**Task 1 (Group Task) — Multivariate Statistical Analysis**

**Question 1:**

Use R to carry out Principal Component Analysis (PCA) on the dataset using only the five measurements for sparrows. Produce and interpret relevant plots such as screeplot, biplot, and loadings plot. Also, produce and interpret a biplot using PC2 and PC3 as the axes.

**Code & Outputs:**

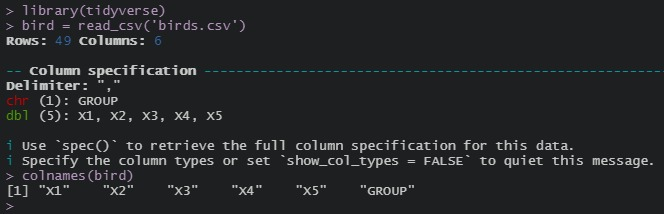
#######group task qs 1####

# install tidyverse and ggfortify packages

library(tidyverse)

bird = read\_csv('birds.csv')

colnames(bird)



* Firstly, please sort the rows of the dataset, first by *died/survival* and then *total length*.

# Arrange rows by died/survival and then total length

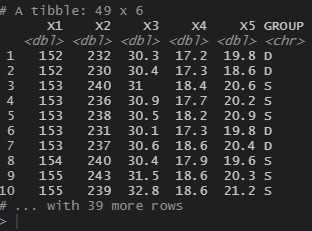
bird = bird %>%

group\_by(GROUP) %>%

arrange(X1) %>%

ungroup()

bird



#Task 1 , question 1

#PCA using 5 measurements for sparrows

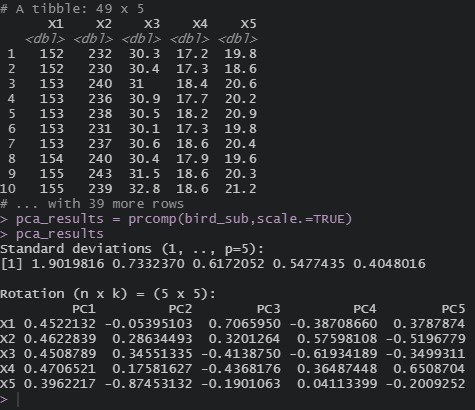
**# PCA**

bird\_sub = select(bird,-GROUP)

bird\_sub

pca\_results = prcomp(bird\_sub,scale.=TRUE)

pca\_results



#Biplot

library(ggfortify)

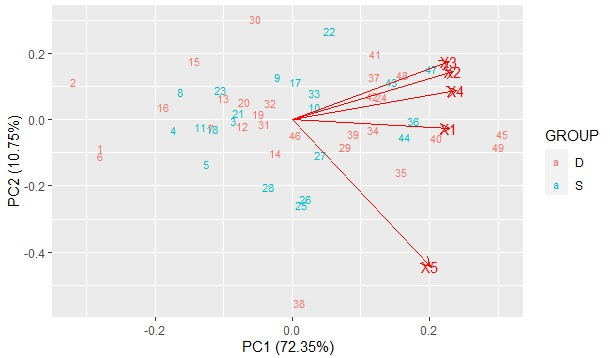
autoplot(pca\_results,

label=TRUE,label.size=**3**,shape=FALSE,

loadings=TRUE,loadings.label=TRUE,

data=bird,colour='GROUP')

#Note: PC1 (72.35%) + PC2 (10.75%) gives 83% of variance



**#Scree plot**

var\_explained = **100**\*((pca\_results$sd)^**2**)/(sum((pca\_results$sd)^**2**))

var\_explained

cumsum(var\_explained)

ggplot(NULL,aes(x=**1**:**5**,y=var\_explained)) +

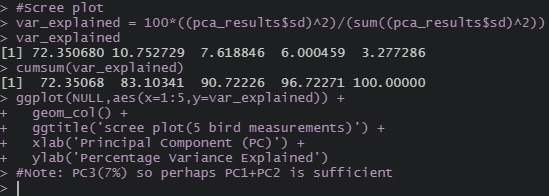
geom\_col() +

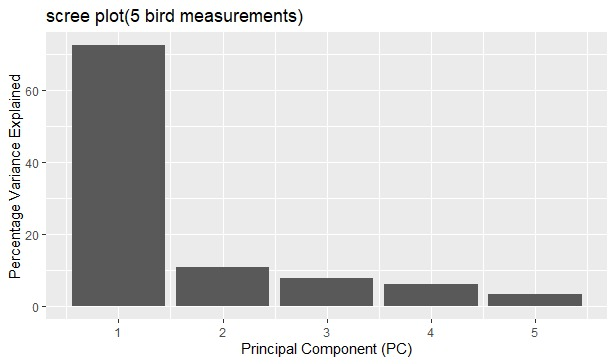
ggtitle('scree plot(5 bird measurements)') +

xlab('Principal Component (PC)') +

ylab('Percentage Variance Explained')

