Jamaica"   "Florida"   "Gabon"
## [13] "Nigeria"   "Senegal"   "Liberia"   "Sleone"    "Ghana"     "Principe"
```

2.1 Performing K-means clustering.

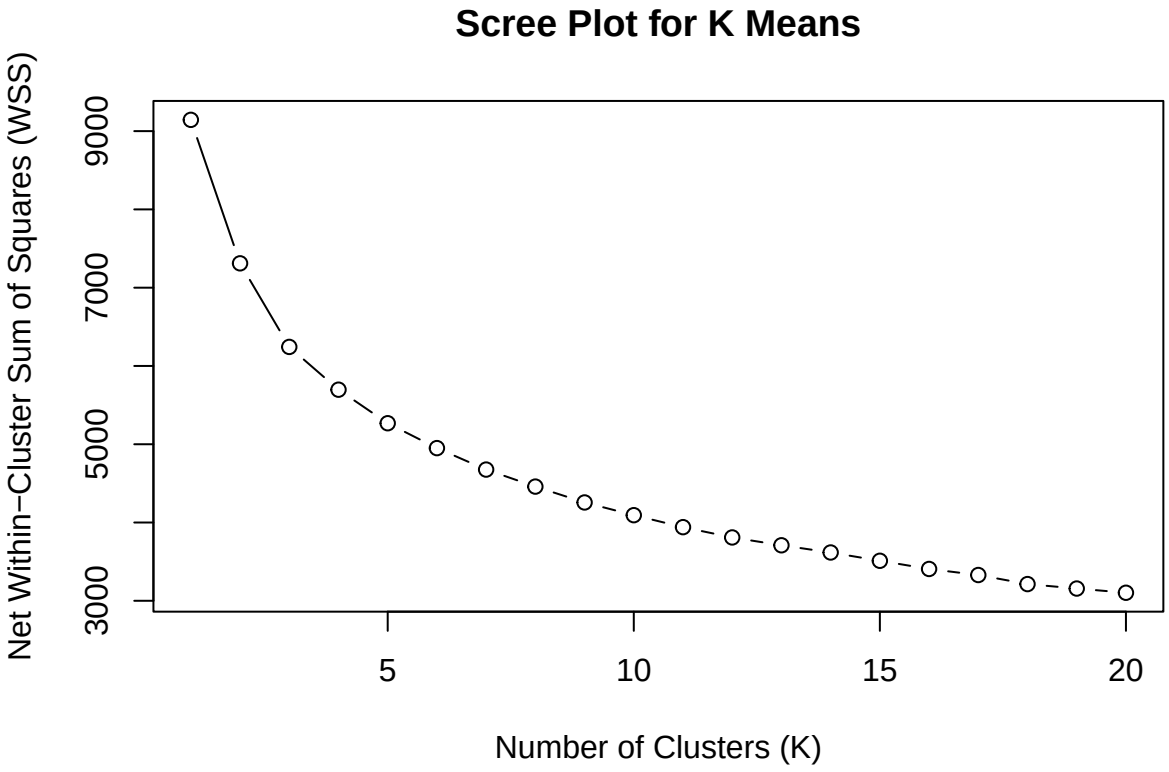
Let's use k-means clustering to partition a low dimensional representation of our dataset into groups. To select an ideal k (number of groups) we plot a range of k's can using a scree plot to determine if a certain value for k causes within-cluster error to plateau significantly.

```
num.df <- subset(snail.df, select = -1)
pr.out <- prcomp(num.df, scale = TRUE) ## distance based method so we scale values


pca.df <- as.data.frame(pr.out$x)
```

```
wss <- numeric()

for (k in 1:20) {
  km <- kmeans(pca.df, centers = k, nstart = 20)
  wss[k] <- km$tot.withinss
}
```

