# Clustering Project

Ayomide Afolabi

11/25/2020

## INTRODUCTION

Water pollution is one of the biggest challenges worldwide.Tracing the source of pollutant helps to give an insight on how to solve this problem.In this project,hierarchical clustering method was used on sterols composition of animals fecal in creating a model that will be use for classifying source of pollutants.

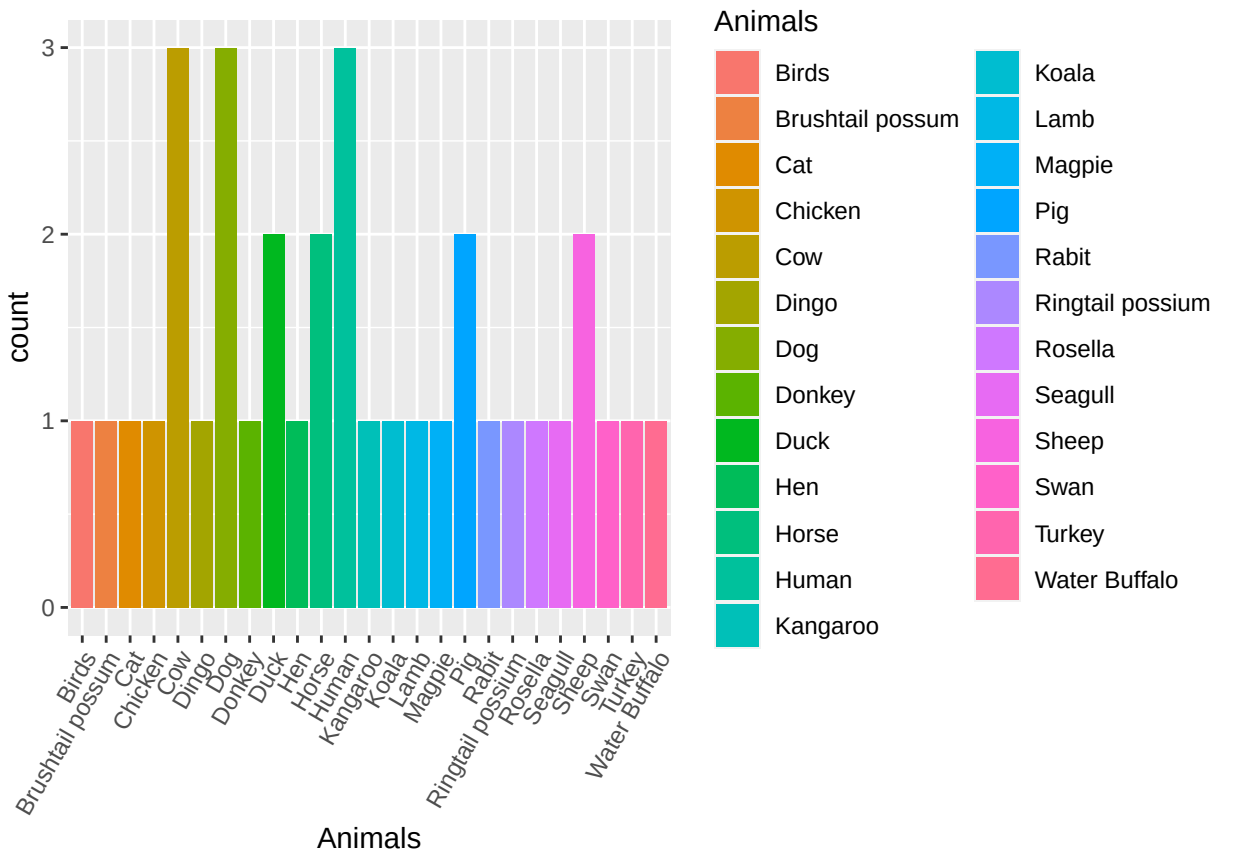
### loading data and data preprocessing

Data used for this project were sourced from different peer-reviewed journals.The right data were combined together and used in this project.

```
setwd("~/GitHub/GEO-SCIENCE-ML-MODEL")
Sterols <- read.csv(file="Sterols.csv",fileEncoding = "UTF-8-BOM")
Sterols1 <- Sterols
rownames(Sterols) <- make.names(Sterols[,1],unique=TRUE)
Sterols <- Sterols[,-1]
Sterols <- as.data.frame(Sterols)
Sterols[is.na(Sterols)] <- 0
Sterols1[is.na(Sterols1)] <- 0
Sterols <- scale(Sterols)
Sterols1$Animals <- as.factor(Sterols1$Animals)
Sterols1$Cholesterol <- as.numeric(Sterols1$Cholesterol)
```

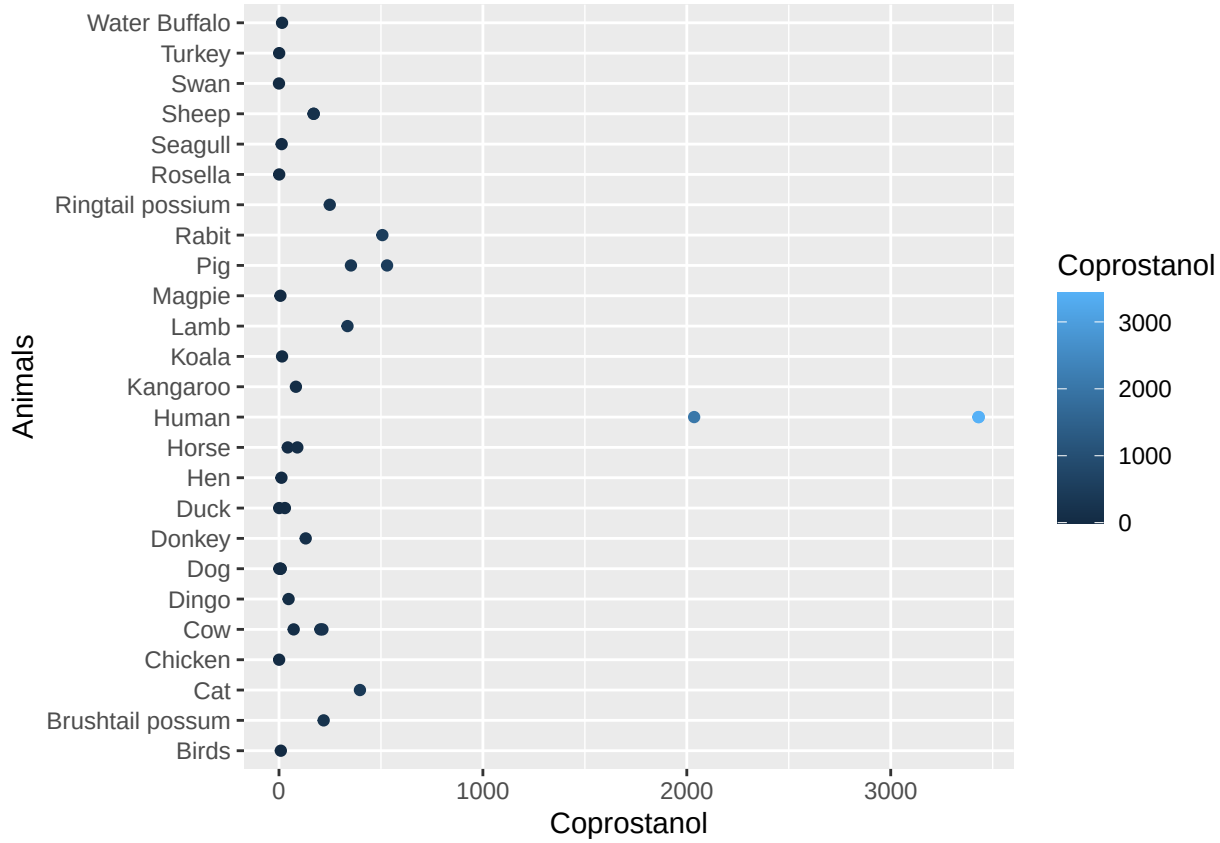
## Exploration Data Analysis

```
# frequency distribution of the observation
library(ggplot2)
ggplot(data=Sterols1)+
  geom_bar(mapping=aes(x=Animals,fill=Animals))+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



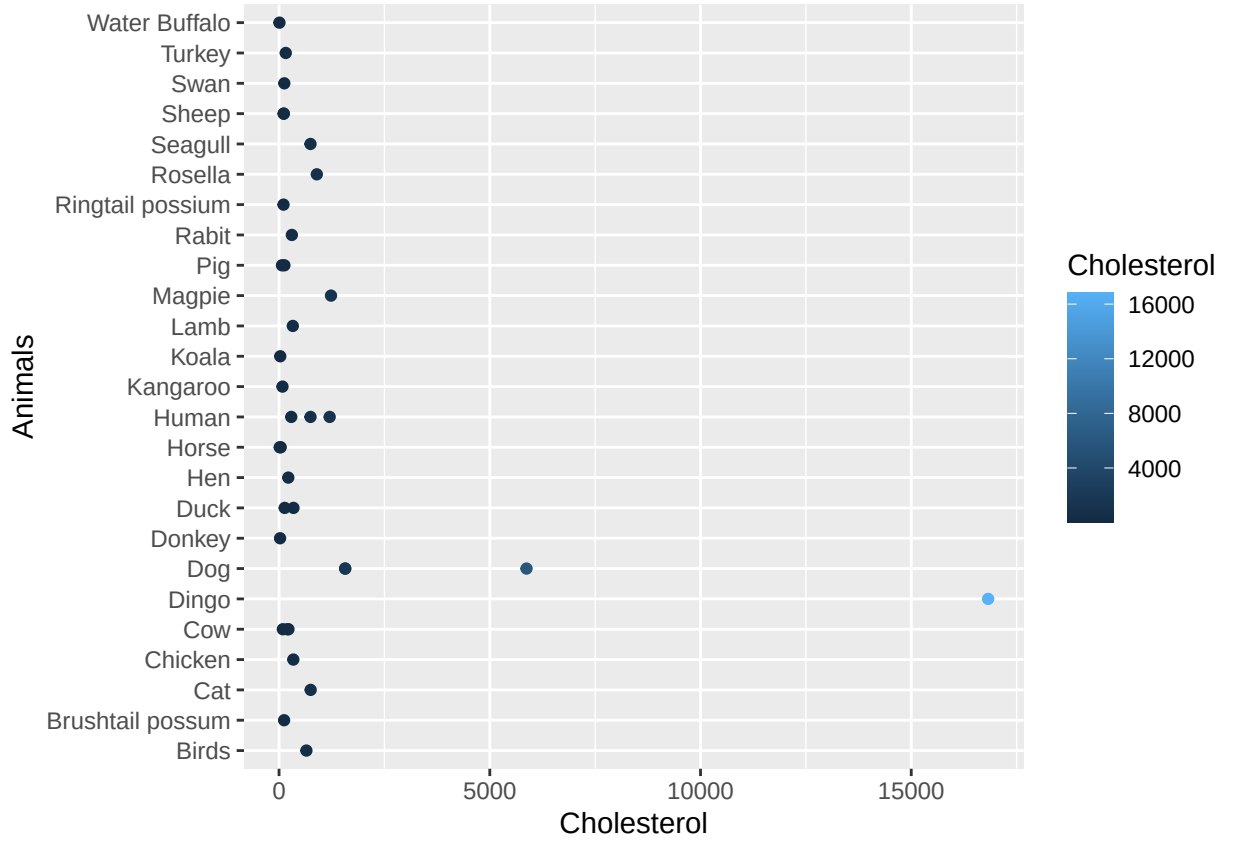

The bar chart distribution shows sterols concentration coming from 3 cows ,3 Humans, 3 dogs, 2 ducks, 2 horse, 2 pigs, 2 sheep and others animals