#Note: PC3(7%) so perhaps PC1+PC2 is sufficient



****

**#loading plot**

loadings = **as**.data.frame(pca\_results$rotation[,**1**:**5**])

loadings$symbol = row.names(loadings)

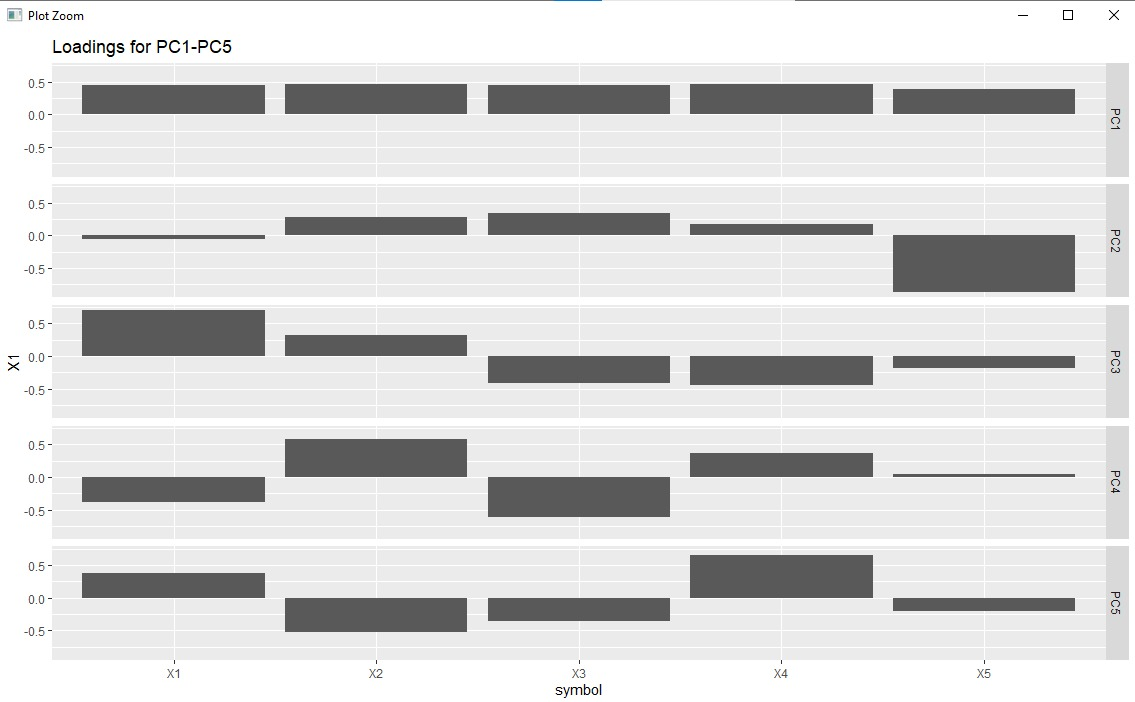
loadings = gather(loadings,key='component',value = 'X1',-symbol)

ggplot(loadings,aes(x=symbol,y=X1))+

geom\_bar(stat = 'identity') +

facet\_grid(component~.) +

ggtitle('Loadings for PC1-PC5')



**#Biplot using PC2 and PC3**

autoplot(pca\_results,x=**2**,y=**3**,

label=TRUE,label.size=**3**,shape=FALSE,

loadings=TRUE,loadings.label=TRUE,

data=bird,colour='GROUP')

#Note: PC2 (10.75%) + PC3 (7.62%)



**Question 2:**

Use R to carry out Cluster Analysis on the dataset using hierarchical clustering on the five measurements for sparrows. Cluster the sparrows and the five measurements separately. Compare results using different distance metrics (Manhattan, Euclidean, etc) and hierarchical clustering methods (single linkage, Ward’s method, etc). Produce, interpret, and evaluate a small selection of relevant plots, including dendrograms.

**Clustering:**

Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal.

There are different clustering methods to analyse the data. Here we use hierarchical clustering to cluster the sparrows (Died/Survival) and the five measurements.

