EDA Human Activity Analysis using Smartphone Dataset

Anandu R

7/29/2020

EDA Human Activity Analysis using Smartphone Dataset

The dataset has information on various sensors on mobile phone, the data set has been split into training(70%) and testing(30%), with activities manually labelled with the associated activity and subject under consideration in the training data set.

We use this data to predict the activity of the user based on readings from the mobile sensors

Loading the data

Downloading

```
fileUrl =
  "https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda"
if (!file.exists('./samsungData.rda')){
  download.file(fileUrl,'./samsungData.rda')
}
```

Loading

```
load("samsungData.rda")
```

Exploring the data

```
head(samsungData)[c(1:3,10:12)]
```

```
##   tBodyAcc-mean()-X tBodyAcc-mean()-Y tBodyAcc-mean()-Z tBodyAcc-max()-X
## 1         0.2885845       -0.02029417        -0.1329051       -0.9347238
## 2         0.2784188       -0.01641057        -0.1235202       -0.9430675
## 3         0.2796531       -0.01946716        -0.1134617       -0.9386916
## 4         0.2791739       -0.02620065        -0.1232826       -0.9386916
## 5         0.2766288       -0.01656965        -0.1153619       -0.9424691
## 6         0.2771988       -0.01009785        -0.1051373       -0.9424691
##   tBodyAcc-max()-Y tBodyAcc-max()-Z
## 1       -0.5673781       -0.7444125
## 2       -0.5578513       -0.8184087
```

```
## 3       -0.5578513       -0.8184087
## 4       -0.5761589       -0.8297115
## 5       -0.5691738       -0.8247053
## 6       -0.5656839       -0.8227661
```

```
names(samsungData)[seq(1,ncol(samsungData),length.out = 10)]
```

```
##  [1] "tBodyAcc-mean()-X"             "tGravityAcc-entropy()-X"
##  [3] "tBodyGyro-std()-Y"             "tBodyGyroJerk-arCoeff()-X,3"
##  [5] "tBodyGyroMag-arCoeff()2"       "fBodyAcc-bandsEnergy()-33,48"
##  [7] "fBodyAccJerk-meanFreq()-Z"     "fBodyGyro-min()-Z"
##  [9] "fBodyGyro-bandsEnergy()-49,64" "activity"
```

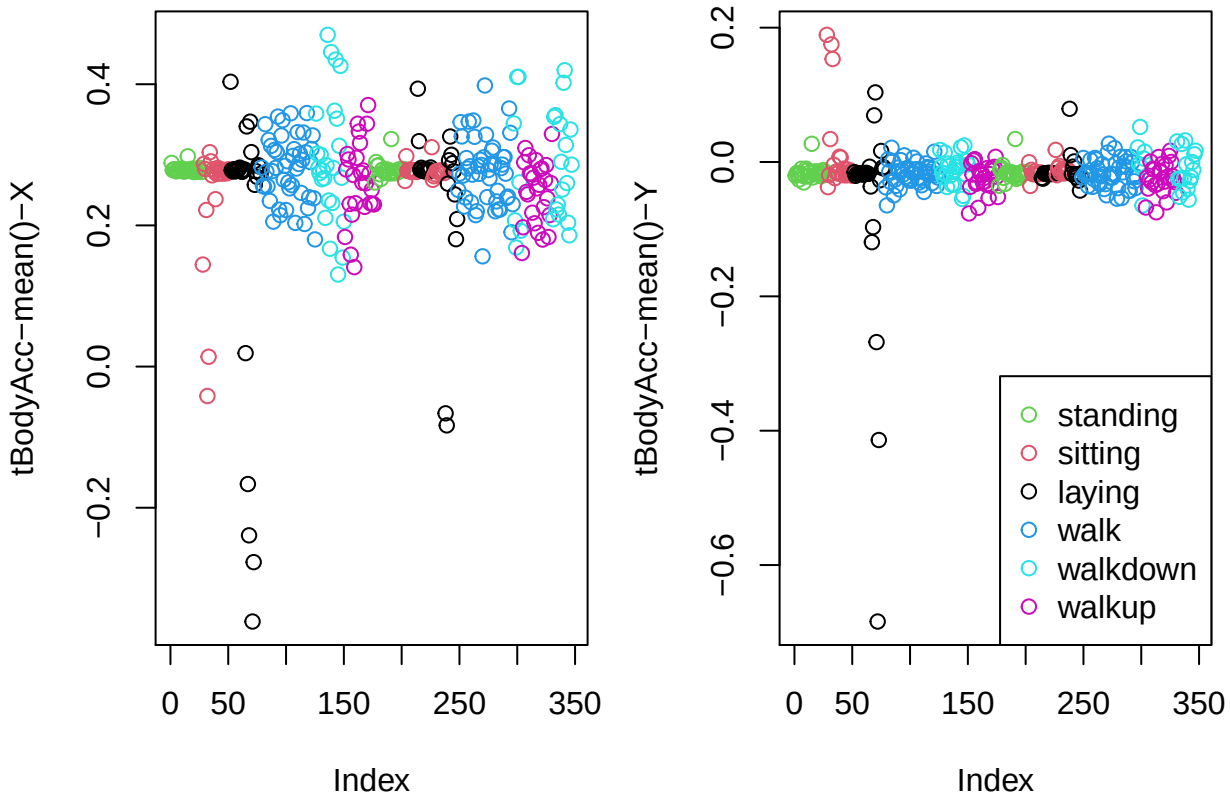
Clean variable names

```
data <- samsungData
variableNames = c(names(samsungData))
cleanVariableNames <- c(paste0("v",c(1:561)),"subject","activity")
names(data) <- cleanVariableNames
```