```
#Coprostanol concentration is high in humans
ggplot(data=Sterols1,mapping=aes(x=Coprostanol,y=Animals))+
  geom_point(aes(colour=Coprostanol))
```



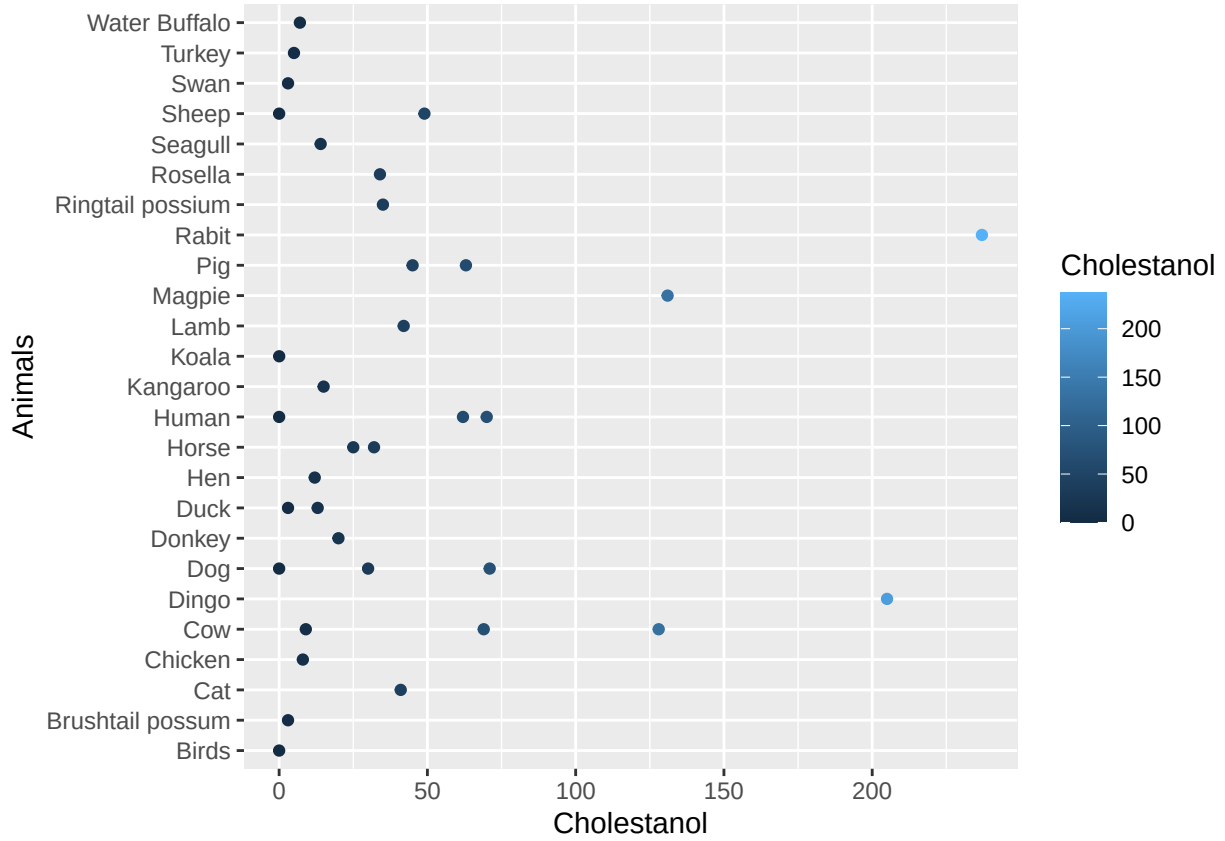

Humans have the highest concentration of coprostanol concentration in their fecal as shown in the diagram compared to other animals