**Hierarchical clustering:**

* Hierarchical Clustering groups (Agglomerative or also called as Bottom-Up Approach) or divides (Divisive or also called as Top-Down Approach) the clusters based on the distance metrics.
* In Agglomerative clustering, each data point acts as a cluster initially, and then it groups the clusters one by one.
* Divisive is the opposite of Agglomerative, it starts off with all the points into one cluster and divides them to create more clusters. These algorithms create a distance matrix of all the existing clusters and perform the linkage between the clusters depending on the criteria of the linkage.
* The clustering of the data points is represented by using a dendrogram. There are different types of linkages: –
* Single Linkage: – In single linkage the distance between the two clusters is the shortest distance between points in those two clusters.
* Complete Linkage: – In complete linkage, the distance between the two clusters is the farthest distance between points in those two clusters.
* Average Linkage: – In average linkage the distance between the two clusters is the average distance of every point in the cluster with every point in another cluster.
* **Ward’s method** : – (Minimum variance method) is an alternative to single-linkage [clustering](https://www.statisticshowto.com/clustering/). Popular in fields like linguistics, it’s liked because it usually creates compact, even-sized clusters.Like most other clustering methods, Ward’s method is computationally intensive. However, Ward’s has significantly fewer computations than other methods. The drawback is this usually results in less than optimal clusters. That said, the resulting clusters are usually good enough for most purposes.

**Distance Metrics:**

Distance metrics are a key part of several [machine learning algorithms](https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?utm_source=blog&utm_medium=4-types-of-distance-metrics-in-machine-learning). These distance metrics are used in both supervised and unsupervised learning, generally to calculate the similarity between data points.

An effective distance metric improves the performance of our machine learning model, whether that’s for classification tasks or clustering.

Four Types of Distance Metrics in Machine Learning

1. Euclidean Distance
2. Manhattan Distance
3. Makowski Distance
4. Hamming Distance

Here we use the most commonly used distance metrics – Euclidean and Manhattan.

* Euclidean Distance: Euclidean Distance represents the shortest distance between two points. Most machine learning algorithms including K-Means use this distance metric to measure the similarity between two data points.
* Manhattan Distance: Manhattan Distance is the sum of absolute differences between points across all the dimensions.
* Manhattan distance captures the distance between two points by **aggregating the pairwise absolute difference**between each variable while Euclidean distance captures the same by **aggregating the squared difference**in each variable.
* Therefore,**if two points are close on most variables, but more discrepant on one of them, Euclidean distance will exaggerate that discrepancy, whereas Manhattan distance will shrug it off, being more influenced by the closeness of the other variables.**

**Code & Output & Plots:**

**# #### QS 2################**

# install Cluster package

# start - Euclidean ,ward method -- D = 0.93

library(cluster)

BBB = scale(bird\_sub)

D = dist(BBB,method="euclidean")

cluster\_results = agnes(D,method="ward")

plot(cluster\_results,which.plots=**2**,

main='cluster the sparrows using 5 measurements(Euclidean/ward)')

rect.hclust(cluster\_results, k=**5**, border=**3**)

plot(**as**.dendrogram(cluster\_results),horiz=FALSE)

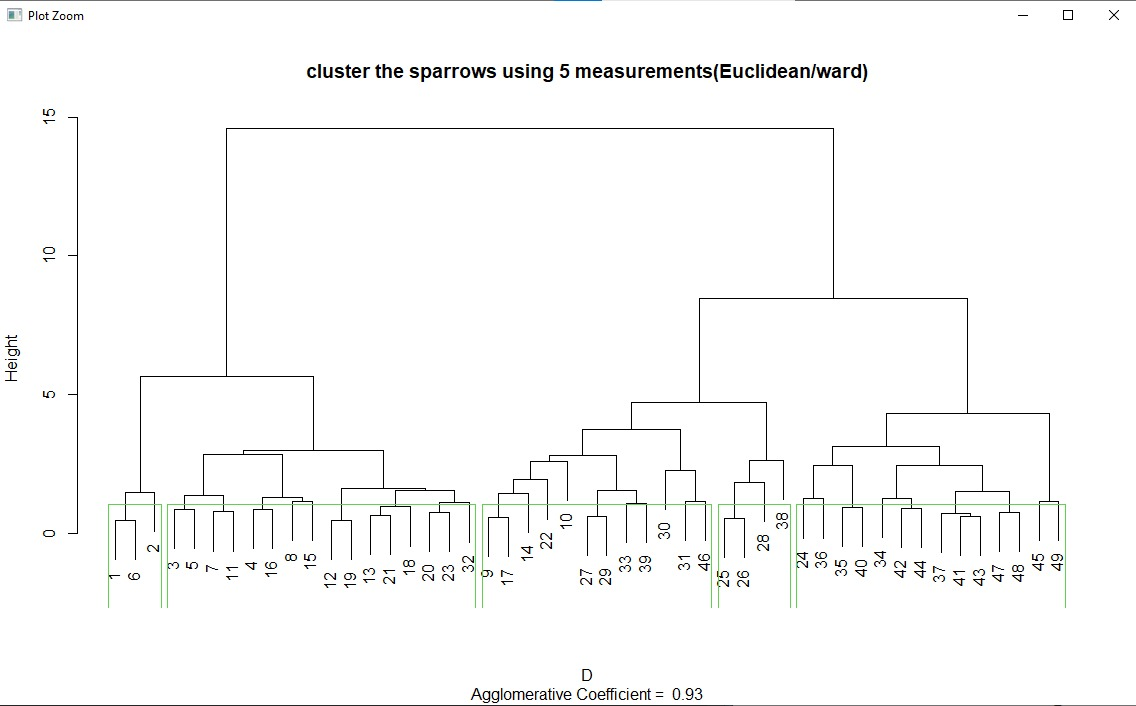
clusters = cutree(cluster\_results,k=**5**)

