ClusterAnalysis\_\_Skunky\_vs\_Sweet

Valerie Martin

10/27/2019

setwd("~/Desktop/Skypilot 2017")  
library(ggplot2)  
library(plyr)  
library(nlme)  
library(car)

## Loading required package: carData

library(emmeans)  
library(multcompView)  
library(multcomp)

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ tibble 2.1.3 ✔ purrr 0.3.2  
## ✔ tidyr 0.8.3 ✔ dplyr 0.8.3  
## ✔ readr 1.3.1 ✔ stringr 1.4.0  
## ✔ tibble 2.1.3 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::arrange() masks plyr::arrange()  
## ✖ dplyr::collapse() masks nlme::collapse()  
## ✖ purrr::compact() masks plyr::compact()  
## ✖ dplyr::count() masks plyr::count()  
## ✖ dplyr::failwith() masks plyr::failwith()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::id() masks plyr::id()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::mutate() masks plyr::mutate()  
## ✖ dplyr::recode() masks car::recode()  
## ✖ dplyr::rename() masks plyr::rename()  
## ✖ dplyr::select() masks MASS::select()  
## ✖ purrr::some() masks car::some()  
## ✖ dplyr::summarise() masks plyr::summarise()  
## ✖ dplyr::summarize() masks plyr::summarize()

library(cluster)   
library(factoextra)

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

stigma <- read.csv("Skypilot 2017 (MASTER) Tab2 08-12-19\_\_Purity added.csv")  
cols <- c("Specimen.No.", "Year", "Era", "Species", "Scent.Morph", "Date", "Location", "Altitude..m.",  
 "Habitat..NMS.", "Corolla.Flare\_1", "Sepal.Length\_1", "Tube.Length\_1", "Lobe.Length\_1", "Lobe.Width\_1",   
 "No.PVgrains\_stigma", "No.PVgrains\_off", "No.OtherGrains", "No.OtherTypes")  
stigma.sm <- subset(stigma, select = cols)  
  
#-----# summarize data & add columns where appropriate#-----#  
ddply(stigma.sm, .(Era,Location, Altitude..m., Scent.Morph), summarise, con=mean(No.PVgrains\_stigma, na.rm=T), het=mean(No.OtherGrains, na.rm=T))