```r
#Cholesterol concentration is high in humans and dogs
ggplot(data=Sterols1,mapping=aes(x=Cholesterol,y=Animals))+
  geom_point(aes(colour=Cholesterol))
```



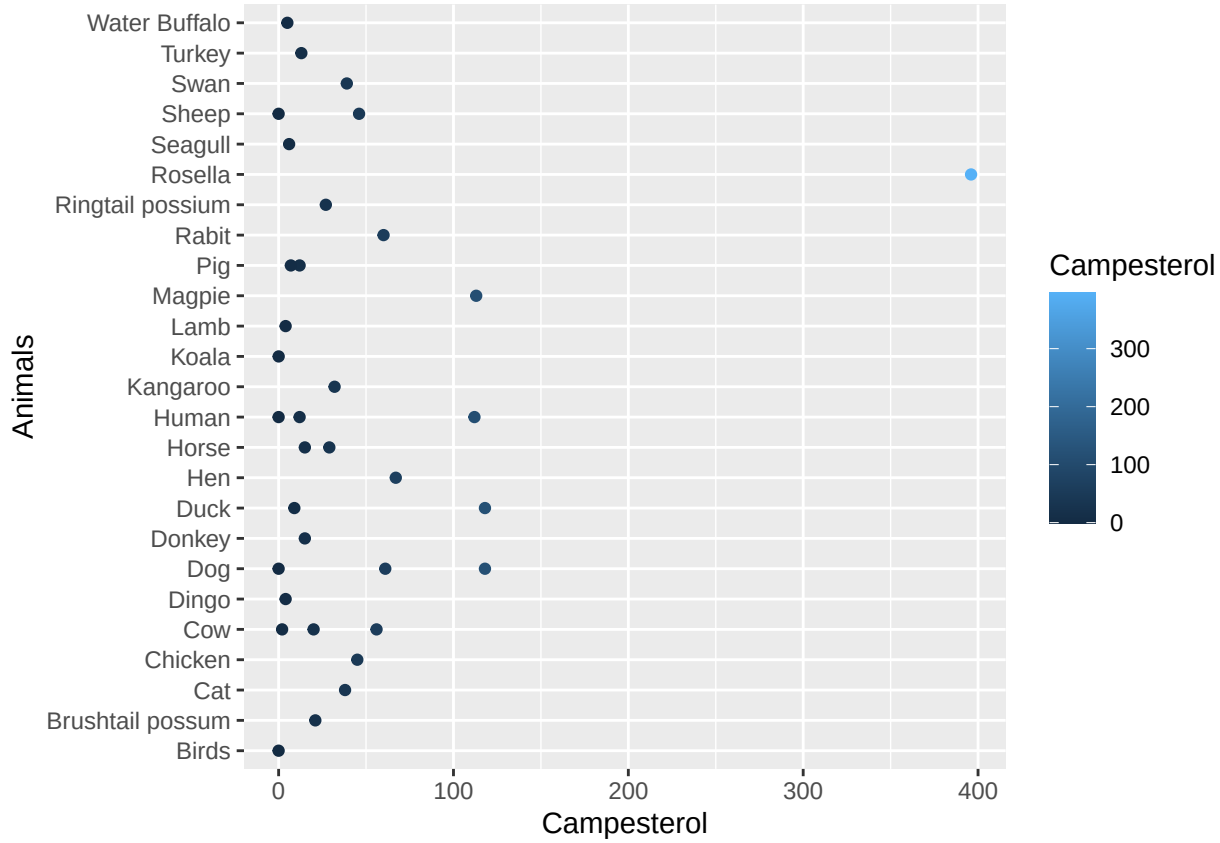

Humans and dogs have a very high concentration of Cholesterol concentration in their fecal as shown in the diagram compared to other animals

```
#Cholestanol concentration is high in Rabbit,Dingo,Magpie
ggplot(data=Sterols1,mapping=aes(x=Cholestanol,y=Animals))+
  geom_point(aes(colour=Cholestanol))
```



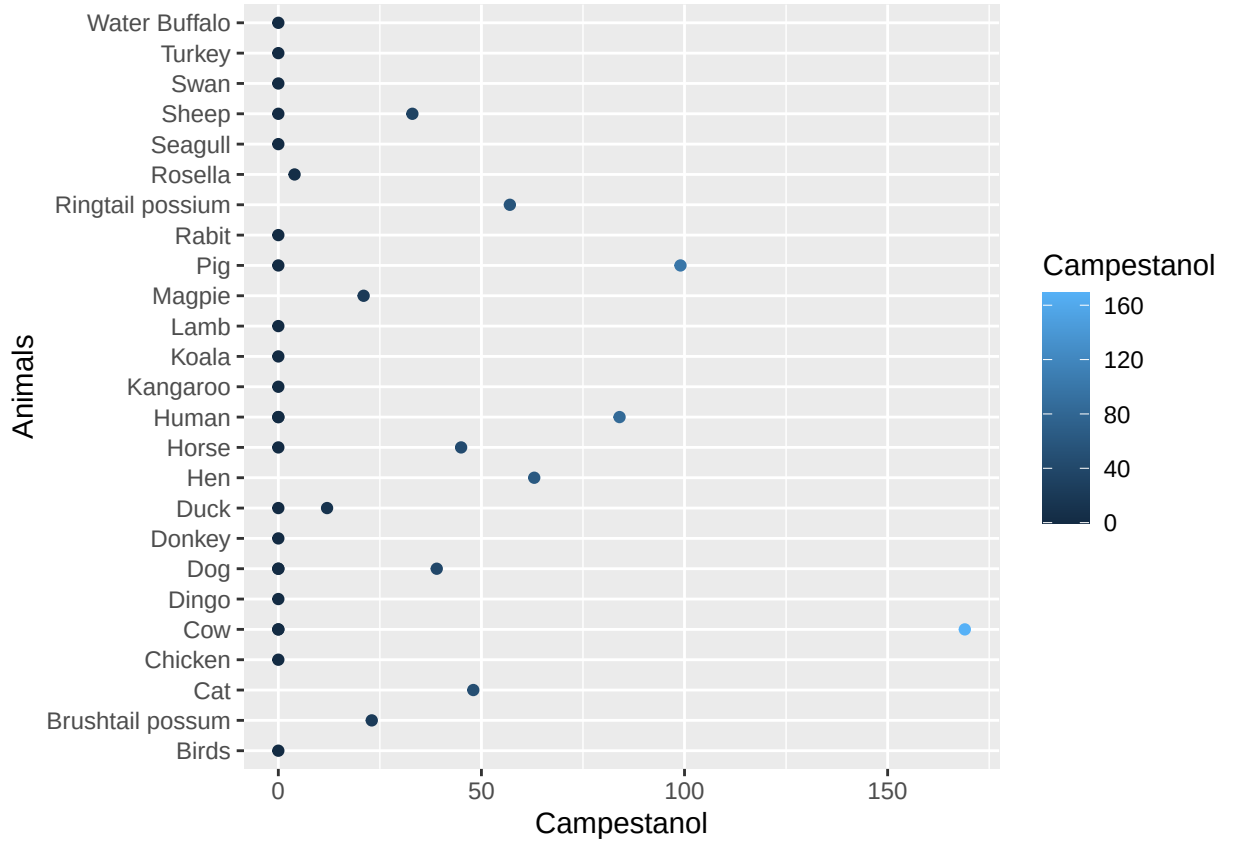

Rabbit,Dingo and Magpie have a very high concentration of Cholestenol concentration in their fecal as shown in the diagram

```
#Campesterol concentration is high in Rosella
ggplot(data=Sterols1,mapping=aes(x=Campesterol,y=Animals))+
  geom_point(aes(colour=Campesterol))
```

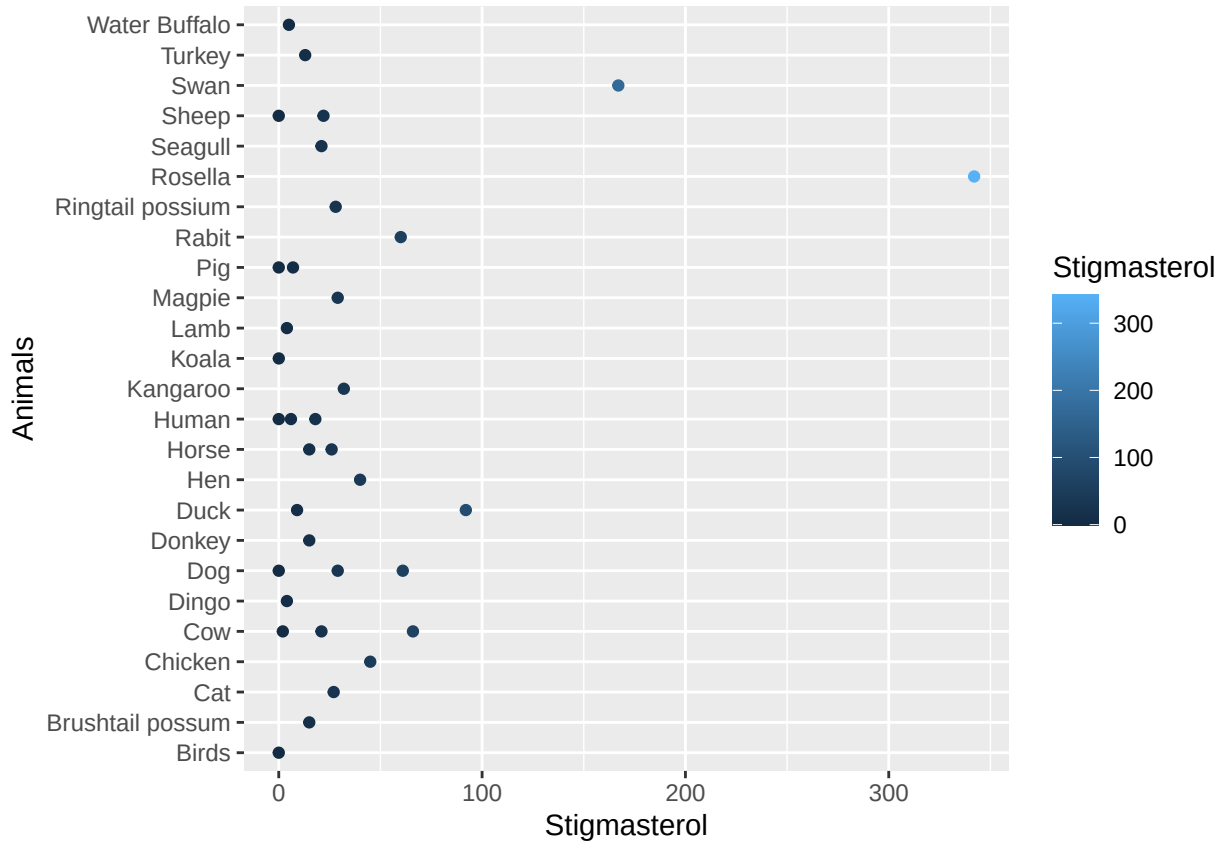



Rosella has the highest concentration of Campesterol concentration in their fecal as shown in the diagram compared to other animals