#1) compare results using different methods (Manhattan,Euclidean etc) and -

# -hierarchical clustering methods(single linkage,ward's method etc)')

# 2)Produce, interpret, and evaluate a -

# -small selection of relevant plots, including dendrograms



**#Euclidean ,single method -- D = 0.57**

library(cluster)

BBB = scale(bird\_sub)

D = dist(BBB,method="euclidean")

cluster\_results = agnes(D,method="single")

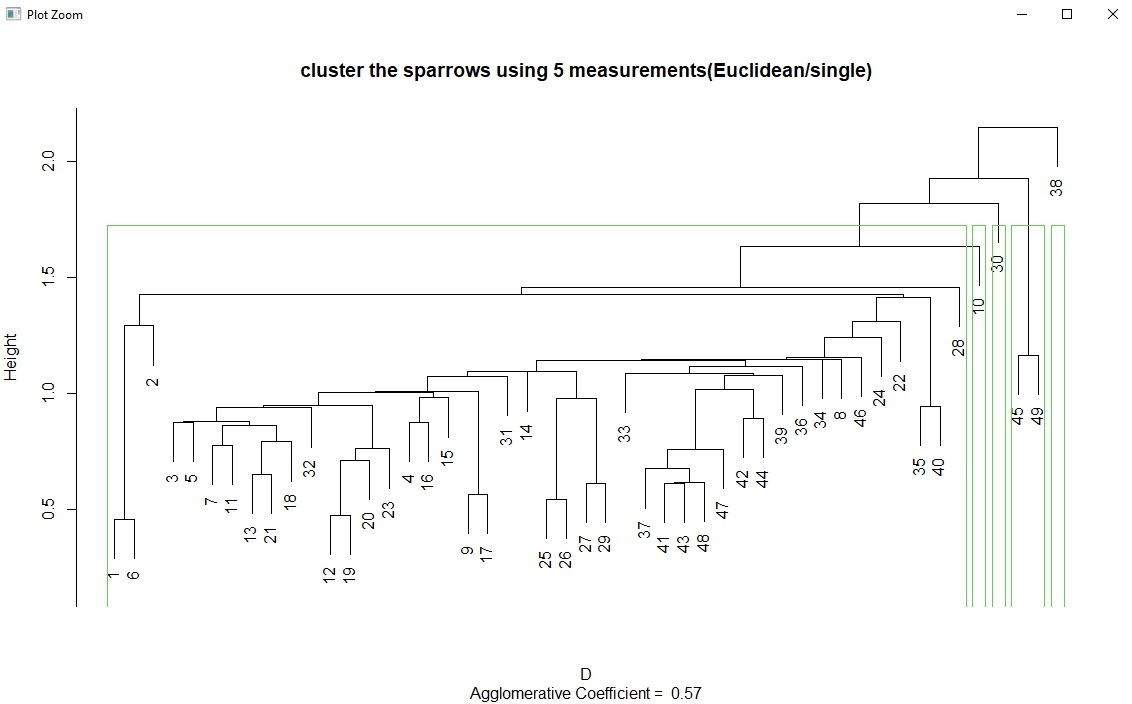
plot(cluster\_results,which.plots=**2**,

main='cluster the sparrows using 5 measurements(Euclidean/single)')

rect.hclust(cluster\_results, k=**5**, border=**3**)

plot(**as**.dendrogram(cluster\_results),horiz=FALSE)

clusters = cutree(cluster\_results,k=**5**)



