|  |
| --- |
| workdir <- "C:\Path\To\Dataset\Directory"  datafile <- file.path(workdir,"epic\_recipes.txt")  data <- read.table(datafile, fill=TRUE, col.names=1:max(count.fields(datafile)),  na.strings=c("", "NA"), stringsAsFactors = FALSE)    a <- aggregate(data[,-1], by=list(data[,1]), paste, collapse=",")  a$combined <- apply(a[,2:ncol(a)], 1, paste, collapse=",")  a$combined <- gsub(",NA","",a$combined) ## this column contains the totality of all ingredients for a cuisine    cuisines <- as.data.frame(table(data[,1])) ## Number of recipes for each cuisine  freq <- lapply(lapply(strsplit(a$combined,","), table), as.data.frame) ## Frequency of ingredients  names(freq) <- a[,1]  prop <- lapply(seq\_along(freq), function(i) {  colnames(freq[[i]])[2] <- names(freq)[i]  freq[[i]][,2] <- freq[[i]][,2]/cuisines[i,2] ## proportion (normalized frequency)  freq[[i]]}  )  names(prop) <- a[,1] ## this is a list of 26 elements, one for each cuisine    final <- Reduce(function(...) merge(..., all=TRUE, by="Var1"), prop)  row.names(final) <- final[,1]  final <- final[,-1]  final[is.na(final)] <- 0 ## If ingredient missing in all recipes, proportion set to zero  final <- t(final) ## proportion matrix    s <- sort(apply(final, 2, sd), decreasing=TRUE)  ## Selecting ingredients with maximum variation in frequency among cuisines and  ## Using standardized proportions for final analysis  final\_imp <- scale(subset(final, select=names(which(s > 0.1))))    ## heatmap  library(gplots)  heatmap.2(final\_imp, trace="none", margins = c(6,11), col=topo.colors(7),  key=TRUE, key.title=NA, keysize=1.2, density.info="none")    ## PCA and biplot  p <- princomp(final\_imp)  biplot(p,pc.biplot=TRUE, col=c("black","red"), cex=c(0.9,0.8),  xlim=c(-2.5,2.5), xlab="PC1, 39.7% explained variance", ylab="PC2, 24.5% explained variance") |

https://i1.wp.com/pixel.wp.com/b.gif?resize=1%2C1&ssl=1