```
#Campestanol
ggplot(data=Sterols1,mapping=aes(x=Campestanol,y=Animals))+
  geom_point(aes(colour=Campestanol))
```

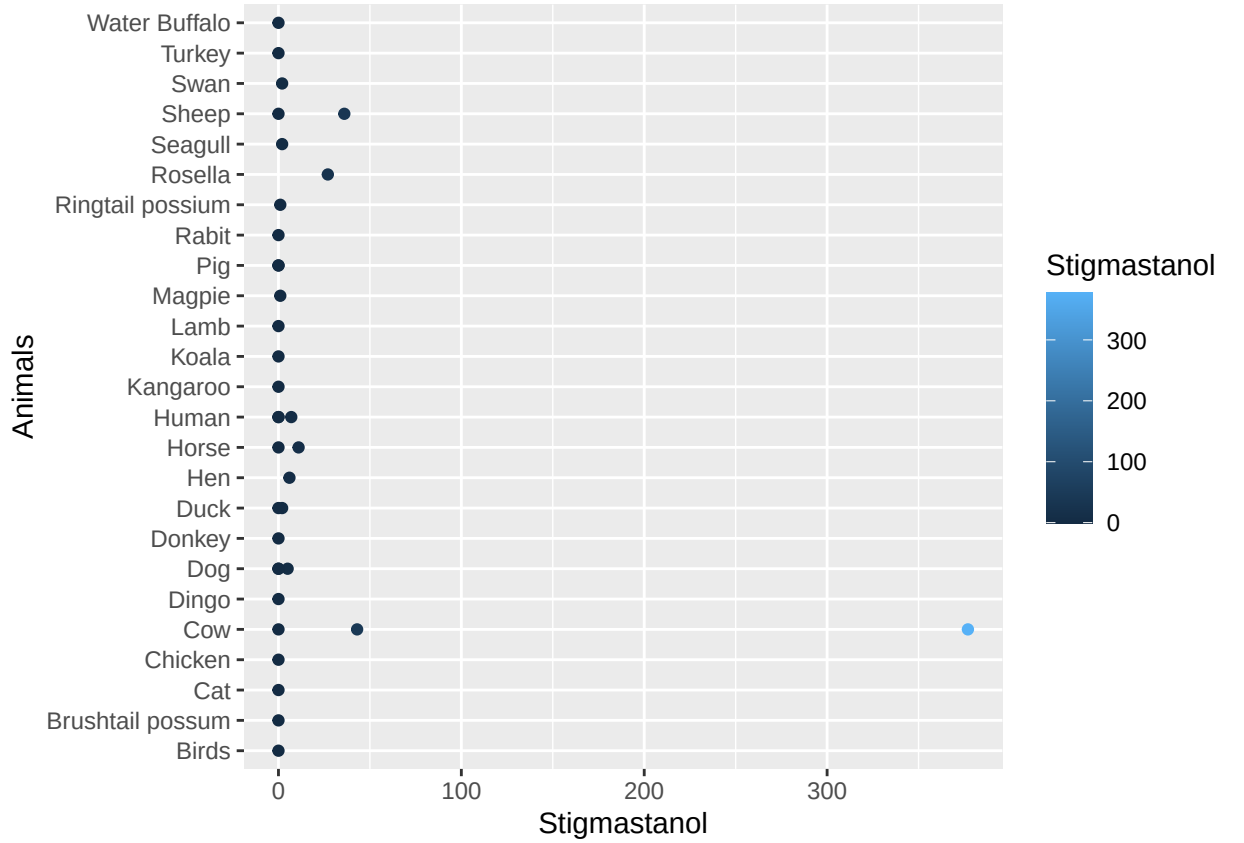



```
#Stigmasterol concentration is high in swan, rosella
ggplot(data=Sterols1,mapping=aes(x=Stigmasterol,y=Animals))+
  geom_point(aes(colour=Stigmasterol))
```



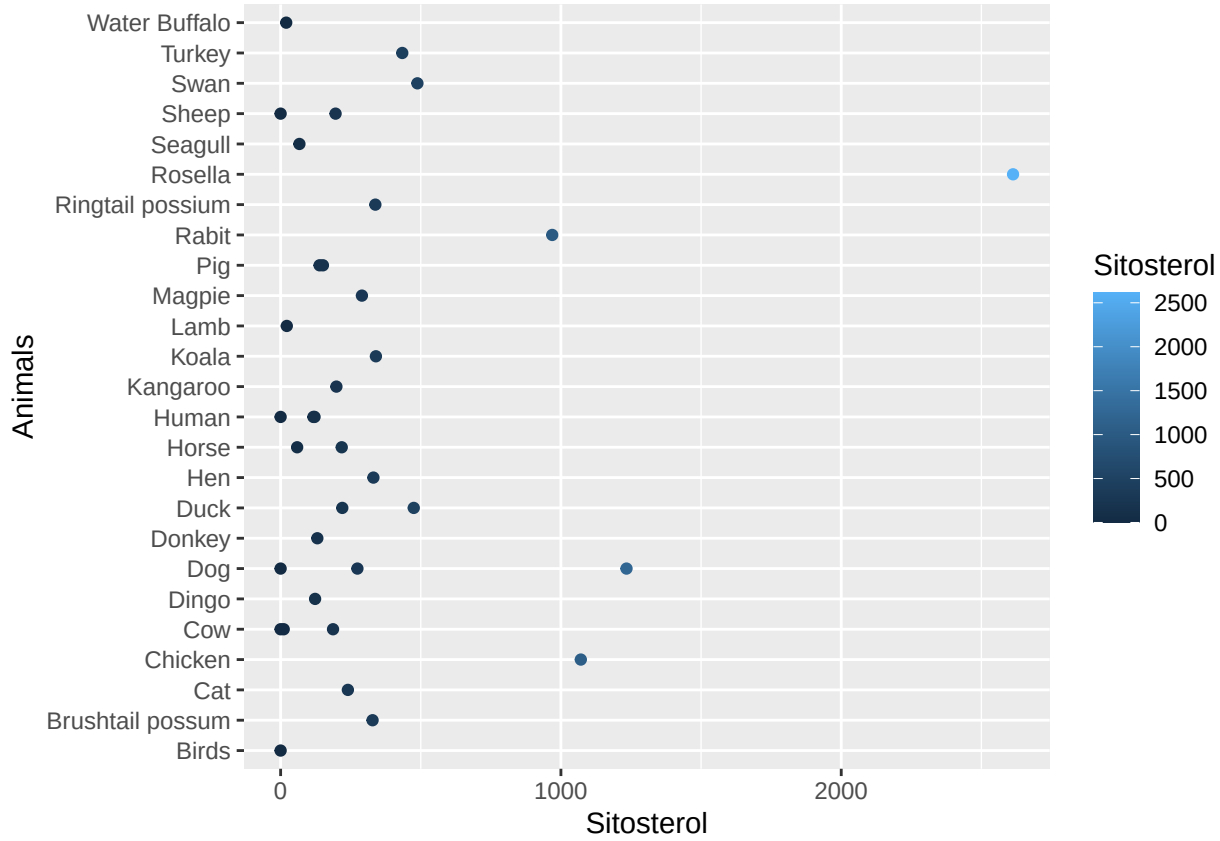

Swan and Magpie have a very high concentration of Stigmasterol concentration in their fecal as shown in the diagram compared to other animals