**#Euclidean ,average method -- D = 0.77**

library(cluster)

BBB = scale(bird\_sub)

D = dist(BBB,method="euclidean")

cluster\_results = agnes(D,method="average")

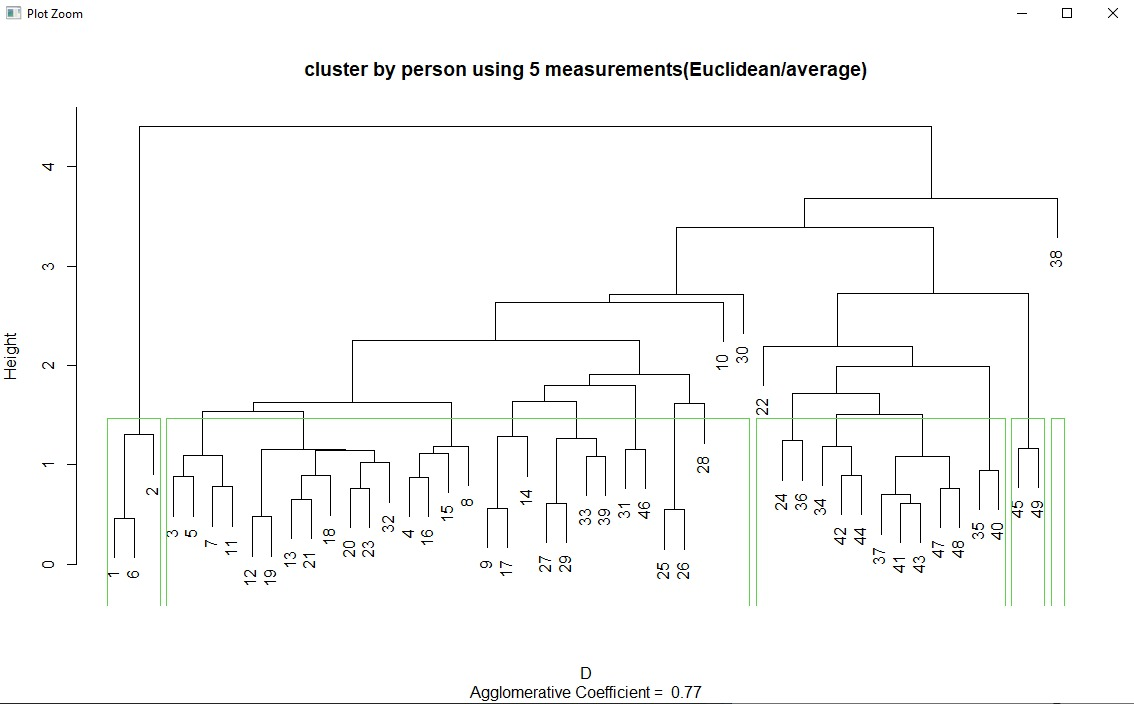
plot(cluster\_results,which.plots=**2**,

main='cluster by person using 5 measurements(Euclidean/average)')

rect.hclust(cluster\_results, k=**5**, border=**3**)

plot(**as**.dendrogram(cluster\_results),horiz=TRUE)

clusters = cutree(cluster\_results,k=**5**)



**# Manhattan , ward method , D = 0.94**

library(cluster)

BBB = scale(bird\_sub)

D = dist(BBB,method="manhattan")

cluster\_results = agnes(D,method="ward")

