1.1 Number of Clusters

As per the nature of the columns in EastWestAirlinesCluster.xls, The data is mixed data. Hence the categorical variables are converted into binary and then clustering is applied.

As per the generated dendogram below, if slicing is done at height 100 then there are 3 clusters formed.

As per the dendogram branching is done into 2 parts and left branch alone formed one cluster and 2nd branch(Right Branch) again splitted into two more branches. The two branches of right branch are formed into an cluster. Hence in total 3 clusters with maximum number of customers.

R code illustrating the procees to create Hierarchical Clustering

*## Libraries*

**library**("dummies")

**library**("dendextend")

**library**("dendextendRcpp")

**library**("gridExtra")

**library**("cluster")

**library**("factoextra")

**library**("MASS")

**library**("fpc")

*## Set the working directory to where the Excel file is*

setwd('J:\\ISB Business Analytics\\Data Mining\\Data Mining Assignment 1')

*## Input file read*

input <- read.csv("EastWestAirlinesClusterCSV.csv",header=TRUE)

*## 1. Loading and preparing data*

mydatawd <- input[,2:11]

*#Creating Dummy variables for categorical data*

mydata <- dummy.data.frame(mydatawd, names = "cc1\_miles", omit.constants=FALSE )

mydata <- dummy.data.frame(mydatawd, names = "cc2\_miles", omit.constants=FALSE )

mydata <- dummy.data.frame(mydatawd, names = "cc3\_miles", omit.constants=FALSE )

*# Standardize Data*

my\_data <- scale(mydata)