```
#Stigmastanol concentration is high in humans
ggplot(data=Sterols1,mapping=aes(x=Stigmastanol,y=Animals))+
  geom_point(aes(colour=Stigmastanol))
```



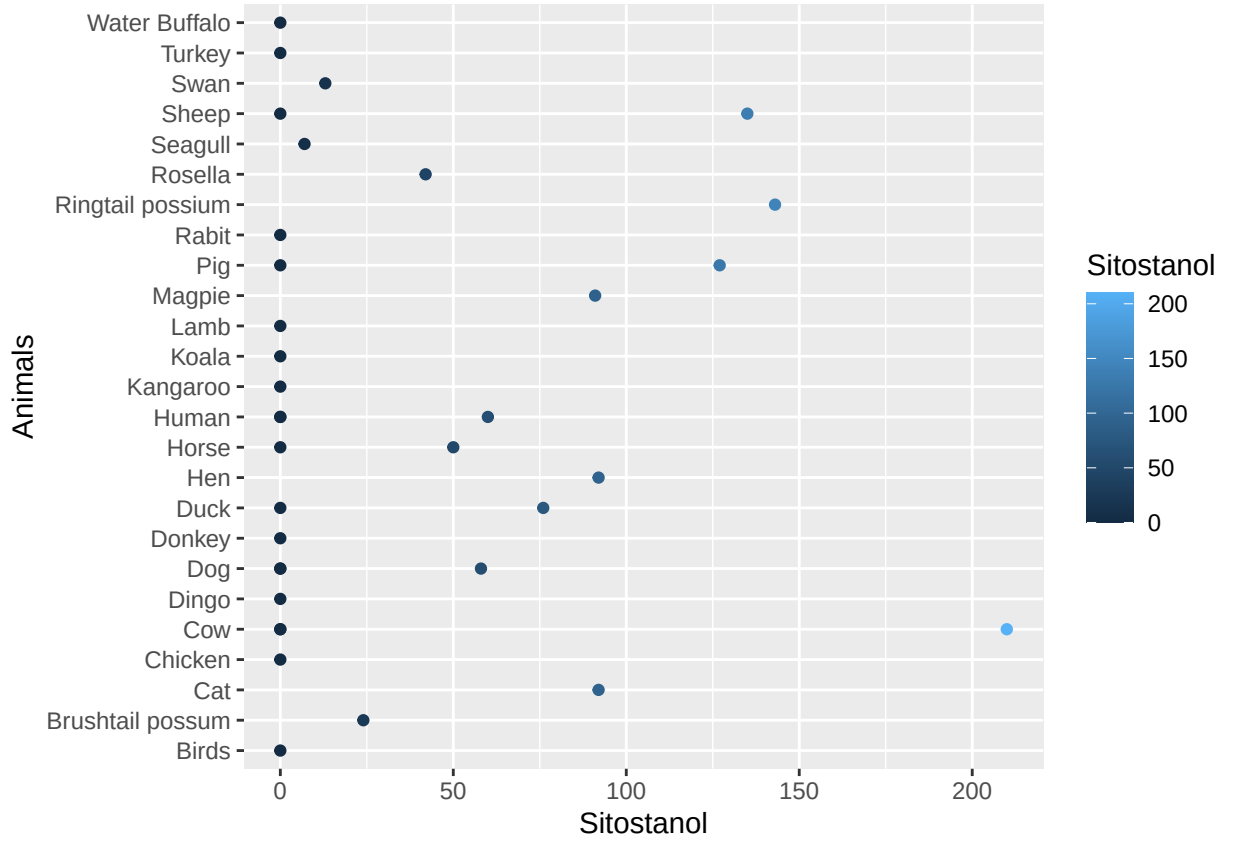

Humans have the highest concentration of stigmastanol concentration in their fecal as shown in the diagram compared to other animals

```
#Sitosterol concentration is high in Rosella,Chicken
ggplot(data=Sterols1,mapping=aes(x=Sitosterol,y=Animals))+
  geom_point(aes(colour=Sitosterol))
```




Rosella and Chicken have a very high concentration of Sitosterol concentration in their fecal as shown in the diagram when compared to other animals

```
#Sitosterol concentration is high in Ringtail possium
ggplot(data=Sterols1,mapping=aes(x=Sitostanol,y=Animals))+
  geom_point(aes(colour=Sitostanol))
```



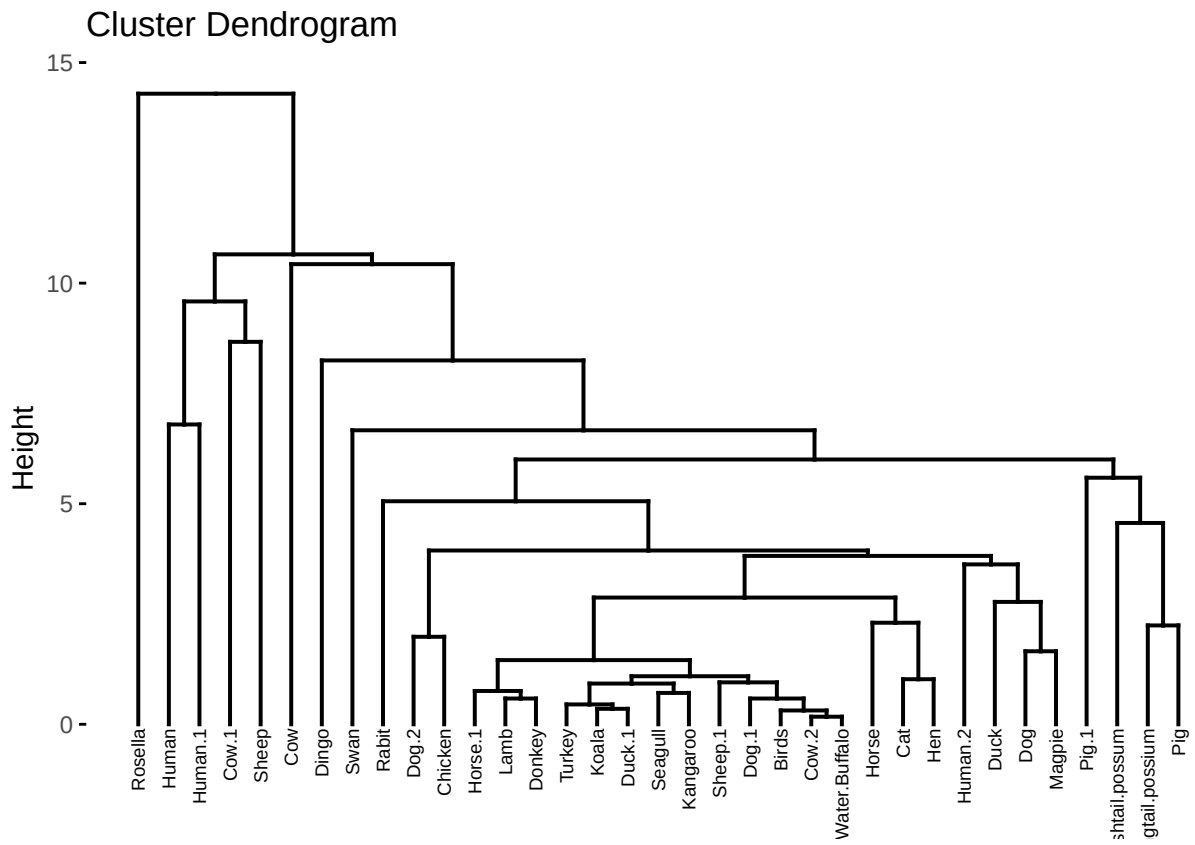

Ringtail possium have the highest concentration of Sitostanol concentration in their fecal as shown in the diagram compared to other animals

## Clustering Analysis

```
# The similarity measures help to indicate clusters that should be combine or not. This is usually comp
data <- Sterols
euclidean_distance <- dist(data, method="euclidean")
#as.matrix(euclidean_distance)[1:35,1:18]
res.hc <- hclust(d=euclidean_distance,method="complete")
library("factoextra")
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