Exploratory Analysis

Plotting average acceleration for first subject

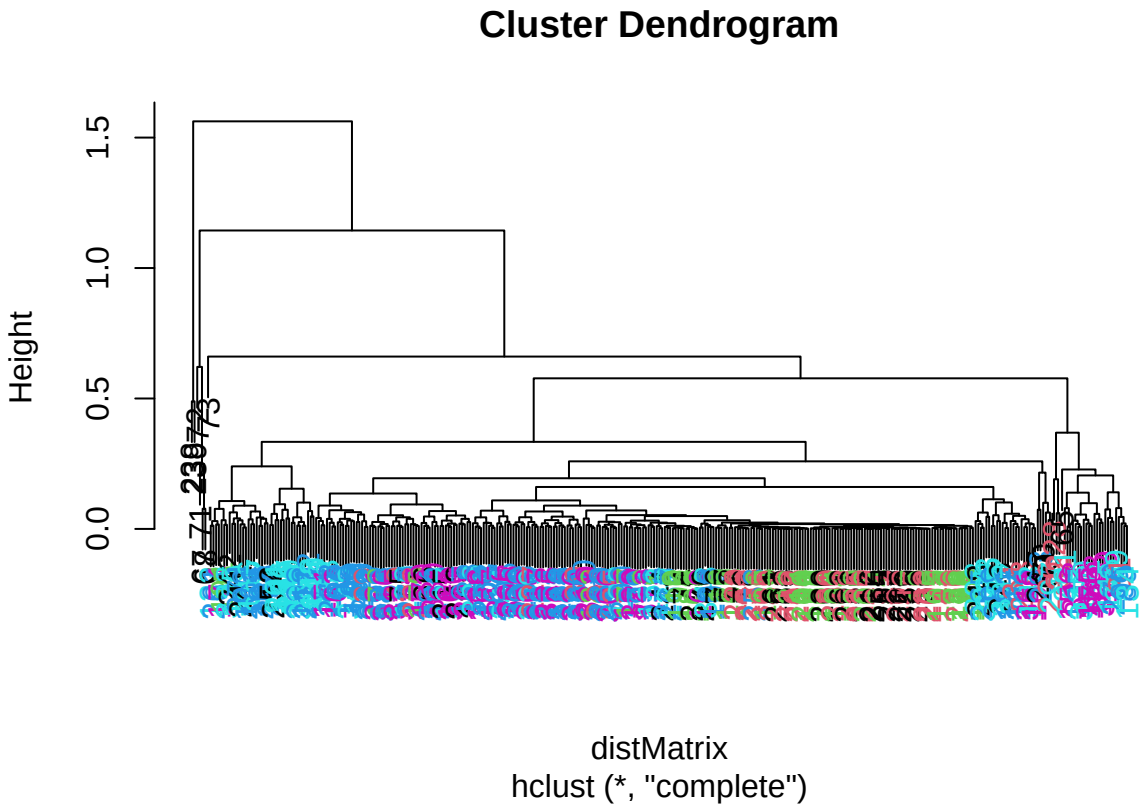
```
par(mfrow=c(1,2),mar = c(5,4,1,1))
# Change Activity variable to factor
data = transform(data, activity = factor(activity))
sub1 = subset(data, subject == 1)
plot(sub1[,1],col = sub1$activity, ylab=variableNames[1])
plot(sub1[,2],col = sub1$activity, ylab=variableNames[2])
legend("bottomright", legend=unique(sub1$activity),col = unique(sub1$activity),pch =1)
```