```
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
```

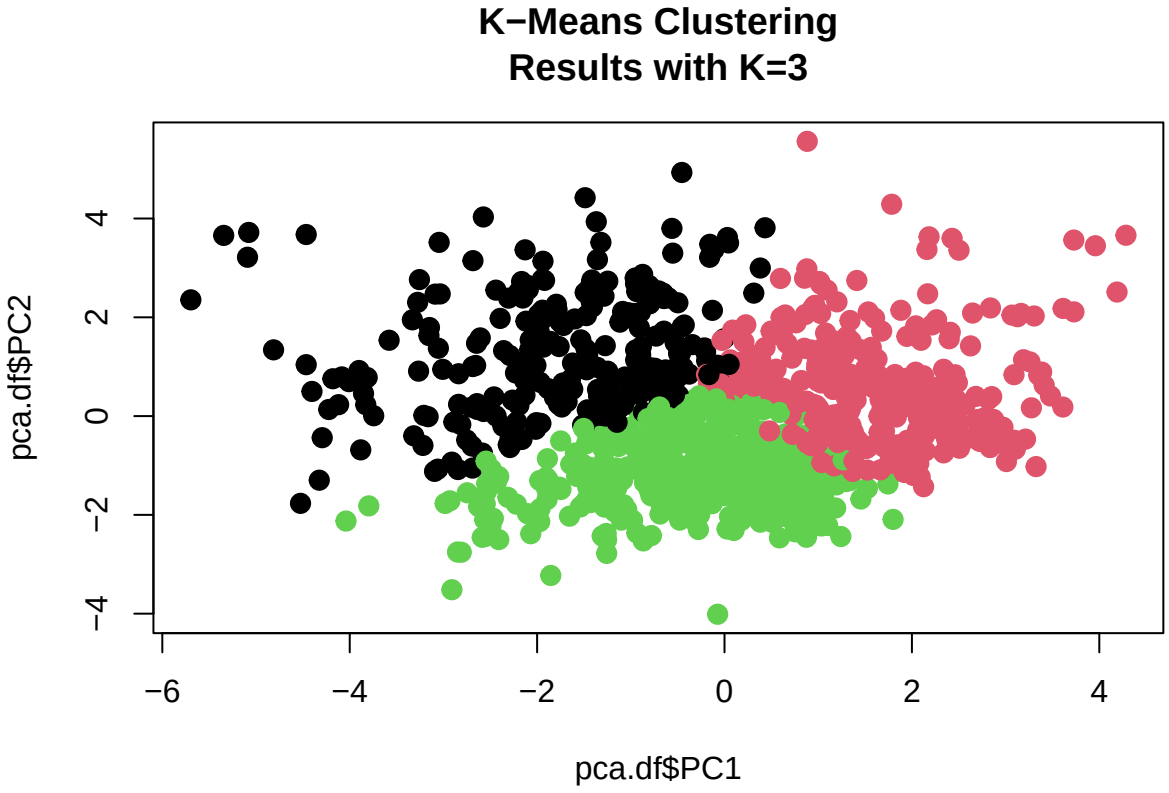
```
plot(1:20, wss, type = "b",
     xlab = "Number of Clusters (K)",
     ylab = "Net Within-Cluster Sum of Squares (WSS)",
     main = "Scree Plot for K Means")
```



There is no sharp tapering off in the plot but the rate of error becomes gradual after k = 3, therefore this will likely be our chosen k.

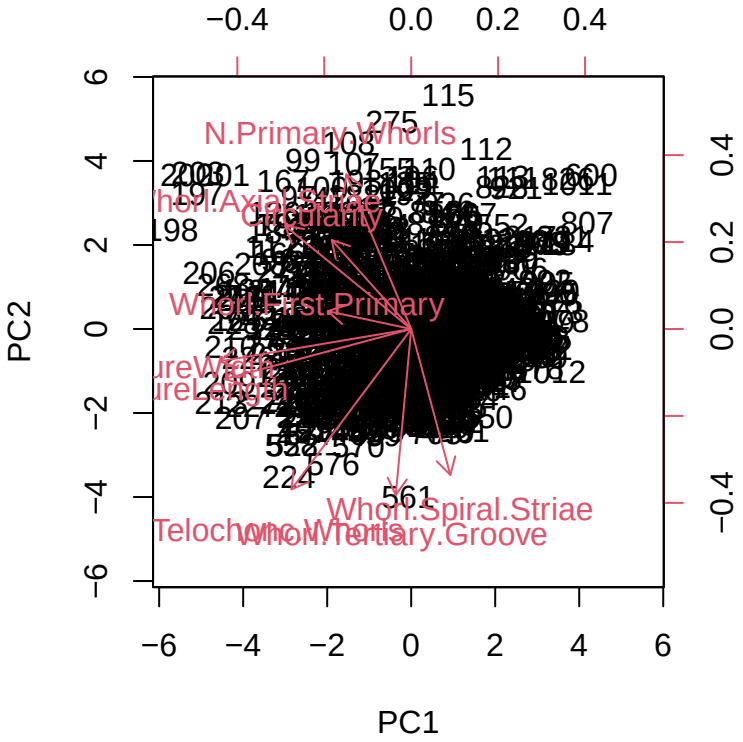
```
k <- 3
km.out <- kmeans(pca.df, k, nstart = 20)
```

```
plot(pca.df$PC1, pca.df$PC2,
col=(km.out$cluster), main=paste0("K-Means Clustering
Results with K=",k), pch=20, cex=2)
```



Let's now make a biplot to determine if certain variables are more closely associated with our clusters.