## Era Location Altitude..m. Scent.Morph con  
## 1 past Cumberland Pass 3620.00 Skunky 59.50000  
## 2 past Cumberland Pass 3620.00 Sweet 52.50000  
## 3 past Cumberland Pass 3710.00 Skunky 58.83333  
## 4 past Cumberland Pass 3710.00 Sweet 70.00000  
## 5 past Niwot Ridge 3500.00 Sweet 190.00000  
## 6 past Niwot Ridge 3660.00 Skunky 107.66667  
## 7 past Niwot Ridge 3660.00 Sweet 34.75000  
## 8 past Niwot Ridge 3750.00 Skunky 231.40000  
## 9 past Niwot Ridge 3750.00 Sweet 179.16667  
## 10 past Penn Mountain 3550.00 Skunky 42.75000  
## 11 past Penn Mountain 3550.00 Sweet 78.50000  
## 12 past Penn Mountain 3600.00 Skunky NaN  
## 13 past Penn Mountain 3600.00 Sweet NaN  
## 14 past Penn Mountain 3660.00 Skunky NaN  
## 15 past Penn Mountain 3660.00 Sweet NaN  
## 16 past Penn Mountain 3700.00 Skunky NaN  
## 17 past Penn Mountain 3700.00 Sweet NaN  
## 18 past Penn Mountain 3780.00 Skunky NaN  
## 19 past Penn Mountain 3780.00 Sweet NaN  
## 20 past Penn Mountain 3800.00 Skunky NaN  
## 21 past Penn Mountain 3800.00 Sweet NaN  
## 22 past Penn Mountain 3880.00 Skunky NaN  
## 23 past Penn Mountain 3880.00 Sweet NaN  
## 24 past Penn Mountain 3950.00 Skunky 171.50000  
## 25 past Penn Mountain 3950.00 Sweet 79.00000  
## 26 past Penn Mountain 3960.00 Skunky 40.50000  
## 27 past Penn Mountain 3960.00 Sweet 82.50000  
## 28 past Penn Mountain 4025.00 Skunky 39.00000  
## 29 past Penn Mountain 4025.00 Sweet 47.00000  
## 30 past Rocky Mtn NP 3565.00 Skunky NaN  
## 31 past Rocky Mtn NP 3565.00 Sweet NaN  
## 32 past Rocky Mtn NP 3670.00 Skunky NaN  
## 33 past Rocky Mtn NP 3670.00 Sweet NaN  
## 34 past Rocky Mtn NP 3825.00 Skunky 11.25000  
## 35 past Rocky Mtn NP 3825.00 Sweet 56.66667  
## 36 past San Francisco Peaks 3596.64 Skunky NaN  
## 37 past San Francisco Peaks 3596.64 <NA> NaN  
## 38 past San Francisco Peaks 3657.60 Skunky NaN  
## 39 past San Francisco Peaks 3703.32 Skunky NaN  
## 40 past San Francisco Peaks 3703.32 <NA> NaN  
## 41 past San Francisco Peaks 3749.04 Skunky NaN  
## 42 past San Francisco Peaks 3749.04 Sweet NaN  
## 43 past San Francisco Peaks 3749.04 <NA> NaN  
## 44 past San Francisco Peaks NA <NA> NaN  
## 45 present Cumberland Pass 3678.00 Skunky 93.40000  
## 46 present Cumberland Pass 3678.00 Sweet 33.75000  
## 47 present Cumberland Pass 3761.00 Skunky 63.20000  
## 48 present Cumberland Pass 3761.00 Sweet 100.60000  
## 49 present Niwot Ridge 3471.00 Skunky 112.87500  
## 50 present Niwot Ridge 3471.00 Sweet 115.85714  
## 51 present Niwot Ridge 3746.00 Sweet 181.14286  
## 52 present Penn Mountain 3573.00 Skunky 146.28571  
## 53 present Penn Mountain 3573.00 Sweet 93.71429  
## 54 present Penn Mountain 3955.00 Skunky 103.42857  
## 55 present Penn Mountain 3955.00 Sweet 100.16667  
## 56 present Rocky Mtn NP 3586.00 Skunky 63.60000  
## 57 present Rocky Mtn NP 3586.00 Sweet 113.60000  
## 58 present Rocky Mtn NP 3739.00 Skunky 6.00000  
## 59 present Rocky Mtn NP 3739.00 Sweet 73.14286  
## het  
## 1 140.000000  
## 2 40.000000  
## 3 260.833333  
## 4 301.800000  
## 5 744.000000  
## 6 244.333333  
## 7 11.750000  
## 8 73.800000  
## 9 134.333333  
## 10 74.250000  
## 11 103.000000  
## 12 NaN  
## 13 NaN  
## 14 NaN  
## 15 NaN  
## 16 NaN  
## 17 NaN  
## 18 NaN  
## 19 NaN  
## 20 NaN  
## 21 NaN  
## 22 NaN  
## 23 NaN  
## 24 403.000000  
## 25 229.000000  
## 26 174.500000  
## 27 247.250000  
## 28 25.000000  
## 29 36.000000  
## 30 NaN  
## 31 NaN  
## 32 NaN  
## 33 NaN  
## 34 256.000000  
## 35 646.000000  
## 36 NaN  
## 37 NaN  
## 38 NaN  
## 39 NaN  
## 40 NaN  
## 41 NaN  
## 42 NaN  
## 43 NaN  
## 44 NaN  
## 45 156.400000  
## 46 205.250000  
## 47 303.400000  
## 48 191.400000  
## 49 263.000000  
## 50 284.285714  
## 51 91.000000  
## 52 78.285714  
## 53 81.000000  
## 54 631.857143  
## 55 110.500000  
## 56 652.200000  
## 57 426.400000  
## 58 6.666667  
## 59 164.571429

stigma.sm$No.TotalPVgrains <- stigma.sm$No.PVgrains\_stigma + stigma.sm$No.PVgrains\_off  
stigma.sm$Purity.tot <- stigma.sm$No.TotalPVgrains/(stigma.sm$No.TotalPVgrains + stigma.sm$No.OtherGrains)  
stigma.sm$Purity.tot <- as.numeric(as.character(stigma.sm$Purity.tot))  
stigma.sm$Purity.on <- stigma.sm$No.PVgrains\_stigma/(stigma.sm$No.PVgrains\_stigma + stigma.sm$No.OtherGrains)  
stigma.sm$Corolla.Length <- stigma.sm$Tube.Length\_1 + stigma.sm$Lobe.Length\_1  
  