Clustering the data on average acceleration

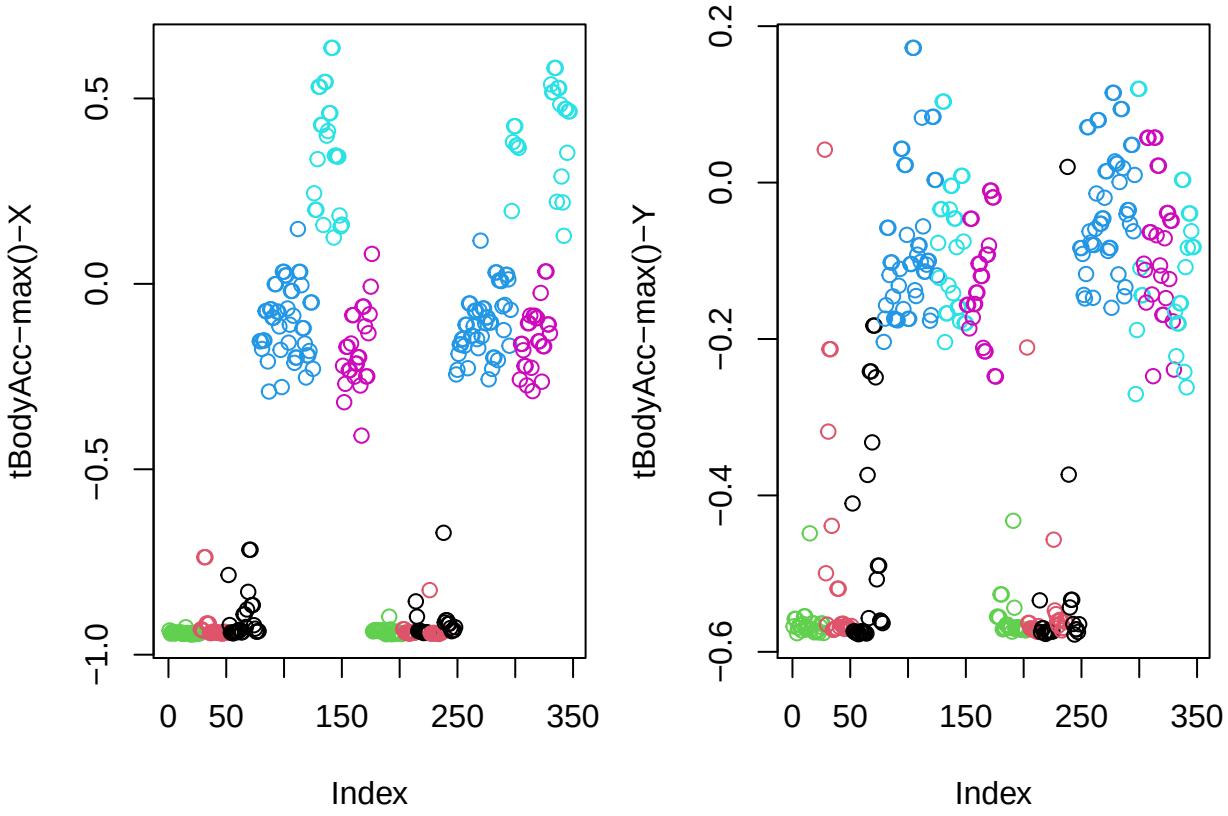
Average acceleration data is stored in the first 3 columns/variables v1-v3

```
source("myplclust.R")
distMatrix = dist(sub1[,1:3])
hclustering = hclust(distMatrix)
myplclust(hclustering, lab.col = c(unclass(sub1$activity)))
```



