Statistics 452: Statistical Learning and Prediction

Chapter 10, part 2.5: Clustering with Categorical X

Brad McNeney

2018-11-19

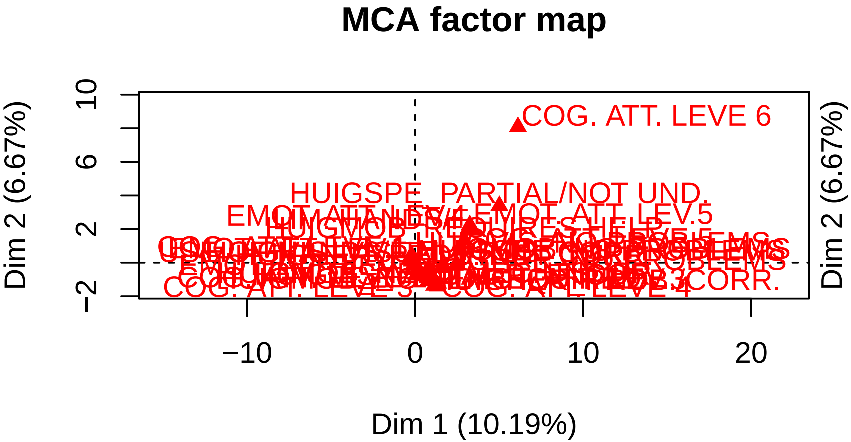
HUI Data - Read and Process as Before

```r
hui <- read.csv("../../Project652/HUI.csv")
recode_ns <- function(x) {
  x[x=="NOT STATED"] <- NA
  x <- droplevels(x)
  x
}
for(i in 1:ncol(hui)) {
  hui[,i] <- recode_ns(hui[,i])
}
hui <- na.omit(hui)
library(dplyr)
hsub <- select(hui,starts_with("HUI"))
names(hsub)
```

```
## [1] "HUIDCOG" "HUIGDEX" "HUIDEMO" "HUIGHER" "HUIGMOB" "HUIGSPE" "HUIGVIS"
```

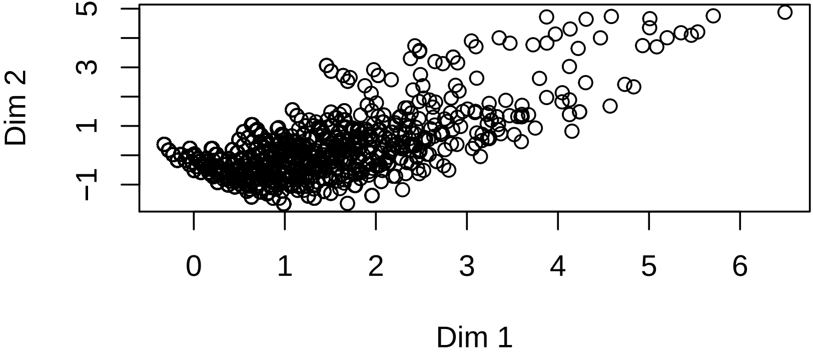
Extract PCs from MCA

```r
library(FactoMineR)
res.mca <- MCA(hsub)
```



```r
huiPCs <- res.mca$ind$coord
```

```r
plot(huiPCs[,1:2])
```