#-----# subsets for analysis #-----#  
# subset to complete cases  
stigma.sm <- subset(stigma.sm, stigma.sm$Corolla.Flare\_1 > 0 & stigma.sm$Sepal.Length\_1 > 0 & stigma.sm$Tube.Length\_1 > 0 &   
 stigma.sm$Lobe.Length\_1 >0 & stigma.sm$Lobe.Width\_1 >0) #eliminates all flowers without complete measurements  
# subset by morph  
stigma.sw <- subset(stigma.sm, Scent.Morph =="Sweet") #subset to sweet morph  
stigma.sk <- subset(stigma.sm, Scent.Morph =="Skunky") #subset to skunky morph  
nohab <- subset(stigma.sm, select = -c(stigma$Habitat..NMS.)) #subset to eliminate habitat  
stigma.new <- subset(stigma.sm, Era == "present") #selects for new era  
stigma.sw.new <- subset(stigma.sw, Era == "present") #selects for new sweet  
stigma.sw$No.Totalgrains <- stigma.sw$No.TotalPVgrains+stigma.sw$No.OtherGrains  
stigma.sk.new <- subset(stigma.sk, Era == "present") #selects for new skunky  
stigma.sk$No.Totalgrains <- stigma.sk$No.TotalPVgrains+stigma.sk$No.OtherGrains

## Cluster analysis using k-clustering

#### In k-clustering, we choose the number of clusters

Based on tips I found on this website: <https://uc-r.github.io/kmeans_clustering>

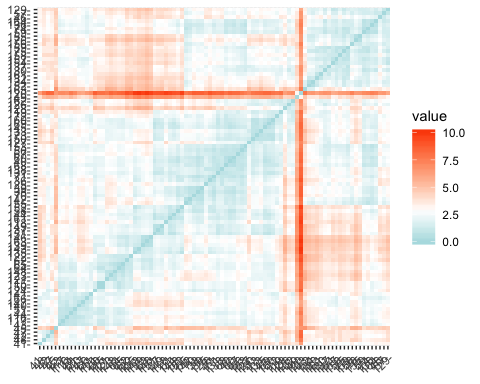
#### Here we see that there are two outliers, individuals 49 and 120.

These appear to be simply larger flowers, not necessarily abnormal (I changed an outlier initially found that had a sepal length of 131 - this was changed in the CSV file to 1.31)

stigma.new.flwrSizeOnly <- subset(stigma.new, select = c("Corolla.Flare\_1", "Sepal.Length\_1", "Tube.Length\_1", "Lobe.Length\_1", "Lobe.Width\_1"))  
  
str(stigma.new.flwrSizeOnly)

## 'data.frame': 89 obs. of 5 variables:  
## $ Corolla.Flare\_1: num 0.92 0.81 1.17 1.25 1.17 ...  
## $ Sepal.Length\_1 : num 0.93 1.26 0.98 1 1.13 ...  
## $ Tube.Length\_1 : num 1.53 1.72 1.45 1.39 1.62 ...  
## $ Lobe.Length\_1 : num 0.55 0.65 0.66 0.66 0.63 0.875 0.75 0.584 0.624 0.665 ...  
## $ Lobe.Width\_1 : num 0.65 0.44 0.6 0.59 0.55 0.83 0.507 0.654 0.66 0.592 ...

stigma.new.flwrSizeScaled <- scale(stigma.new.flwrSizeOnly)  
  
distance <- get\_dist(stigma.new.flwrSizeScaled)  
fviz\_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))



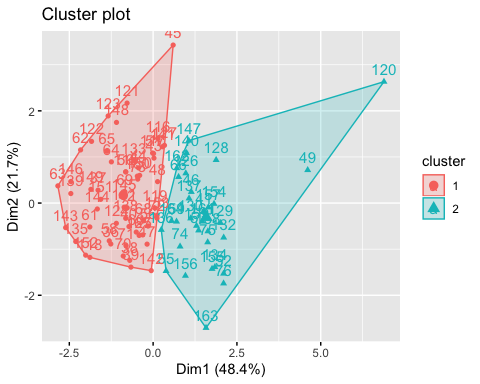
k2 <- kmeans(stigma.new.flwrSizeScaled, centers = 2, nstart = 25)  
str(k2)