Plotting max acceleration for first subject

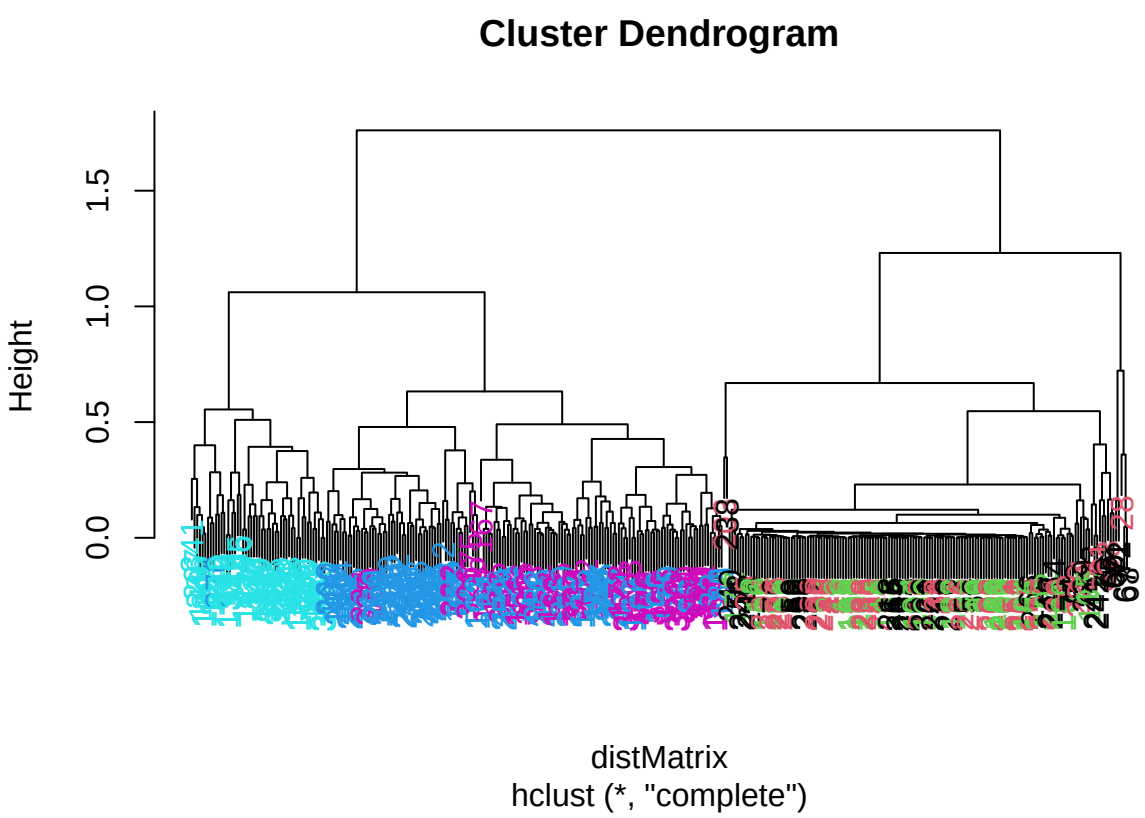
```
par(mfrow=c(1,2),mar = c(5,4,1,1))
plot(sub1[,10],col = sub1$activity, ylab=variableNames[10])
plot(sub1[,11],col = sub1$activity, ylab=variableNames[11])
```