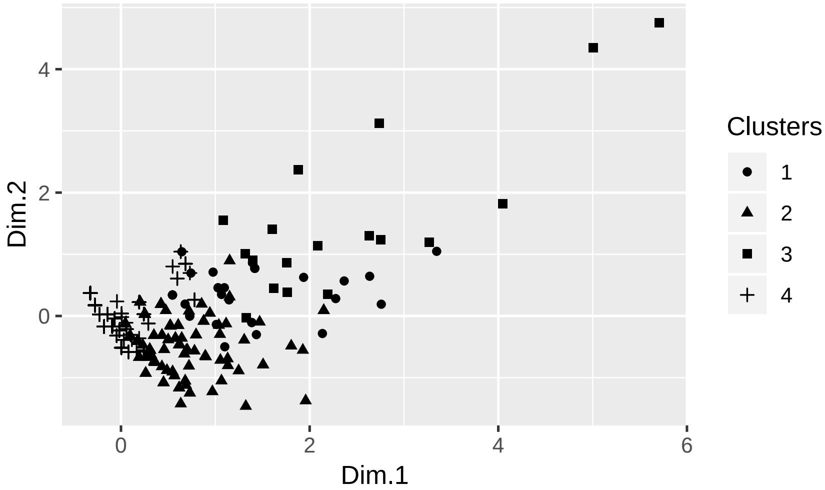


Cluster on PCs

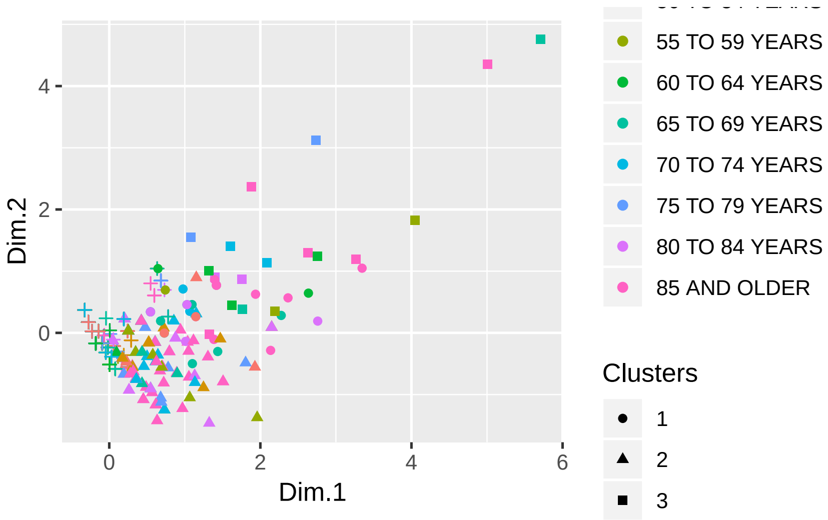
- Will work with a small subset (first 1000 people) to keep computation and overplotting down.

```
n <- 1000
hk <- kmeans(huiPCs,centers=4,nstart=10)
# plot for a small sample of the dataset
huiPCs <- data.frame(huiPCs[1:n,])
huiPCs$Clusters <- factor(hk$cluster[1:n])
huiPCs$age <- hui$DHHGAGE[1:n]
huiPCs$sex <- hui$DHH_SEX[1:n]
```

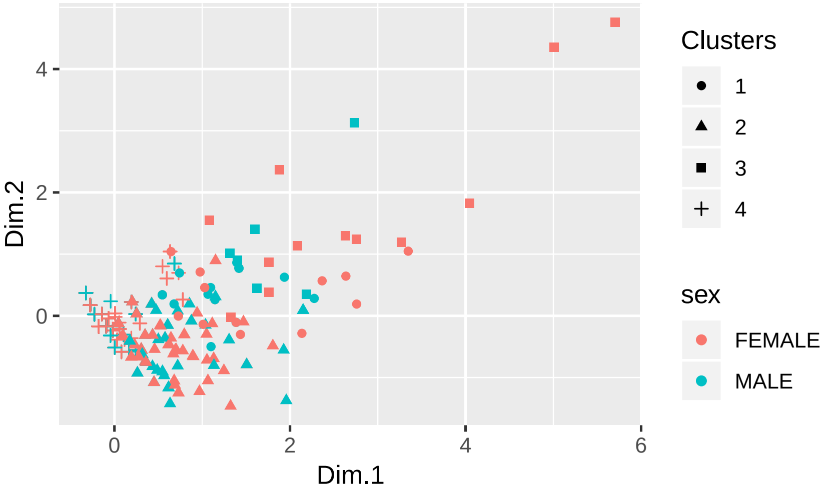
```r
library(ggplot2)
ggplot(huiPCs,
       aes(x=Dim.1,y=Dim.2,shape=Clusters)) +
  geom_point()
```



```
ggplot(huiPCs,
       aes(x=Dim.1,y=Dim.2,color=age,shape=Clusters)) +
  geom_point()
```



```
ggplot(huiPCs,
       aes(x=Dim.1,y=Dim.2,color=sex,shape=Clusters)) +
  geom_point()
```


Dim.2
Dim.1
Clusters
● 1
▲ 2
■ 3
sex
● FEMALE
● MALE