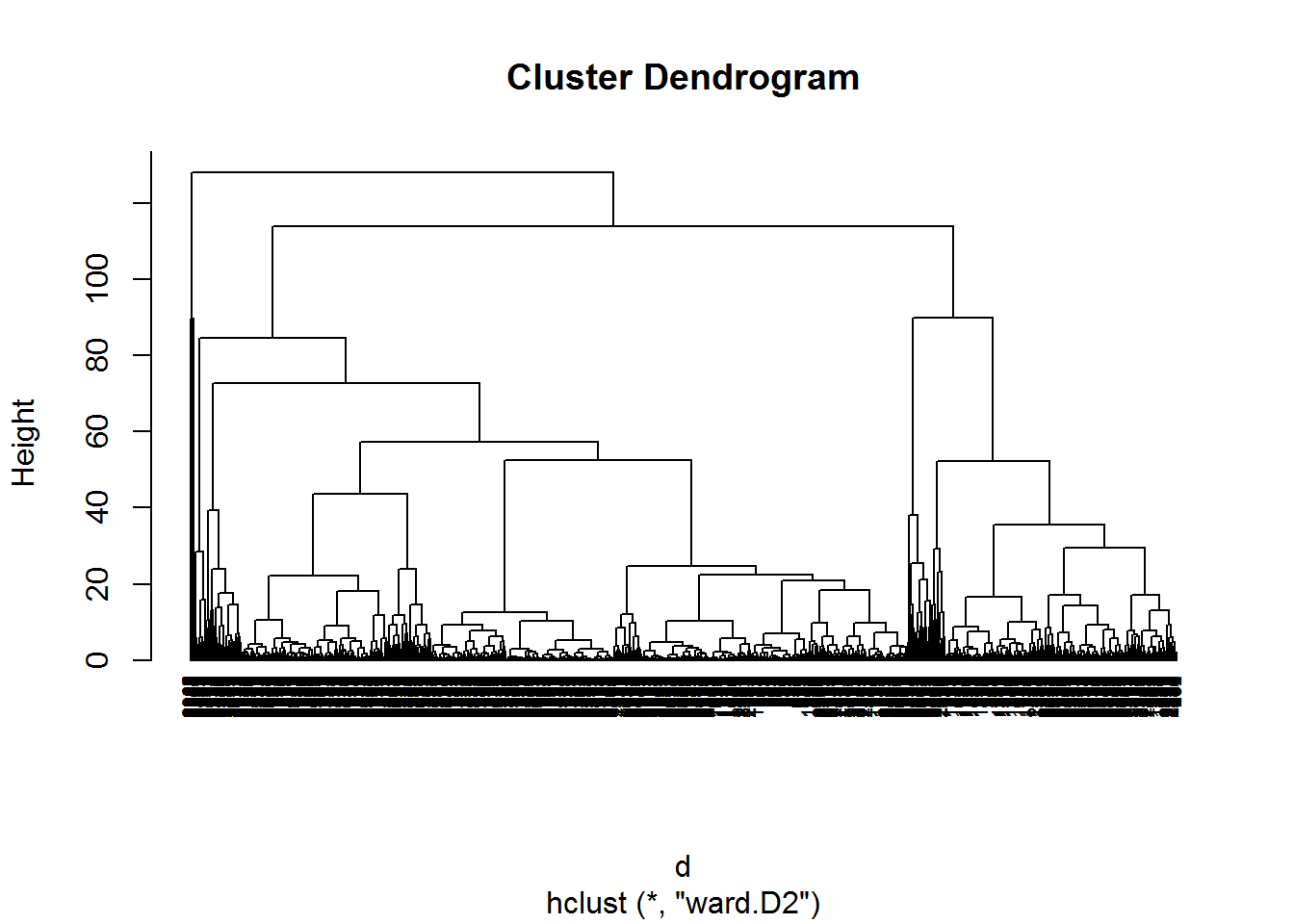
*# 2. Compute dissimilarity matrix*

d <- dist(my\_data, method = "euclidean")

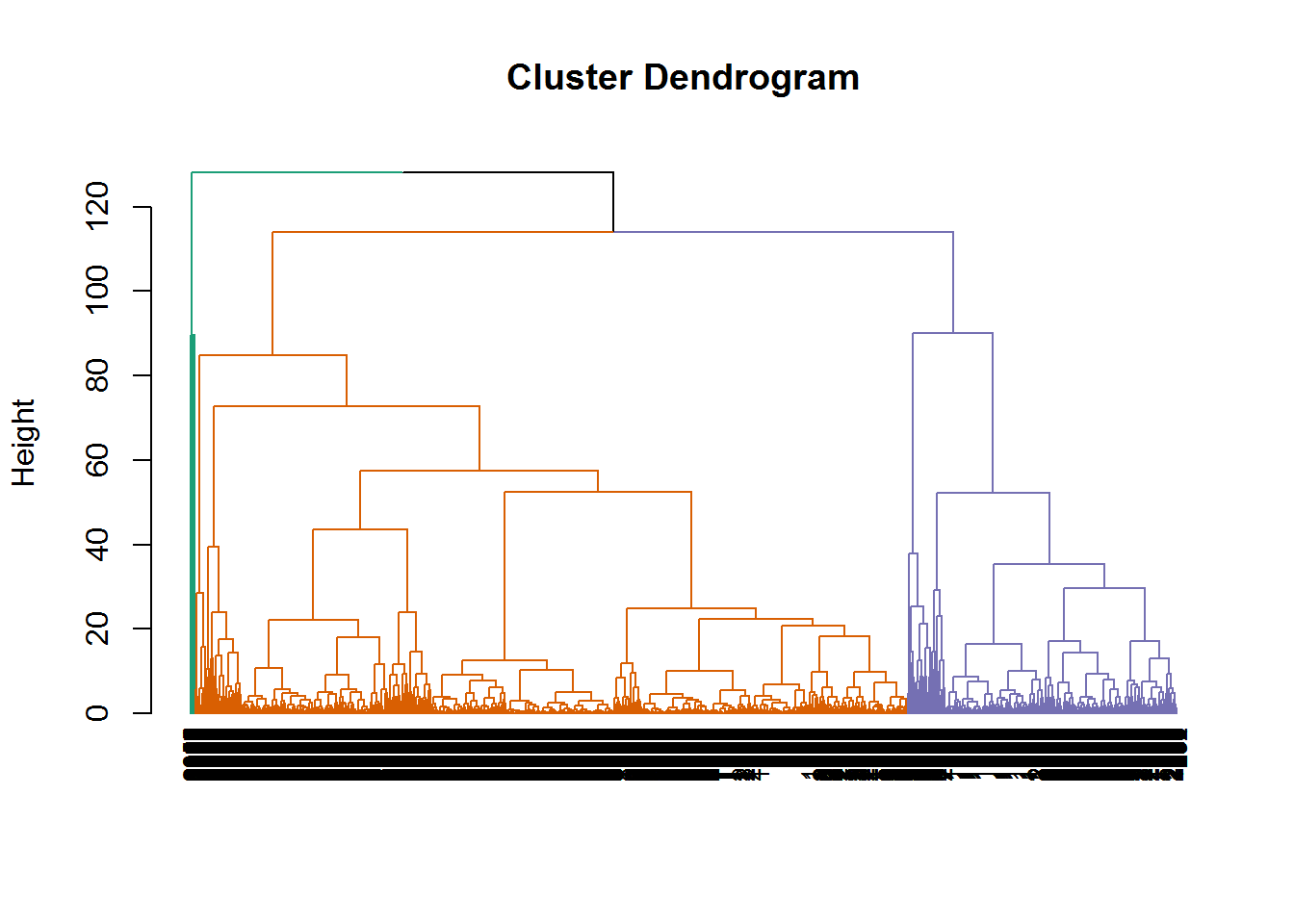
*# Hierarchical clustering using Ward's method*