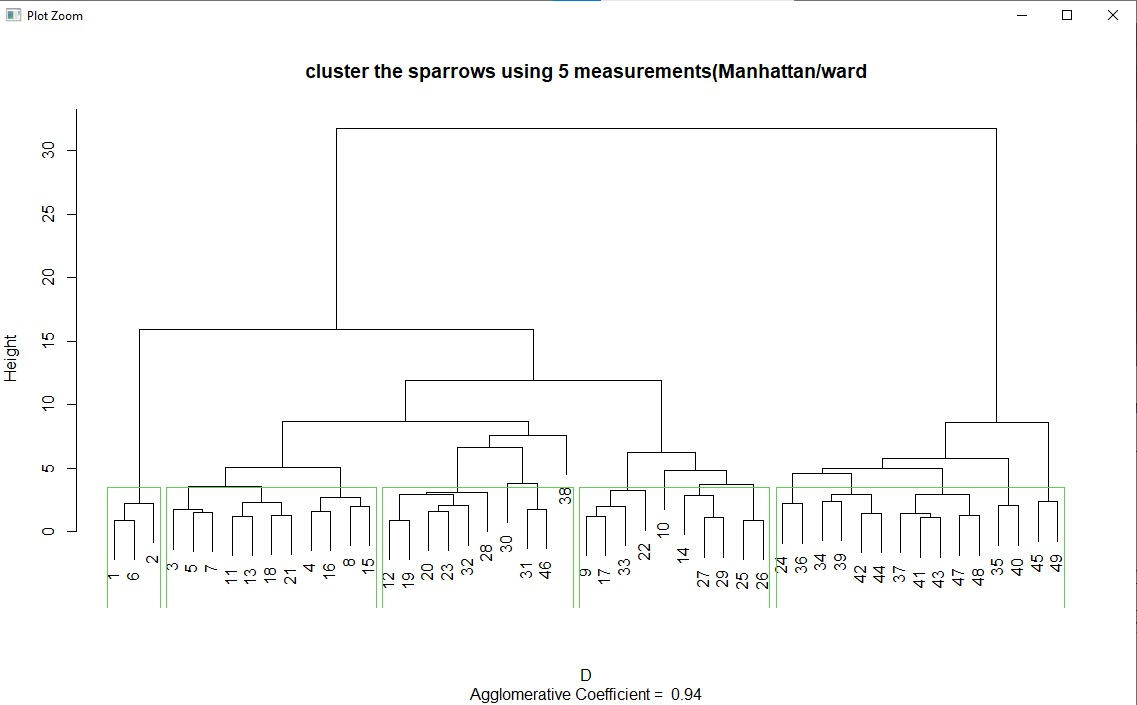
plot(cluster\_results,which.plots=**2**,

main='cluster the sparrows using 5 measurements(Manhattan/ward')

rect.hclust(cluster\_results, k=**5**, border=**3**)

plot(**as**.dendrogram(cluster\_results),horiz=TRUE)

clusters = cutree(cluster\_results,k=**5**)



**# Manhattan , single method , D = 0.62**

library(cluster)

BBB = scale(bird\_sub)

D = dist(BBB,method="manhattan")

cluster\_results = agnes(D,method="single")

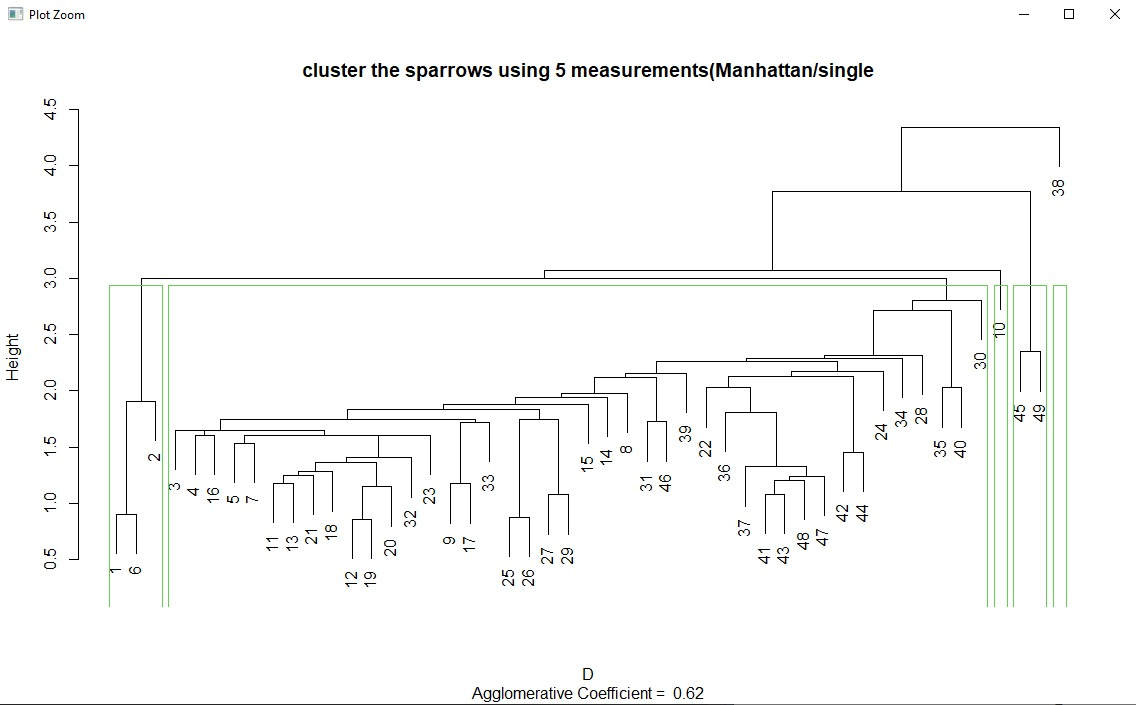
plot(cluster\_results,which.plots=**2**,

main='cluster the sparrows using 5 measurements(Manhattan/ward')

rect.hclust(cluster\_results, k=**5**, border=**3**)

plot(**as**.dendrogram(cluster\_results),horiz=TRUE)

clusters = cutree(cluster\_results,k=**5**)