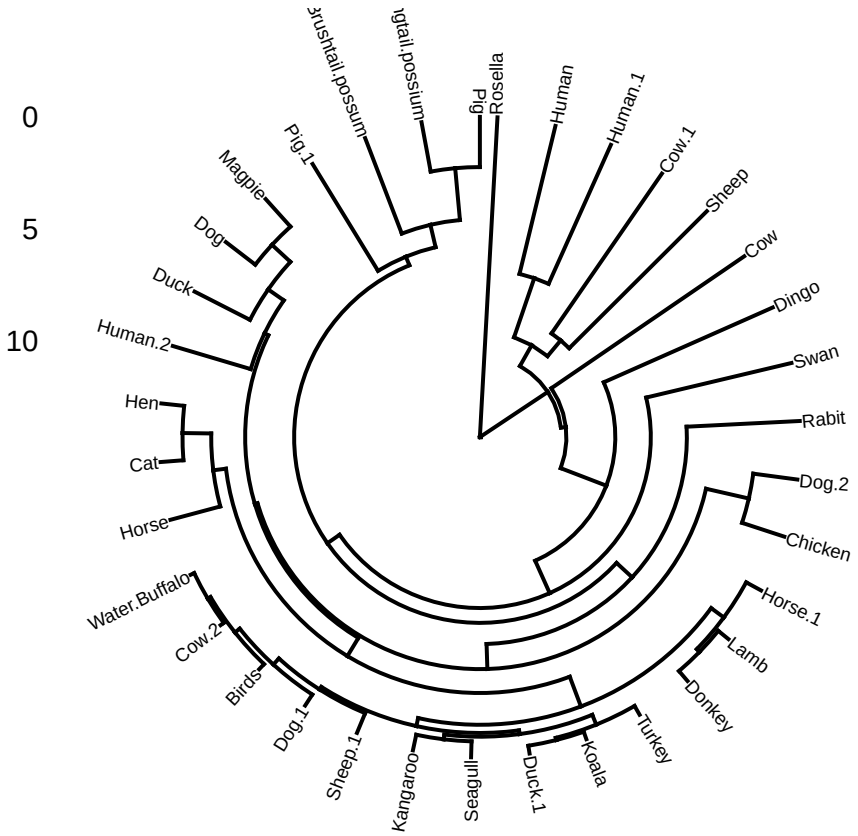
```
fviz_dend(res.hc,
cex = 0.5, # label size
k_colors = "jco",
color_labels_by_k = TRUE, # color labels by groups
rect = TRUE,# Add rectangle around groups
rect_border="jco",rect_fill=TRUE
)
```




The cluster dendogram shown below has some deviation, especially one of the human fecal samples not falling in the right cluster. Also, three cow fecal samples falls under different cluster.

The cluster diagram can also be view from other angles as shown below

```
fviz_dend(res.hc, cex = 0.5, color_labels_by_k = TRUE,
k_colors = "jco", type = "circular")
```




```
library(igraph)
```

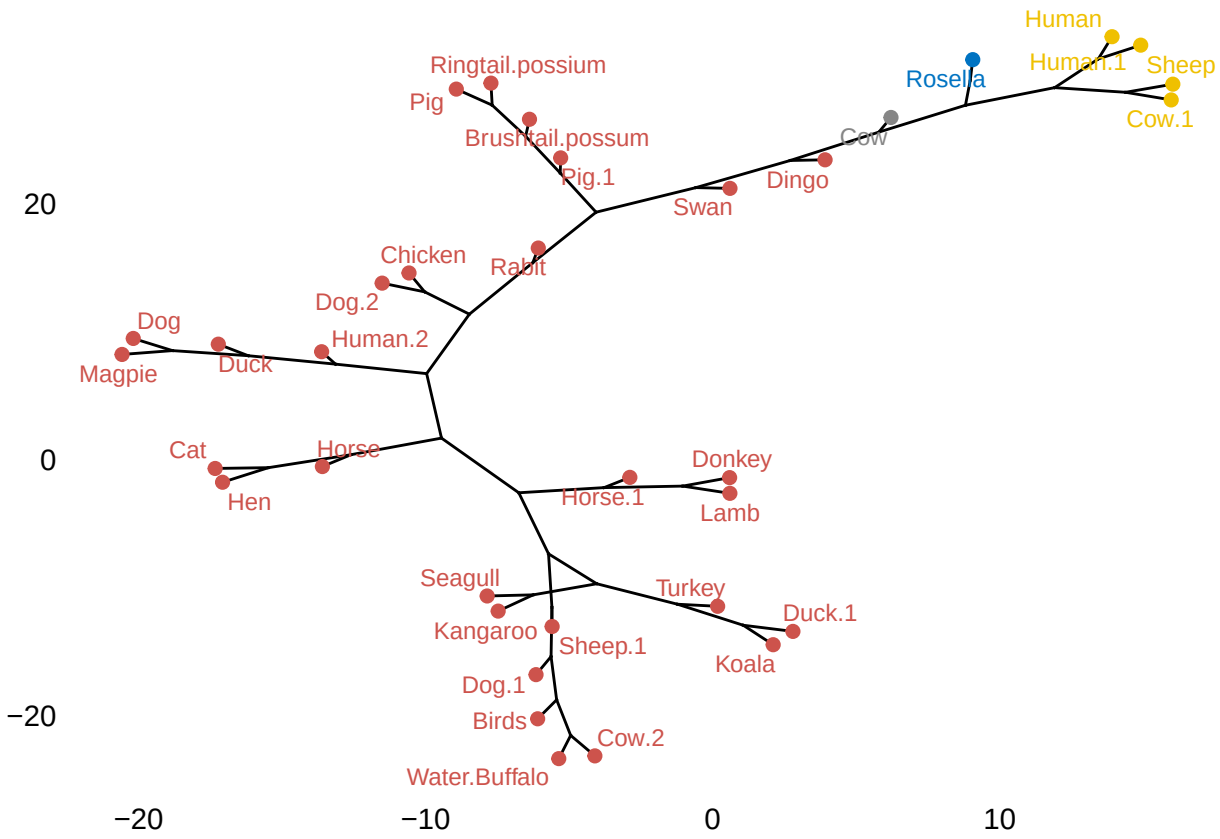
```
## Warning: package 'igraph' was built under R version 4.0.3
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

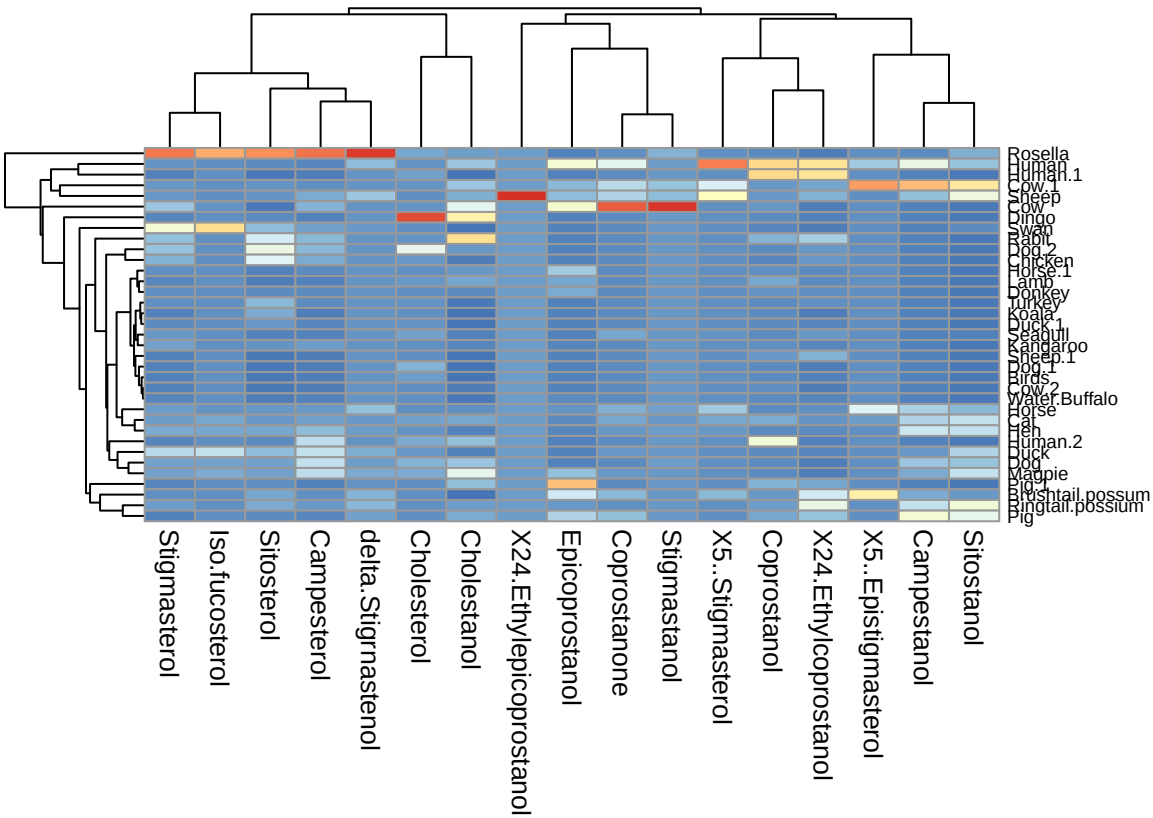
```
require("igraph")
fviz_dend(res.hc, k = 4,color_labels_by_k = TRUE, k_colors = "jco",
type = "phylogenic", repel = TRUE)
```



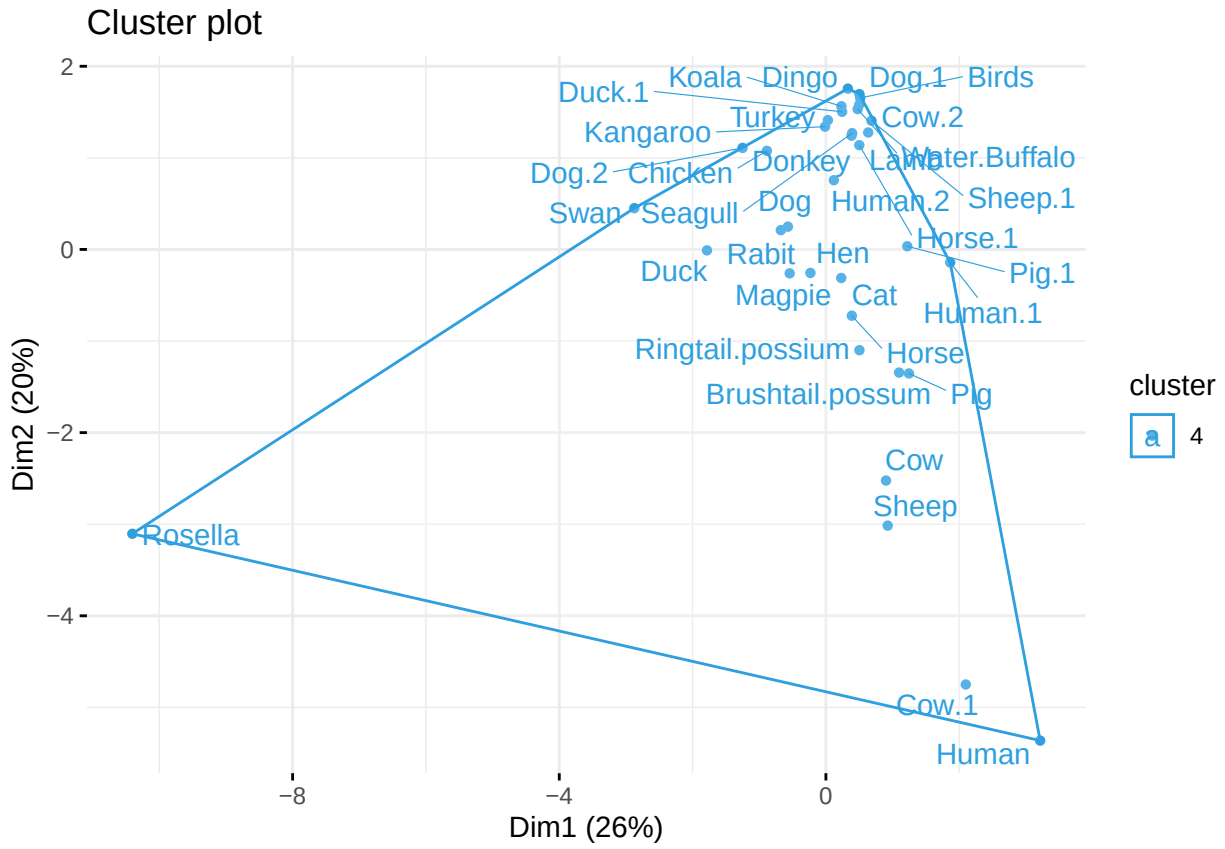