res.hc <- hclust(d, method = "ward.D2" )

Cluster Dendogram in Ward’s Method



## Different clusters with differentation of color



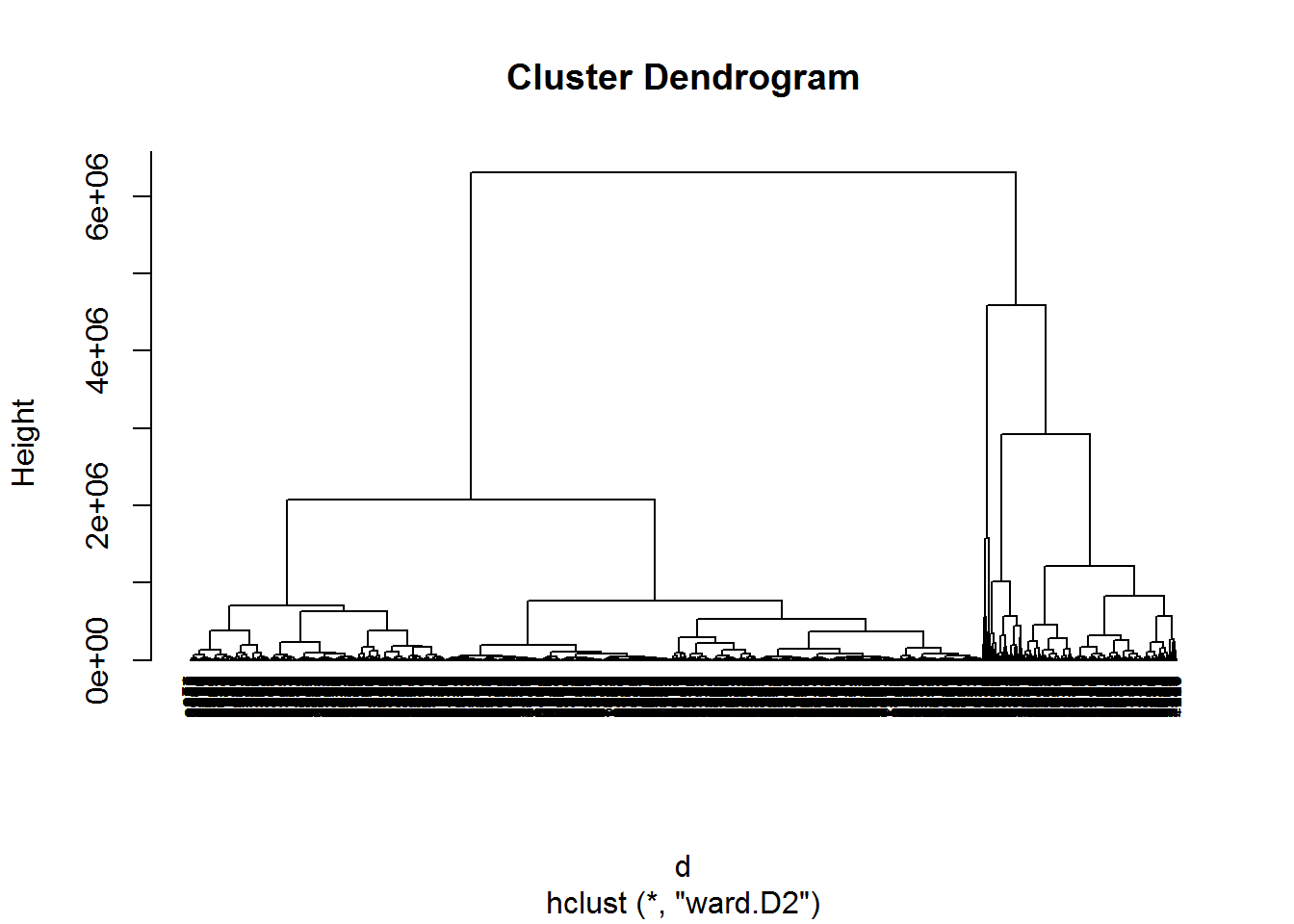
## 1.2 Standardization of Data

Stanardization/Normalization of data is not done then below issues will influence the model.

