**FUNDAÇÃO GETÚLIO VARGAS**

**Business Analytics – Big Data**

**Prof João Fernando Serrajordia Rocha de Mello**

**Disciplina: Modelagem Estatística Avançada – Trabalho substitutivo**

**Grupo: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Para as questões que seguem, trabalharemos com os dados HBAT.**

1. **Sobre *outliers* em análise de agrupamentos, marque a alternativa falsa - apenas uma (1,0 pt).** 
   1. **( V )Em métodos hierárquicos, podem acabar restando como um agrupamento de uma única observação.**
   2. **( F ) Em métodos não hierárquicos, não têm influência, pois embora puxem o centroide, o método possui robustez e o resultado final não é influenciado.**
   3. **( V ) Quando se detectam correlações lineares entre as variáveis, a distância de *Mahalanobis* é interessante para a detecção de *outliers* porque considera esse tipo de relação.**
   4. **( V ) Componentes principais pode ser utilizada para a detecção visual de *outliers* multivariados através da construção de gráficos em 2 ou 3 dimensões, plotando as 2 ou 3 componentes respectivamente.**
   5. **( V ) Pontos discrepantes devem ser cuidadosamente analisados antes de serem retirados da amostra. Pontos “longe” da média podem ser uma característica natural da distribuição de um grupo específico. V**
2. **Qual é o maior valor da distância de *Mahalanobis* no problema da HBAT, considerando a distância de cada observação com relação ao centroide geral (média geral)? (2,0 pts)**
   1. **( X ) 30.67**
   2. **( ) 25.34 1**
   3. **( ) 7.85**
   4. **( ) 17.15**
   5. **( ) 16.69**
3. **Sobre o tamanho amostral da HBAT e padronização de variáveis, marque a alternativa falsa: (2,0 pts)**
   1. **( F ) Grupos pequenos, representando 10% ou menos da população, são facilmente detectados com 100 observações, daí a escolha pelo tamanho amostral.**
   2. **( V ) É considerado adequado pois o pesquisador está interessado em grandes grupos para campanhas de marketing, que serão detectáveis por técnicas de agrupamentos.**
   3. **( V ) As variáveis são escalas de medidas com campo de variação semelhante, sendo razoável não utilizar padronização neste caso.**
   4. **( V ) Grupos com frequência menor que 10% podem aparecer como outliers na análise.**
4. **Sobre a estratégia de construção dos agrupamentos, marque a alternativa falsa (2,0pts):**
   1. **( V ) Métodos não hierárquicos funcionam bem para grandes amostras, mas requerem estratégias de definição do número de clusters.**
   2. **( V ) Métodos hierárquicos são bons para definir o número de clusters, e avaliar diversas possibilidades de agrupamentos em uma única análise, mas são difíceis de visualizar para grandes amostras.**
   3. **( F ) A distância de Mahalanobis não pode ser utilizada quando se tem uma estrutura de correlação linear natural nos dados, sendo nesses casos preferível utilizar-se a distância de Manhattan.**
   4. **( V ) Para validar o agrupamento, uma ideia interessante é observar os perfis dos grupos conforme variáveis que não foram utilizadas no processo.**
   5. **( V ) Se for conduzida uma análise fatorial no mesmo conjunto de dados, os escores fatoriais podem ser utilizados para interpretar os grupos com maior facilidade.**
5. **Aplicando o algoritmo hclust com distância euclidiana, excluindo-se as variáveis 11, 15 e 17, com método de aglomeramento pelo centroide aos dados da HBAT (3,0 pts):**
   1. **Quantos grupos se obtém ao se utilizar um ponto de corte h=1,2?**

**97 grupos**

* 1. **Em se utilizando a solução de 10 agrupamentos, qual a frequência de indivíduos no menor grupo?**

**1 indivíduo**

* 1. **Qual é o grupo com menor pontuação média para percepção de qualidade do produto?**

**É o grupo 9 com média 5 para percepção de qualidade do produto**

1. **Questão desafio (1,0 pts)**
   1. **Faça um gráfico de dispersão com as duas primeiras componentes principais considerando as variáveis X19 a X23 identificando os grupos através de cores diferentes no gráfico. Mencione as configurações utilizadas na análise de componentes.**

****

**################**

**### SCRIPT – R ###**

**################**

**################**

**### DIRETORIO ###**

**################**

**getwd()**

**setwd('/Volumes/KINGSTON/FGV/Clusters e matrizes')**

**###############**

**### LIBRARY ###**

**###############**

**library("dplyr")**

**library("ggplot2")**

**library("corrplot")**

**library("psych")**

**#####################**

**### BASE DE DADOS ###**

**#####################**

**load('./Aula 04/hbat.Rdata')**

**#################**

**### QUESTAO 2 ###**

**#################**

**q2 <- hbat[,7:19]**

**# Validacao da amostra**

**KMO(q2)**

**# Exclusao das variaveis que nao tiveram uma boa aderencia da amostragem**

**q2 <- select(q2,-x11,-x15,-x17)**

**distancia\_mahalanobis <- mahalanobis(q2, colMeans(q2),**

**cov(q2))**

**max(distancia\_mahalanobis)**

**#################**

**### QUESTAO 5 ###**

**#################**

**q5 <- q2**

**q5\_dist <- dist(q5, method = "euclidean")**

**cluster <- hclust(q5\_dist, method = "centroid")**

**plot(cluster)**

**abline(h = 1.2, col = "red", lty = 5)**

**cluster$height <- sort(cluster$height)**

**grupos <- cutree(cluster, h = 1.2)**

**table(grupos)**

**grupos <- cutree(cluster, k = 10)**

**table(grupos)**

**rect.hclust(cluster, k=10, border="red")**

**q5$grupo <- grupos**

**q5 %>%**

**group\_by(grupo) %>%**

**summarise(media = mean(x6)) %>%**

**arrange(media)**

**#################**

**### QUESTAO 6 ###**

**#################**

**q6 <- select(hbat,x19:x22)**

**head(q6)**

**pca <- prcomp(q6, scale. = T)**

**pca$rotation**

**pca$scale**

**componentes <- as.data.frame(pca$x)**

**componentes$grupo <- grupos**

**ggplot(componentes, aes(PC1, PC2)) +**

**geom\_point(aes(color = grupo)) +**

**theme\_minimal()**