Clustering the data on max acceleration

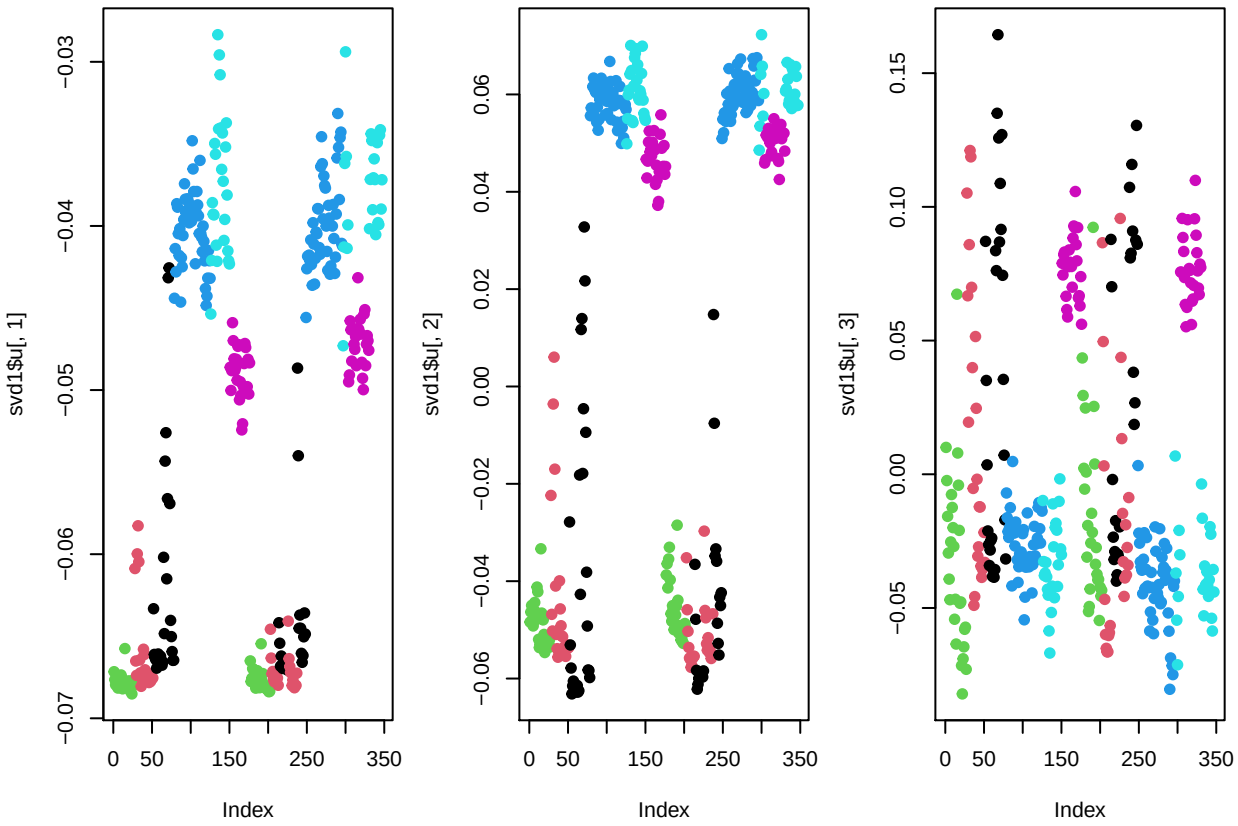
Average acceleration data is stored in the first 3 columns/variables v1-v3

```
source("myplclust.R")
distMatrix = dist(sub1[,10:12])
hclustering = hclust(distMatrix)
myplclust(hclustering, lab.col = c(unclass(sub1$activity)))
```



Performing Singular Value Decomposition

```
svd1 = svd(sub1[,-c(562,563)])
par(mfrow=c(1,3),mar = c(5,4,1,1))
plot(svd1$u[,1], col = sub1$activity, pch =19)
plot(svd1$u[,2], col = sub1$activity, pch =19)
plot(svd1$u[,3], col = sub1$activity, pch =19)
```