-> Distance measure will be wrongly clacualted, if all the variables are not with equal weight.

-> Laregest scale dominating the measure

The below dendogarm generated with the given data set without standardization is as below. As per the insights large value units are influencing the small values.



## 1.3 Cluster Centroids

The below are the metrics of cluster centriods.

### **Cluster 2 - Label - High Networth frequent fliers**

Cluster 2 metrics are leading in progressive way except non-flight bonus transactions.

These are the segment of customers, who are associated with EastWest Airlines since long time.

### **Cluster 3 - Label - Non Frequent travellers**

Cluster 3 metrics depicts that the flight miles and flight transactions in last 12 months is zero and their

non-flight bonus transactions are leading than the other clusters.

### **Cluster 1 - Label - Middle class travellers**

Cluser 1 metrics depicts that the level of spending is average to cluster 2 and 3.

### **Centriod Metrics - Column versus Clusters**

## [,1] [,2] [,3]

## Balance 73547.288828 115682.00 89401.333

## Qual\_miles 144.071142 295.50 0.000

## cc1\_miles 2.057114 3.75 3.000

## cc2\_miles 1.014529 1.00 1.000

## cc3\_miles 1.009519 3.00 2.000

## Bonus\_miles 17064.150802 74597.50 47920.000

## Bonus\_trans 11.576653 32.75 17.000

## Flight\_miles\_12mo 460.035822 825.00 0.000

## Flight\_trans\_12 1.373246 2.75 0.000

## Days\_since\_enroll 4117.809118 5058.75 3863.333

## 1.4 Stability of Clusters Validation

The Process to check for the stability of clusters check. Create different Dendogarms with the sample of 95% data. Here Dend1 and Dend2 are with sample of 95% data.

*# Hierarchical clustering using Ward's method*

Dend1 <- as.dendrogram(res.hc)

*#Random Sample1 with 95% of data*

input2=input[sample(nrow(my\_data),replace=F,size=0.95\*nrow(input)),]

d <- dist(input2, method = "euclidean")

res2.hc <- hclust(d, method = "ward.D2" )

Dend2 <- as.dendrogram(res2.hc)

*#Random Sample2 with 95% of data*

input3=input[sample(nrow(my\_data),replace=F,size=0.95\*nrow(input)),]

d <- dist(input3, method = "euclidean")

res3.hc <- hclust(d, method = "ward.D2" )

Dend3 <- as.dendrogram(res3.hc)