## List of 9  
## $ cluster : Named int [1:89] 1 1 1 1 1 2 1 1 1 1 ...  
## ..- attr(\*, "names")= chr [1:89] "36" "37" "38" "39" ...  
## $ centers : num [1:2, 1:5] -0.5663 1.0087 -0.0238 0.0425 -0.2065 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:2] "1" "2"  
## .. ..$ : chr [1:5] "Corolla.Flare\_1" "Sepal.Length\_1" "Tube.Length\_1" "Lobe.Length\_1" ...  
## $ totss : num 440  
## $ withinss : num [1:2] 174 135  
## $ tot.withinss: num 309  
## $ betweenss : num 131  
## $ size : int [1:2] 57 32  
## $ iter : int 1  
## $ ifault : int 0  
## - attr(\*, "class")= chr "kmeans"

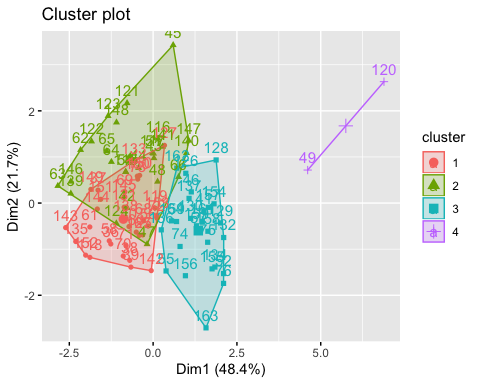
k2

## K-means clustering with 2 clusters of sizes 57, 32  
##   
## Cluster means:  
## Corolla.Flare\_1 Sepal.Length\_1 Tube.Length\_1 Lobe.Length\_1 Lobe.Width\_1  
## 1 -0.566267 -0.02384571 -0.2065211 -0.4865976 -0.4732365  
## 2 1.008663 0.04247516 0.3678657 0.8667520 0.8429526  
##   
## Clustering vector:  
## 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 55   
## 1 1 1 1 1 2 1 1 1 1 2 1 1 2 1 1 2 2   
## 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73   
## 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1   
## 74 75 76 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129   
## 2 2 2 1 1 1 1 1 2 1 1 1 1 1 2 1 2 2   
## 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147   
## 2 1 2 1 2 1 2 2 1 1 2 1 1 1 1 1 1 2   
## 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164   
## 1 1 1 1 1 1 2 2 2 2 2 2 1 2 2 2 2   
##   
## Within cluster sum of squares by cluster:  
## [1] 174.4549 134.8189  
## (between\_SS / total\_SS = 29.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

fviz\_cluster(k2, data = stigma.new.flwrSizeScaled)



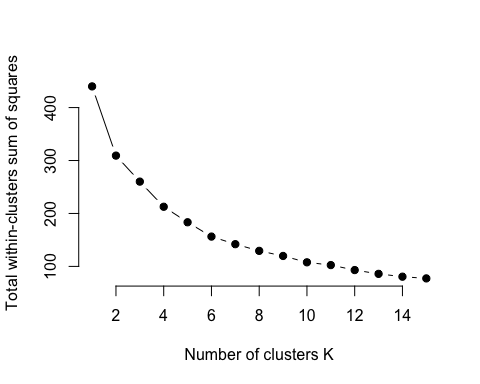
k2 <- kmeans(stigma.new.flwrSizeScaled, centers = 4, nstart = 25)  
fviz\_cluster(k2, data = stigma.new.flwrSizeScaled)



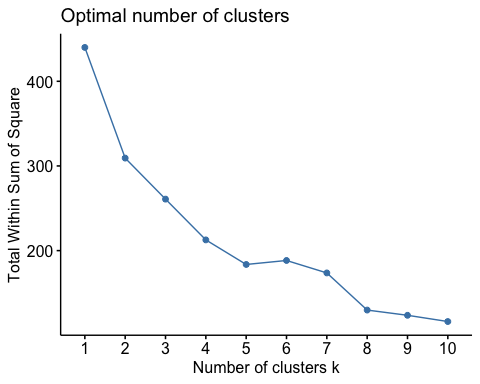
#### As I have to choose the number of cluster groups, I attempted to look at the sum of squares found for several values of k.