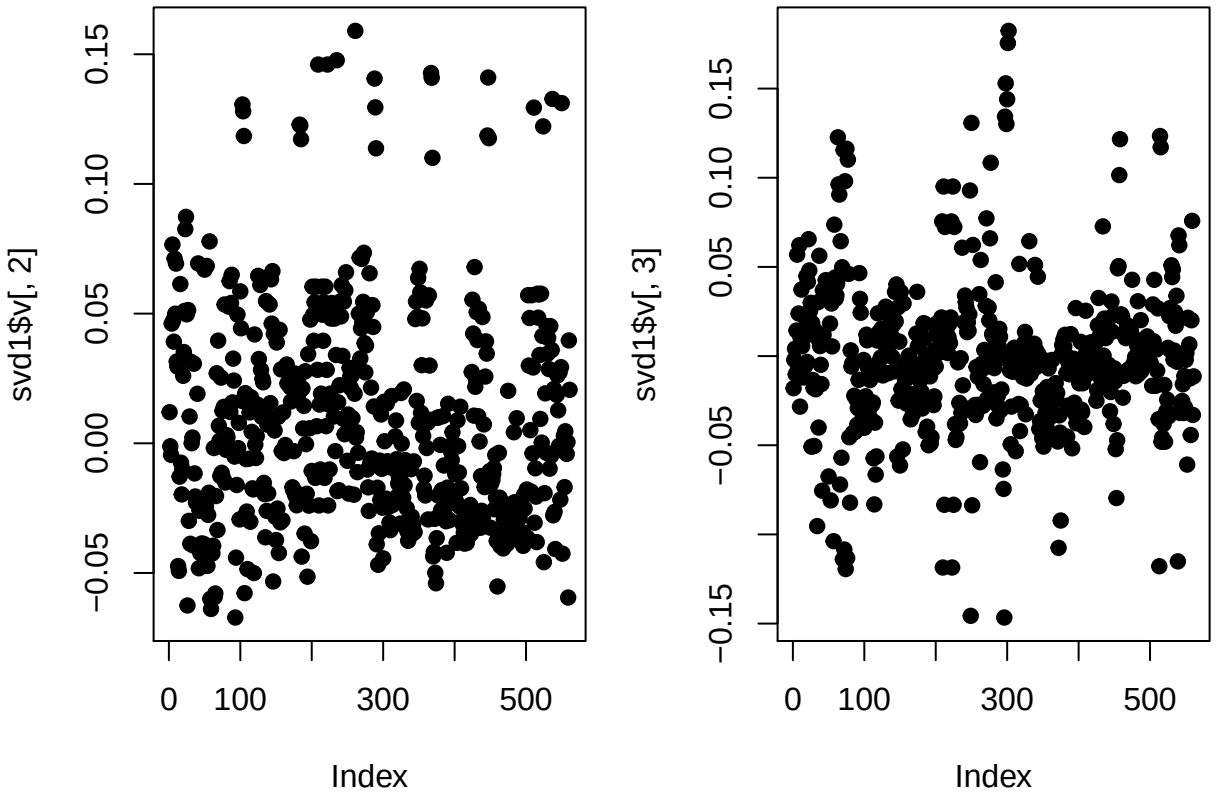


In the first left singular vector the moving activities have been separated from the stationary activities,the second singular vector doesn't seem to be all that different from the first but is definitely much better than the first left singular vector to separate the moving category from the stationary, whereas in the third singular vector we observe that the magenta color-coded activity distinguishly separated from the rest of theactivity.

Therefore we examine the Second third singular vector to find out the variable which contributes most to this peculiar behaviour.

Inorder to do so we examine the corresponding right singular vector - the second and third right singular vectors

```
par(mfrow=c(1,2),mar = c(5,4,1,1))
plot(svd1$v[,2],pch=19)
plot(svd1$v[,3],pch=19)
```



We use the which.max function to locate the index which has maximum value for the singular vectors 2 and 3 respectively.