### **Global Comparison of Dendograms:**

*# Global Comparison of Dendograms*

*# Total Population Metrics*

Dend1

## 'dendrogram' with 2 branches and 3999 members total, at height 128.0329

*# Random Sample 1 Metrics*

Dend2

## 'dendrogram' with 2 branches and 3799 members total, at height 6296485

*# Random Sample 2 Metrics*

Dend3

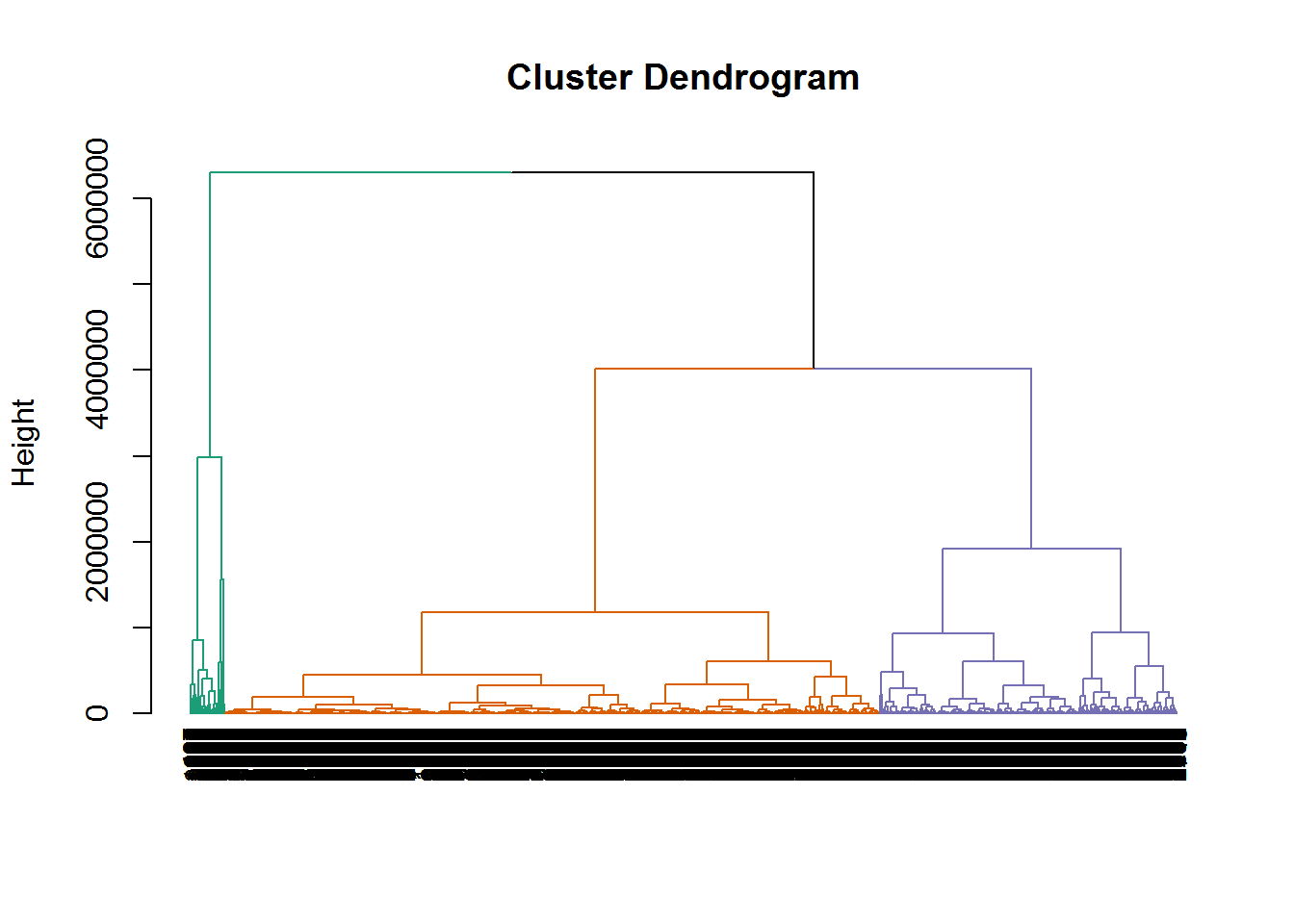
## 'dendrogram' with 2 branches and 3799 members total, at height 5987322