There was not a clear “elbow” here, demarcating which k value would give us the lowest sum of squares for the lowest k-value… so I chose to try another method (hierarchical clustering below).

wss <- function(k) {  
 kmeans(stigma.new.flwrSizeScaled, k, nstart = 10 )$tot.withinss  
}  
k.values <- 1:15  
wss\_values <- map\_dbl(k.values, wss)  
  
plot(k.values, wss\_values,  
 type="b", pch = 19, frame = FALSE,   
 xlab="Number of clusters K",  
 ylab="Total within-clusters sum of squares")



fviz\_nbclust(stigma.new.flwrSizeScaled, kmeans, method = "wss")



#### Hierarchical clustering method

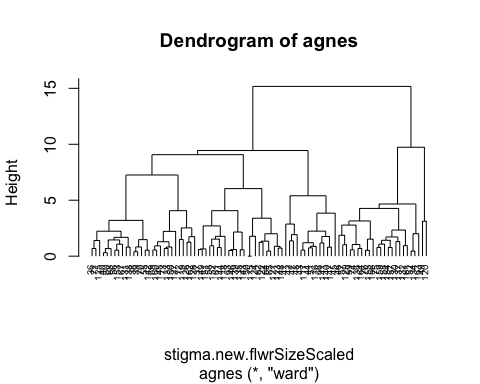
Reference website: <https://uc-r.github.io/hc_clustering>

Here, I compare different types of hierarchical clustering methods to establish which has the largest cluster structure value. The Ward method seems to work best, so I look at how many individuals are in each of the two largest clusters.

m <- c( "average", "single", "complete", "ward")  
names(m) <- c( "average", "single", "complete", "ward")  
  
# function to compute coefficient  
ac <- function(x) {  
 agnes(stigma.new.flwrSizeScaled, method = x)$ac  
}  
  
map\_dbl(m, ac)

## average single complete ward   
## 0.8450557 0.7209022 0.9029666 0.9355533

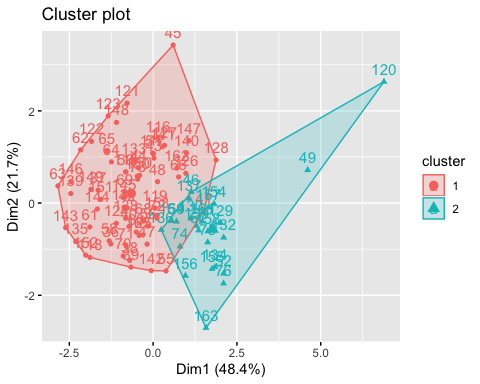
hc3 <- agnes(stigma.new.flwrSizeScaled, method = "ward")  
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of agnes")



# Cut tree into 2 groups  
sub\_grp <- cutree(hc3, k = 2)  
  
# Number of members in each cluster  
table(sub\_grp)

## sub\_grp  
## 1 2   
## 65 24

fviz\_cluster(list(data = stigma.new.flwrSizeScaled, cluster = sub\_grp))



#### Before I go any further, I wanted to discuss the following from the “Additional Comments” section in the hierarchical clustering tutorial:

"Clustering can be a very useful tool for data analysis in the unsupervised setting. However, there are a number of issues that arise in performing clustering. In the case of hierarchical clustering, we need to be concerned about:

What dissimilarity measure should be used? What type of linkage should be used? Where should we cut the dendrogram in order to obtain clusters? Each of these decisions can have a strong impact on the results obtained. In practice, we try several different choices, and look for the one with the most useful or interpretable solution. With these methods, there is no single right answer - any solution that exposes some interesting aspects of the data should be considered."

I looked at Euclidian distances (the automatic dissimilarity measure built into these functions) and a Ward linkage type.

I’m also not sure how to plot these clusters out while also showing whether the points are from sweet or skunky individuals. I’m thinking I could assign colors? Since I had to remove “Sweet” and “Skunky” from the stigma.new.FlwrSizeOnly dataset, I’m not sure how to go back and plot out these variables with the new dataset.