```
biplot(pr.out, scale = 0)
```



When contrasting plots we can see distinct variations in snails appear in our groups based differences in physical attributes. Our PC2 dimension holds significant weight in aperture length and width. Our k-means cluster has identified a potential type of sub species (green cluster) of snail which had smaller apertures on average. Our red cluster was also associated with key characteristics - such as Circularity, N.Primary.Whorls and Whorl.Axial.Striae. This cluster of snail share distinct features such as consistent aperture width/length ratio, a larger number primary whorls and axial striae appearing on the same whorl. This provides potential evidence for a different sub species of snail. Our black cluster group are associated with N.Telochonc.Whorls, Whorl.Spiral.Striae and Whorl.Tertiary.Groove. Black cluster snails saw similar shell location for spiral striae and tertiary groove as well as a greater number of whorls in the telochonch. These consistent yet distinct shell features hint at another snail subspecies.

It is important to note that k-means does not provide a true classification or confirmation of biological subspecies, but rather an unsupervised grouping based on patterns in the data. Further validation, such as genetic analysis or expert assessment, would be required to confirm whether these clusters correspond to distinct taxonomic groups however we have made strong use of an exploratory tool.

2.2 Hierarchical clustering.

Let's create some dendrograms using hierarchical clustering and the Location variable as our identifier.

```
set.seed(50)
n <- 100
s <- sample(nrow(num.df), n)

sample.df <- num.df[s, ]
```

```
scale.df <- scale(sample.df)
snail.labels <- snail.df[s,]$Location
dist.data <- dist(sample.df)
```

```r
hc1 <- hclust(dist.data, method = "complete")
```

```r
library(dendextend)
```

```
##
## ---------------------
## Welcome to dendextend version 1.19.0
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## You may ask questions at stackoverflow, use the r and dendextend tags:
##   https://stackoverflow.com/questions/tagged/dendextend
##
##  To suppress this message use:  suppressPackageStartupMessages(library(dendextend))
## ---------------------
```

```
##
## Attaching package: 'dendextend'
```

```
## The following object is masked from 'package:stats':
##
##     cutree
```

```r
dend <- as.dendrogram(hc1)
labels(dend) <- snail.labels

pdf("dendrogram1.pdf", width = 20, height = 15)
dend %>%
  set("leaves_pch", 19) %>%
  set("leaves_cex", 2) %>%
  set("leaves_col", 2) %>% # adjust the leaves
  hang.dendrogram(hang_height = .6) %>% # hang the leaves
   plot(main = "Hierarchical Clustering with Complete Linkage",  ylim = c(0, 20))
dev.off()
```

```
## pdf
##   2
```

There are several key insights from the hierarchical clustering analysis of snail samples: A large portion of the observations cluster tightly near the base of the dendrogram, suggesting low overall variability in snail features across geographic locations. At a finer scale, snails tend to group consistently within their respective locations, indicating intralocation similarity in shell characteristics. Several snails from African regions cluster together closely specifically Angola, Ghana, and Sierra Leone clustered closely indicating intralocation similarity. Snails from the Americas (Bahamas, Jamaica, and Cuba) also formed tight clusters, Jamaicas specifically with itself, indicating morphological consistency across those locations. Florida snails exhibited the largest number of observations and the broadest spread, of dissimilarity range suggesting greater morphological diversity in this state. Notably, outlier observations were found in Belize and Principe, many of the Belize snails merged higher up indicating where snails displayed substantial divergence from all other clusters.

Let's change the distance measurement to account for correlation and then plot another hierarchical cluster.

```r
hc2 <- hclust(dist.data, method = "single")
dend <- as.dendrogram(hc2)
labels(dend) <- snail.labels

pdf("dendrogram2.pdf", width = 20, height = 15)
dend %>%
  set("leaves_pch", 19) %>%
  set("leaves_cex", 2) %>%
  set("leaves_col", 2) %>% # adjust the leaves
  hang.dendrogram(hang_height = .6) %>% # hang the leaves
   plot(main = "Hierarchical Clustering with Single Linkage", ylim = c(0, 20))
dev.off()
```

```
## pdf
##   2
```

There are several key insights from the hierarchical clustering analysis of snail samples using single linkage. A large proportion of the observations cluster tightly near the base of the dendrogram, indicating generally low morphological variability across geographic locations. Due to the chaining effect inherent in single linkage, many samples are progressively merged based on minimal pairwise dissimilarities, which gives us less distinct clusters than our complete linkage. Nonetheless, some regional patterns are apparent. Snails from Sierra Leone frequently clustered closely, suggesting strong intralocation similarity, with Angola and Ghana also showing local cohesion in parts of the dendrogram. Jamaica exhibited consistent clustering within its own samples. In contrast, Florida snails were widely distributed throughout the tree and merged at varying heights, indicating variations within that population. Several samples from Belize merged at relatively higher distance levels, highlighting their distinctiveness, while snails from Principe also appeared as clear outliers like in our previous visualisation.