*# Comparison of Population, Random sample 1 and Random sample 2 dendogarms.*

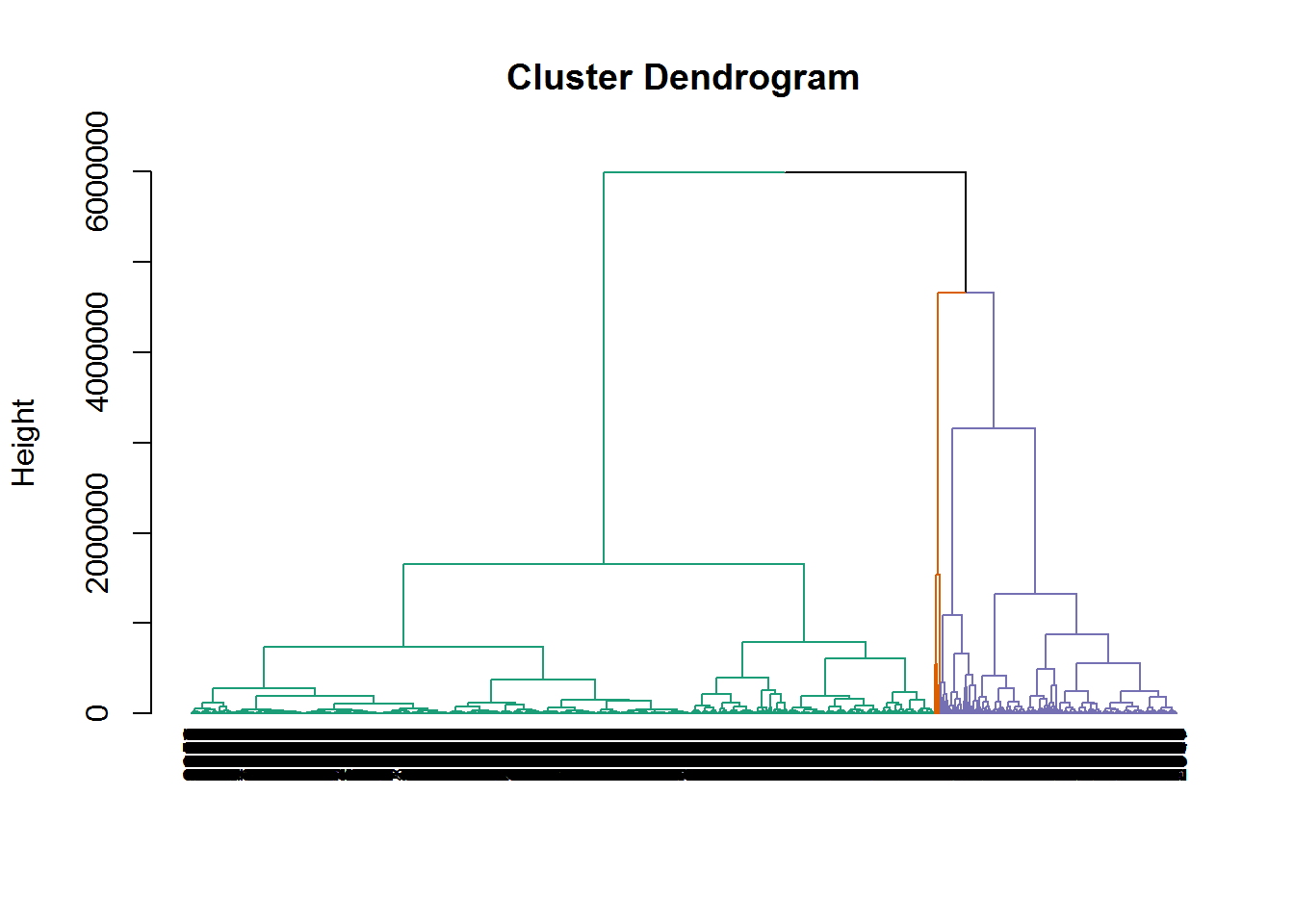
all.equal(Dend3, Dend2, Dend1, use.edge.length = TRUE)

## [1] "Difference in branch heights - Mean relative difference: 0.04946635"

#### **Sample 1 Dendogarm:**



#### **Sample 2 Dendogarm:**



### **Conclusion**

From above stastics of dendogram, dendogram comparison and dendogram visualizations, the dendogarams are not same(The Pictures are not same).For every instance of sample, new insights are emerged.

## 1.5 K-Means Clustering

The below R code depicts the process to get the aggregate metrics for Hierarchial clustering and K-Means.

*## K-means clustering*

set.seed(123)

fit <- kmeans(my\_data, 3) *# 3 cluster solution*

*#Aggregation of k-means*

mydatak <- data.frame(mydata, fit$cluster) *# append cluster membership*

temp <- aggregate(mydatak, by=list(fit$cluster), FUN=mean)

*#to find the size of clusters*

ClusterCo <- aggregate(mydatak, by=list(fit$cluster), FUN=sum)

*#to find the cluster size*

d <- transform(ClusterCo, clusterSize = fit.cluster / Group.1)

d <- transform(d, fit.cluster= fit.cluster/ clusterSize)

temp$clusterSize <- d$clusterSize

temp$clusterPCT <- (d$clusterSize\*100)/3999

*# transpose to change from horizontal to vertical*

temp2 <- t(temp)

round\_df <- **function**(x, digits) {

*# round all numeric variables*

*# x: data frame*

*# digits: number of digits to round*

numeric\_columns <- sapply(x, class) == 'numeric'

x[numeric\_columns] <- round(x[numeric\_columns], digits)

x

}

temp4 <- round\_df(temp2, 2)