```
library("pheatmap")
```

```
## Warning: package 'pheatmap' was built under R version 4.0.3
```

```
pheatmap(data,fontsize_col =10,fontsize_row=7,cellheight = 4 )
```


Rosella
Human
Human.1
Cow.1
Sheep
Cow
Dingo
Swan
Rabbit
Dog.2
Chicken
Horse.1
Lamb
Donkey
Turkey
Koala
Duck.1
Seagull
Kangaroo
Sheep.1
Dog.1
Birds
Cow.2
Water.Buffalo
Horse
Cat
Hen
Human.2
Duck
Dog
Magpie
Pig.1
Brushtail.possum
Ringtail.possum
Pig
5
4
3
2
1
0
Stigmasterol
Iso.fucosterol
Sitosterol
Campesterol
delta.Stigmastenol
Cholesterol
Cholestanol
X24.Ethylepicoprostanol
Epicoprostanol
Coprostanone
Stigmastanol
X5..Stigmasterol
Coprostanol
X24.Ethylcoprostanol
X5..Epistigmasterol
Campestanol
Sitostanol

```
fviz_cluster(list(data = data,cluster = 4),
palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
ellipse.type = "convex", # Concentration ellipse
repel = TRUE, # Avoid label overplotting (slow)
show.clust.cent = FALSE, ggtheme = theme_minimal())
```




```
# compute cophentic distance
res.coph <- cophenetic(res.hc)
cor(euclidean_distance,res.coph)
```

```
## [1] 0.9515707
```

## Cluster Validation

It is necessary to evaluate whether the dataset contains meaningful clusters or not, before applying a clustering method on that particular dataset. This evaluation can simply be done by computing the clustering tendency. The hopkins statistic is used to assess the clustering tendency of a dataset by measuring the probability that a given dataset is generated by a uniform distribution.
If the probability is low,that means the dataset has a high clustering tendency. A threshold of 0.5 is used in most cases.