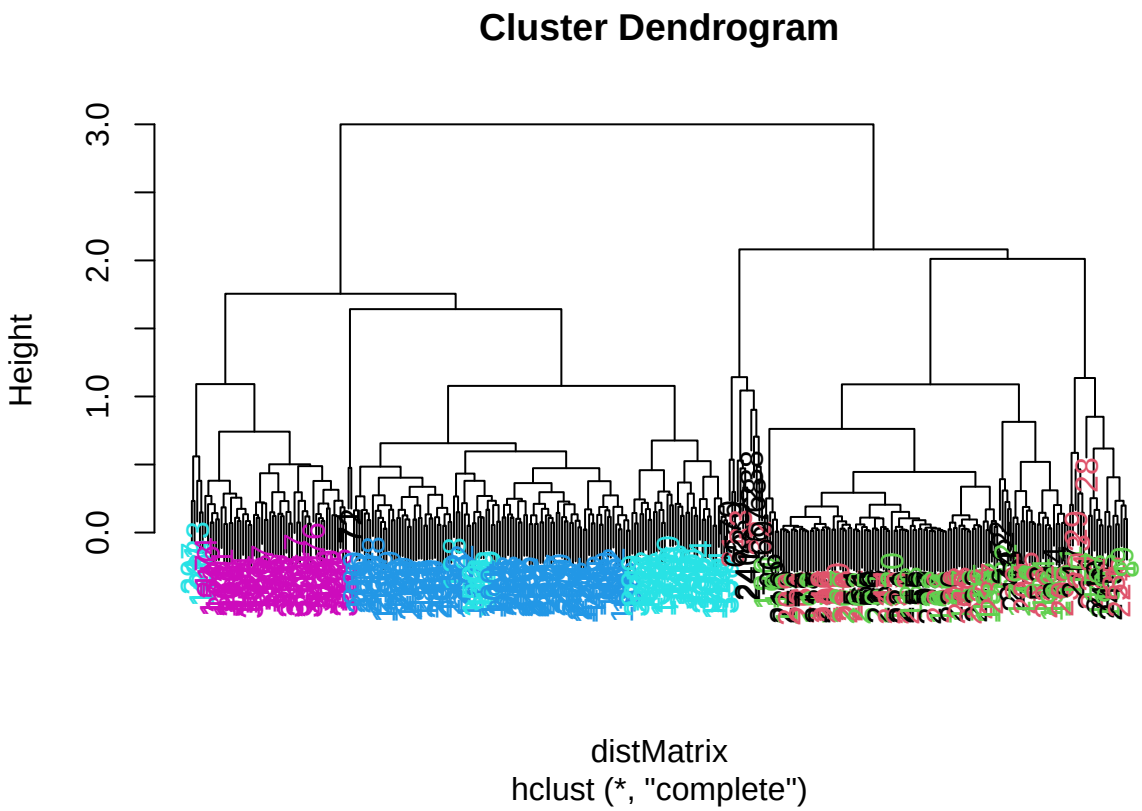
```
maxCon = c(which.max(svd1$v[,2]),which.max(svd1$v[,3]))
variableNames[maxCon]
```

```
## [1] "tBodyGyroJerkMag-entropy()" "fBodyAcc-kurtosis()-Z"
```

Upon inspection the variable in question are the "tBodyGyroJerkMag-entropy()" that influences the second singular vector the most, and "fBodyAcc-kurtosis()-Z" for the third similarly.

And now, we can use these values along with the original 3 max acceleration variables to cluster the data

```
distMatrix = dist(sub1[,c(10:12,maxCon)])
hclustering = hclust(distMatrix)
myplclust(hclustering, lab.col = c(unclass(sub1$activity)))
```



Compared to the initial clustering, Within the moving category there is better separation between the sub-categories, albiet just within the moving category, we need to perform further analysis to find the variable that can be used to separate the sub-category of activities within the stationary category.

Employing K-Means clustering to cluster the data into the various activies.

First try basis clustering

```
kClust = kmeans(sub1[,-c(562,563)],centers = 6, nstart = 250)
table(kClust$cluster,sub1$activity)
```

```
##
##      laying sitting standing walk walkdown walkup
##    1      0       0        0    0       49      0
##    2     29       0        0    0        0      0
##    3      3       0        0    0        0     53
##    4      0       0        0   95        0      0
##    5     18      10        2    0        0      0
##    6      0      37       51    0        0      0
```

Deleting unecessary folder

```r
unlink("./samsungData.rda")
```