**# Manhattan , average method , D = 0.62**

library(cluster)

BBB = scale(bird\_sub)

D = dist(BBB,method="manhattan")

cluster\_results = agnes(D,method="single")

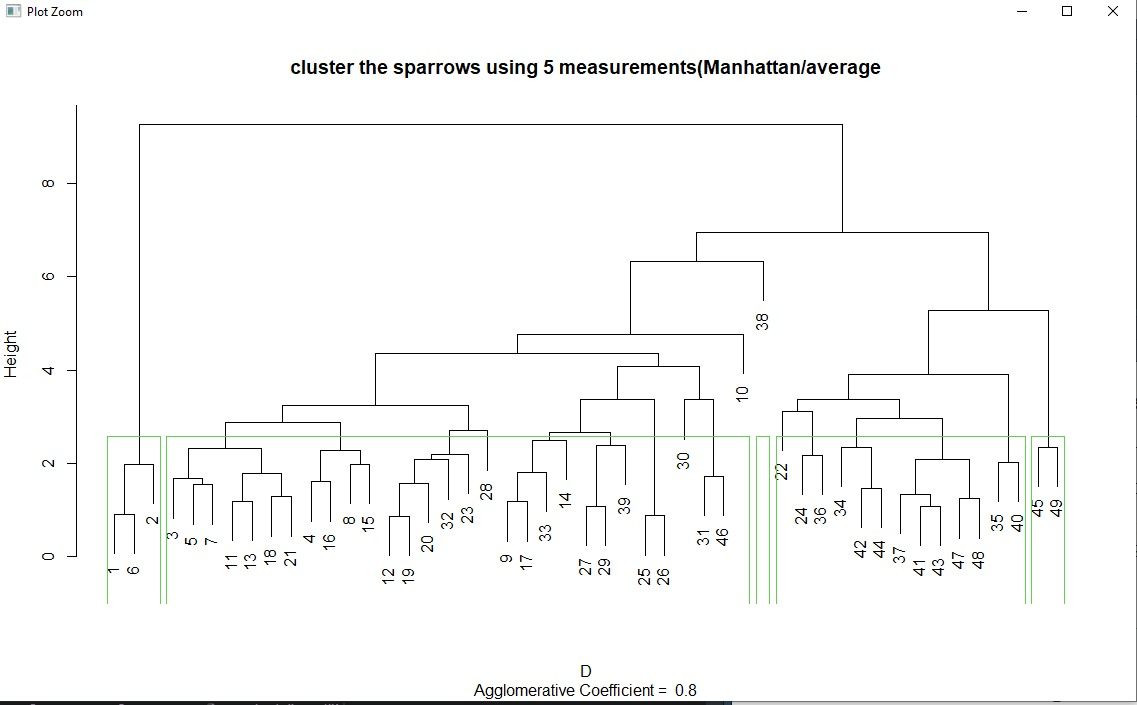
plot(cluster\_results,which.plots=**2**,

main='cluster the sparrows using 5 measurements(Manhattan/average')

rect.hclust(cluster\_results, k=**5**, border=**3**)

plot(**as**.dendrogram(cluster\_results),horiz=TRUE)

clusters = cutree(cluster\_results,k=**5**)



|  |  |  |
| --- | --- | --- |
| **Hierarchical methods** | **Euclidean Distance** | **Manhattan Distance** |
| **Single** | 0.57 | 0.62 |
| **ward** | 0.93 | 0.94 |
| **Average** | 0.77 | 0.8 |

Implemented the two-distance metrics (Euclidean and Manhattan) using three hierarchical methods (single, ward, average) and found Agglomerative Coefficient D. Both results from Euclidean and Manhattan distance metrics looks nearly similar, but Euclidean Distance value is little less when compared to Manhattan Distance.

**Question 3:**

Critically assess the methods applied and insights gained in this group task. Draw overall conclusions from applying both PCA and Cluster Analysis to the dataset, both the insights they have in common and those that appear to conflict. Discuss how you could communicate these conclusions to the general public (not data scientists or bird experts).