```
library(clustertend)
set.seed(123)
hopkins(data, n = nrow(data)-1)
```

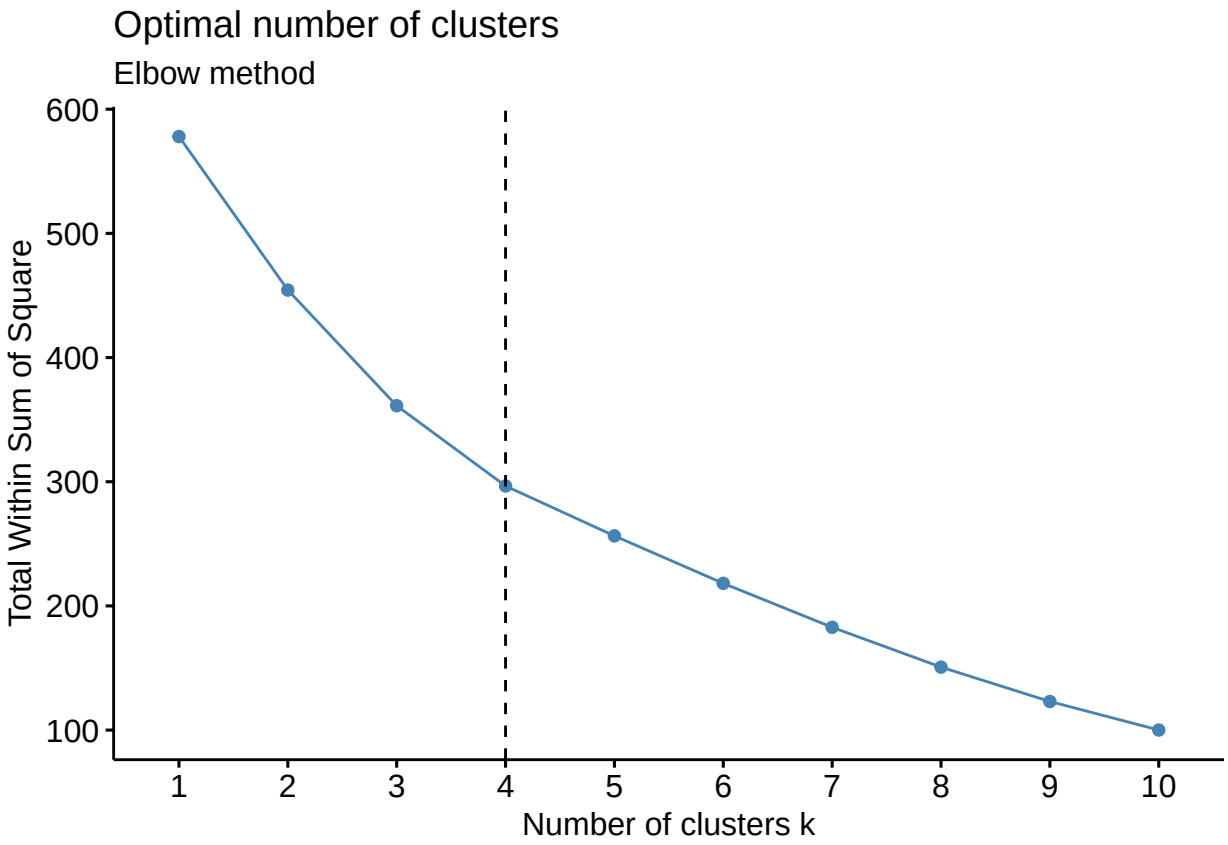
```
## $H
## [1] 0.1835349
```

The hopkins statistics computed is 0.193532794077259.This value is below the threshold of 0.5, which shows the dataset is highly clusterable

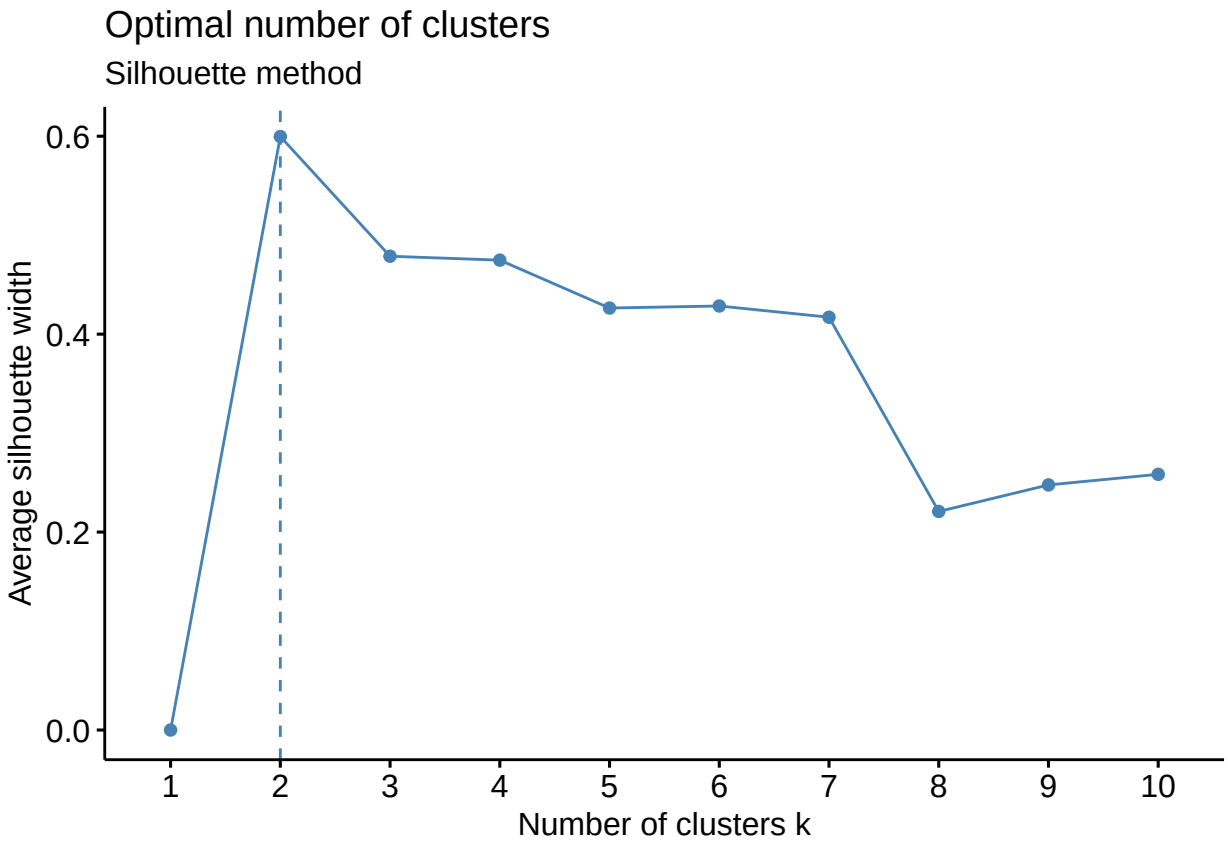
## Optimal Cluster

The optimal cluster to be used for clustering, can be done using different methods. This include elbow method, silhouette method and gap statistic method. The elbow method was used in this case. Using an optimal value helps to reduce error associated with misclassification.

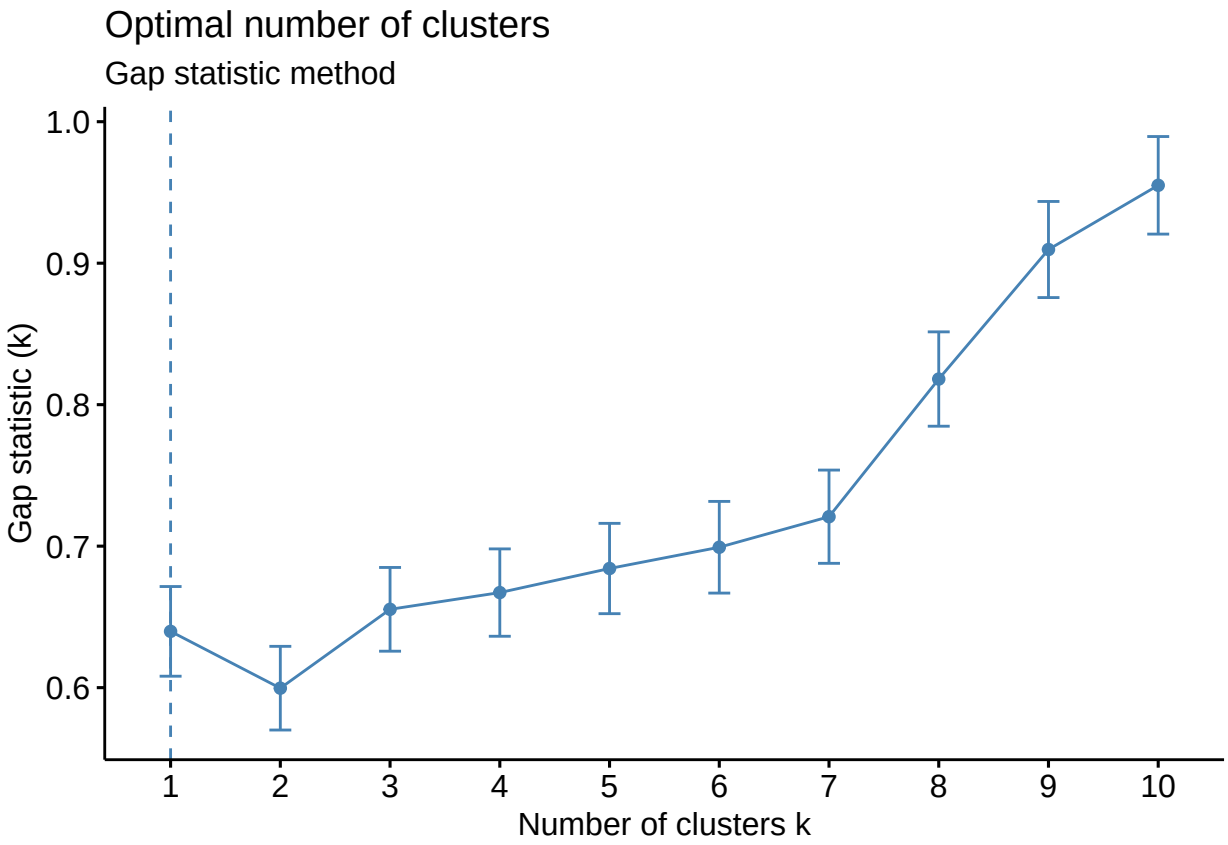
```
library(NbClust)
# Elbow method
fviz_nbclust(Sterols, hcut, method = "wss") +
geom_vline(xintercept = 4, linetype = 2)+
labs(subtitle = "Elbow method")
```




```
# Silhouette method
fviz_nbclust(Sterols, hcut, method = "silhouette")+
labs(subtitle = "Silhouette method")
```




```
set.seed(123)
fviz_nbclust(Sterols, hcut, nstart = 25, method = "gap_stat", nboot = 500)+
labs(subtitle = "Gap statistic method")
```




The elbow method shows that the the optimal clusters should be four. This is finally used in generation the final cluster dendogram.

### Final model

```
fviz_dend(res.hc,  # Cut in four groups
cex = 0.5, # label size
k_colors = "jco",k=4,
color_labels_by_k = TRUE, # color labels by groups
rect = TRUE,# Add rectangle around groups
rect_border="jco",rect_fill=TRUE
)
```