*#Hierarchical Aggregate calculations*

*# Hierarchical clustering using Ward's method*

*#set.seed(123)*

groups <- cutree(res.hc, k=3) *# cut tree into 3 clusters*

membership <-as.matrix(groups)

membership <- data.frame(membership)

names(membership) <- c("cluster")

mydatao <- data.frame(mydata, membership$cluster) *# append cluster membership*

temp <- aggregate(mydatao, by=list(membership$cluster), FUN=mean)

temp2 <- t(temp)

round\_df <- **function**(x, digits) {

*# round all numeric variables*

*# x: data frame*

*# digits: number of digits to round*

numeric\_columns <- sapply(x, class) == 'numeric'

x[numeric\_columns] <- round(x[numeric\_columns], digits)

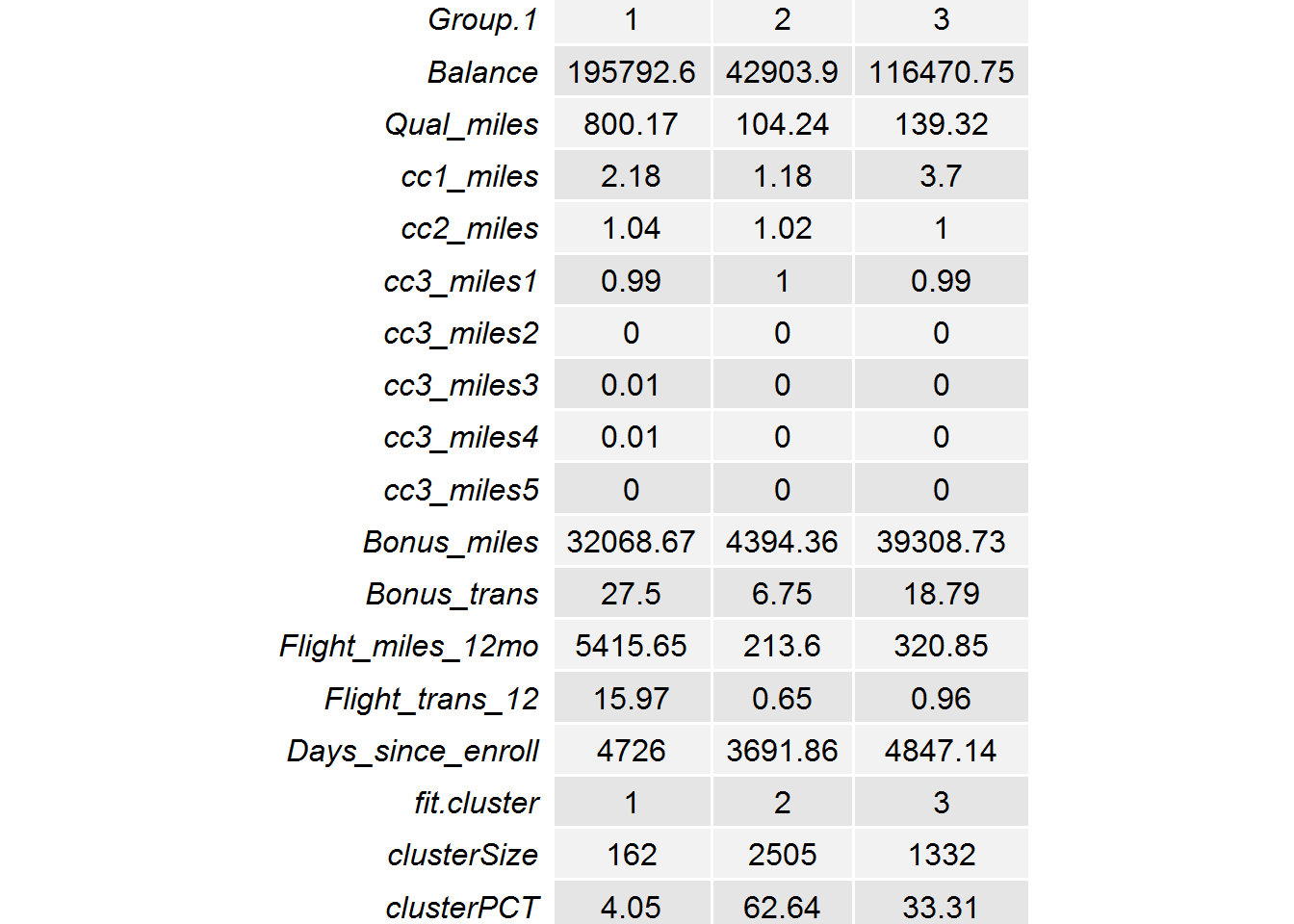
x

}

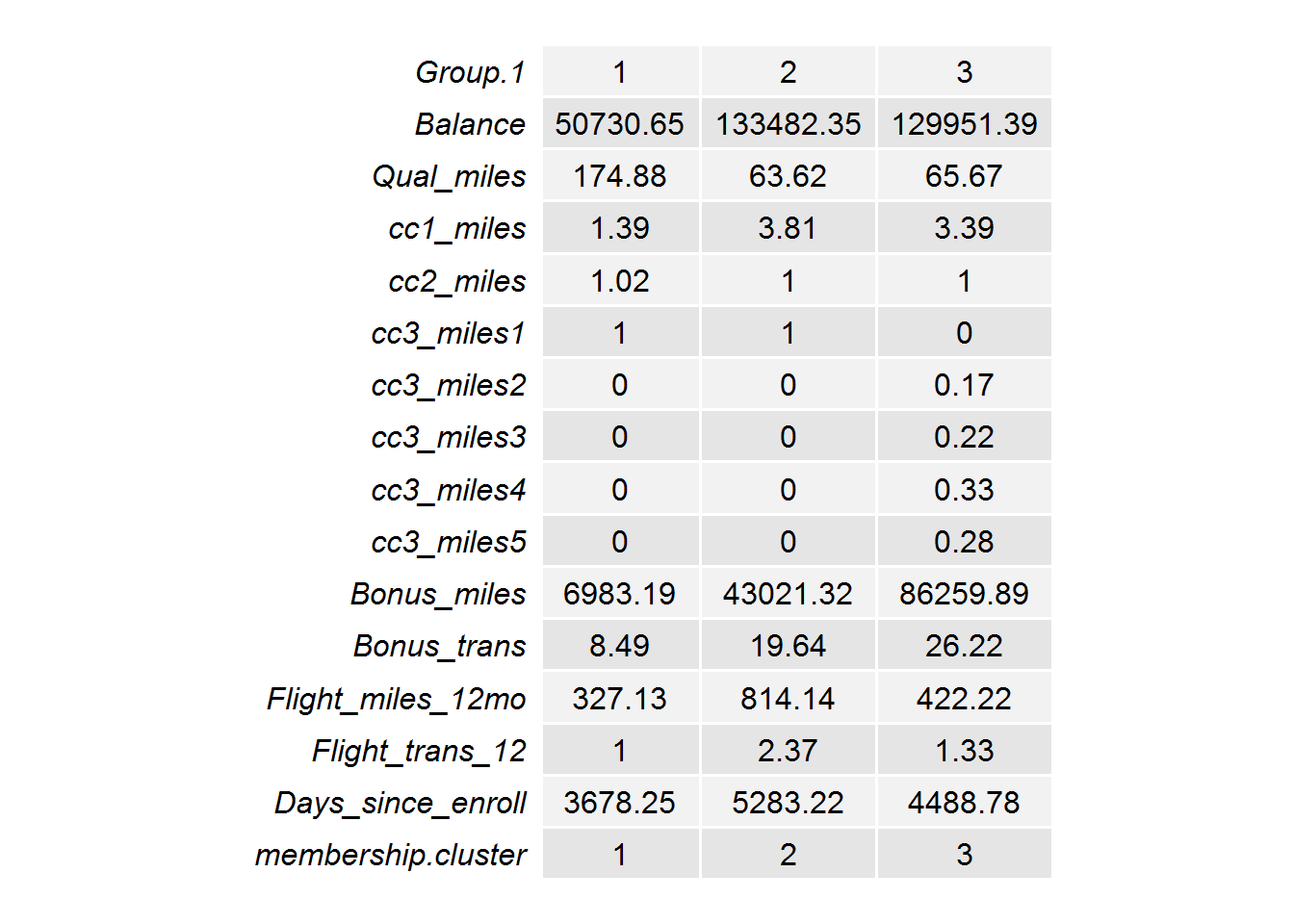
temp5 <- round\_df(temp2, 2)

### **K-Means Aggregate Metrics with Mean**

Along with the primary dataset metrics, each cluster size( clusterSize ) and percentage of cluster size( clusterPCT ) in given population are calculated and tabulated.



### **Hierarchical Cluster Aggregate Metrics with Mean**



### **Conclusion**

From above metrics, none of the clusters are matching, Hence the outcomes of K-means cluster and Hierarchial clustering are not same.