**PCA**

* Principal Component Analysis (PCA) is a useful technique for exploratory data analysis, allowing you to better visualize the variation present in a dataset with many variables. It is particularly helpful in the case of "wide" datasets, where you have many variables for each sample.
* PCA allows you to see the overall "shape" of the data, identifying which samples are similar to one another and which are very different. This can enable us to identify groups of samples that are similar and work out which variables make one group different from another.
* Using Prcomp() method,we can find the standard deviation and correlation coefficient for all the five measurements of all the died/survived sparrows.
* Here PC1 (72.35%) + PC2 (10.75%) gives 83% of variance. we can see that 83% of the variance is measured by the first two PC’s.

Also produced Biplot, Scree plot and Loading plot.

* In the Biplot between PC1 and PC2 ,the axes are seen as arrows originating from the center point. Here, we see that the variables X1,X2,X3,X4 and X5 all contribute to PC1, with higher values in those variables moving the samples to the right on this plot.The x-axis displays the PC1 and the y-axis displays the PC2 with respect to GROUP(died/survival).
* In Scree plot, we’ll calculate the percentage of total variance explained by each principal component. The x-axis displays the principal component and the y-axis displays the percentage of total variance explained by each individual principal component.
* The first principal component explains **72.350680%** of the total variation in the dataset.
* The second principal component explains **10.752729%** of the total variation in the dataset.
* The third principal component explains **7.618846%** of the total variation in the dataset.
* The fourth principal component explains **6.000459%** of the total variation in the dataset.
* The fifth principal component explains **3.277286%** of the total variation in the dataset.

Notice that all of the percentages sum to 100%.

* In loadings plot, it shows the effect of each measurement to construct PCA. Loading means Eigen vector, which means coefficient. Loadings plot means representing coefficient in a plot.PC1 is weighted average of the measurement because all the 5 measurements are positive. For PC2,PC3,PC4 and PC5 some of the measurements are negative as it explains less percentage if the total variation in the dataset.
* In the Biplot between PC2 and PC3 ,the axes are seen as arrows originating from the center point. Here, we see that the variables X1,X2,X3,X4 and X5 all contribute to PC2, with higher values in those variables moving the samples distributed equally on this plot.The x-axis displays the PC2 and the y-axis displays the PC3 with respect to GROUP(died/survival).

**Cluster Analysis:**

There are different clustering methods to analyse the data. Here we use hierarchical clustering to cluster the sparrows (Died/Survival) and the five measurements.

**PCA-Cluster Analysis:**

Cluster Analysis attempts to put the observations of your dataset into groups using some sort of distance metric. Cluster analysis split X’s **rows** into some groups based on relative distance.

Principal Components Analysis (PCA) takes n input variables (Y) and creates a new set of PV variables (Z) that summarize the information in the Y’s more efficiently. The hope is that a small number of Z’s can be used to represent the entire set of Y’s. Thus, PCA is a data reduction technique. Principal component analysis work on X’s **columns** (well, accurately speaking covariance of X) to choose, drop, or reconstruct features.

Cluster Analysis and PCA are obviously very different. However, there is a connection. Cluster Analysis works better if the set of input variables for your cases are uncorrelated. And PC variables are uncorrelated by construction.

So, here’s an approach that you can try. Create PC’s from your set of Y’s. Use the PC variables as inputs to the Cluster Analysis. Create groups using the Cluster Analysis. Once the groups are created, THROW AWAY the PC variables. Compute the means of the Y’s for each of your groups. Interpret the groups by looking at how the means of the Y’s differ among the groups.

This is why they are different fundamentally. They do very different jobs and serve different purposes. But in practice both are often used together to understand data behavior. Thus, the two procedures are very different. However, you can chain them together as:

* Y → PCA → PC Variables → Cluster Analysis → Groups of Cases

This approach is theoretically sound and usually leads to excellent results.

Reference links:

* <https://www.datacamp.com/community/tutorials/pca-analysis-r>
* <https://www.r-tutor.com/gpu-computing/clustering/hierarchical-cluster-analysis>
* <https://www.geeksforgeeks.org/how-to-calculate-manhattan-distance-in-r/>
* <https://www.statisticshowto.com/wards-method/#:~:text=Like%20other%20clustering%20methods%2C%20Ward's,the%20sum%20of%20squares%20index>
* <https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/>
* <https://www.statology.org/scree-plot-r/>
* <https://www.datacamp.com/community/tutorials/pca-analysis-r>
* <https://www.quora.com/What-is-the-difference-between-cluster-analysis-and-principal-component-analysis>