# Perform the below given activities:

# a. Apply PCA to the dataset and show proportion of variance

# b. Perform PCA using SVD approach

# c. Show the graphs of PCA components

# -----------------------------------------------------------------

# Import Zip File

getwd()

setwd("E:\\Acadgild\\Class 11\\Assignment")

data <- read.csv(unzip("epi\_r.csv.zip", list = T)$Name[1])

View(data)

dim(data)

str(data)

# check for NA

sum(is.na(data))

sort(sapply(data, function(x) sum(is.na(x))))

# impute missing values

library(mice)

imputed = mice(data[,c("calories", "sodium", "protein", "fat")], method='cart', m=5)

imputed <- complete(imputed)

# replacing NAs with imputed values

dat <- data

dat$calories <- imputed$calories

dat$protein <- imputed$protein

dat$sodium <- imputed$sodium

dat$fat <- imputed$fat

sum(is.na(dat))

# checking for outliers

library(ggplot2)

ggplot(reshape2::melt(dat[,c("calories", "sodium", "protein", "fat")]),

aes(x= variable, value, fill = variable))+

geom\_boxplot()+facet\_wrap(~variable, scales = 'free\_y')

# yes there are outliers

# removing these outliers

df <- outliers::rm.outlier(dat[,c("calories", "sodium", "protein", "fat")], fill = TRUE)

View(df)

dat$calories <- df$calories

dat$protein <- df$protein

dat$sodium <- df$sodium

dat$fat <- df$fat

# to remove zero variance columns from the dataset, setting variance not equal to zero.

df<- df[ , apply(df, 2, var) != 0]

# principal component analysis

prin\_comp <- prcomp(dat[,-1], scale. = T)

names(prin\_comp)

prin\_comp$rotation[1:5,1:4]

# plot the resultant principal components

biplot(prin\_comp, scale = 0)

#compute standard deviation of each principal component

std\_dev <- prin\_comp$sdev

#compute variance

pr\_var <- std\_dev^2

#check variance of first 10 components

pr\_var[1:10]

#proportion of variance explained

prop\_varex <- pr\_var/sum(pr\_var)

prop\_varex[1:20]

#scree plot

plot(prop\_varex, xlab = "Principal Component",

ylab = "Proportion of Variance Explained",

type = "b")

#cumulative scree plot

plot(cumsum(prop\_varex), xlab = "Principal Component",

ylab = "Cumulative Proportion of Variance Explained",

type = "b")

library(factoextra)

fviz\_eig(prin\_comp)

fviz\_pca\_ind(prin\_comp, col.ind = "cos2", repel = TRUE,

gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))

fviz\_pca\_var(prin\_comp, col.ind = "contrib", repel = TRUE,

gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))