**FUNDAÇÃO GETÚLIO VARGAS**

**Business Analytics – Big Data**

**Prof João Fernando Serrajordia Rocha de Mello**

**Disciplina: Modelagem Estatística Avançada – Trabalho em grupo**

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

**Questão 1 (2,5 pts):** Baseado no arquivo “A02E01 - PCA simulado.r” faça as seguintes modificações:

1. Altere o programa para que x4 tenha desvio padrão 100 e gere novamente o data frame w como a combinação dos objetos x1, x2, x3 e o objeto x4 modificado;

O comando x4<-rnorm(100,sd=100);sd(x4) gera um x4 com desvio padrão próximo a 100, porém não 100.

1. Calcule as matrizes de correlação para os objetos x e w, o que se conclui?
2. Faça os gráficos de matriz de dispersão para os objetos x e w. Visualmente, qual a principal diferença?
3. Realize a análise de componente principal para os objetos x e w utilizando matriz de covariâncias. Compare os desvios padrão das duas estruturas de componentes principais, qual a principal diferença observada? Compare as cargas fatoriais, e indique com uma frase a principal alteração observada.
4. Realize a análise de componentes principais para os objetos x e w, utilizando matriz de correlações. Compare os desvios padrão e as cargas fatoriais. O que se conclui?

**Questão 2 (2,5 pts):** Escolha um arquivo de imagem de sua preferencia (dica: se necessário, reduza a definição da imagem). Execute as tarefas abaixo:

1. Carregue a imagem em “.ppm” e carregue-a no R. Considere a matriz da cor verde;
2. Exiba a imagem correspondente à matriz da cor verde da imagem escolhida;
3. Crie um objeto com a análise SVD dessa matriz. Analisando a matriz D, quantas variáveis são necessárias para expressar a figura escolhida com pelo menos 80% da informação?
4. Faça imagens utilizando 1 variável do SVD, 5 e 10. Faça mais 3 a 5 imagens tentando identificar o menor número de variáveis que exiba a imagem com uma qualidade que você julgue adequada.
5. Quantas variáveis você acredita serem adequadas para exprimir a imagem?

**As questões se baseiam no estudo da Hbat sobre as preferencias de seus clientes (base “hbat.Rdata”). Neste estudo a Hbat selecionou 100 de seus clientes aleatoriamente indicando-lhes um questionário sobre suas preferencias, os quais foram devidamente respondidos. Os dados contidos na tabela são:**

|  |  |
| --- | --- |
| Ordem | Nome |
| 1 | Id |
| 2 | tipo.cliente |
| 3 | tipo.industria |
| 4 | porte.firma |
| 5 | Região |
| 6 | sistema.distr |
| 7 | qualidade |
| 8 | ecomerce |
| 9 | Suporte |
| 10 | solucao.reclam |
| 11 | propaganda |
| 12 | Linha |
| 13 | imagem.equipe |
| 14 | Preco |
| 15 | Garantia |
| 16 | embalagem |
| 17 | encomenda.cobranca |
| 18 | preco.flex |
| 19 | vel.entrega |
| 20 | satisf.cliente |
| 21 | prob.recomendar |
| 22 | prob.recompra |
| 23 | pct.compras |
| 24 | relac.futuro |

**As variáveis 1 a 6 são cadastrais, 7 a 19 procuram medir as percepções dos clientes e 20 a 24 buscam medir o “sucesso”da Hbat com estes clientes. O objetivo imediato é entender as variáveis de percepções dos clientes e sua estrutura de correlação.**

**Questão 3 (1 pts):** Etapa 3 – Geração de fatores (a): Avaliando a medida de adequação da amostra (MSA) das variáveis explicativas (X6 a X19), considerando o limite de 50% para o índice geral e por variável, marque a alternativa verdadeira:

(a) Todas as variáveis possuem bom índice MSA.

(b) inicialmente 3 variáveis apresentam MSA menor que 50%. Retirando-se sequencialmente as duas piores, as demais variáveis apresentam bom MSA.

(c) O MSA como sendo o quadrado médio amostral, indica boa discriminação com relação às variáveis resposta.

(d) No caso da HBAT este indicador não se mostrou conclusivo.

(e) Três variáveis devem ser retiradas considerando-se o critério de 50%.

**Questão 4 (1 pts):** Considerando os dados excluindo-se as variáveis rejeitadas no exercício anterior. Utilizando-se o critério de variância do fator > 1, quantos fatores a análise de autovalores indica que devem ser extraídos? Explique em uma frase qual a lógica de se utilizar este critério.

**Questão 5 (1 pts):** Mantendo os mesmos critérios de seleção e padronização do exercício anterior, gerando os fatores pelo método, utilizando método de rotação Varimax, marque verdadeiro ou falso para as afirmações abaixo:

(a) Possuem, em geral, cargas fatoriais que definem bem os fatores. Apenas uma variável possui expressão (carga maior que 50%) em mais de um fator;

(b) Cargas fatoriais representam a correlação da variável sendo explicada pelo fator. Ou seja, 70% ao quadrado fornece um R2 de aproximadamente 50%, daí o critério.

(c) O fator mais importante (com maior variância) é explicado principalmente por Solução de Reclamações, Preço Competitivo e Imagem da Equipe de Vendas.

(d) Não há variáveis com todas as cargas fatoriais baixas (<40%, ou seja, sem expressão relevante em nenhum fator).

(e) Carga fatorial negativa significa que a expressão da variável é muito baixa, portanto a variável preço deveria ser retirada da análise.

**Questão 6 (2,0 pts):** Na mesma base de dados do exercício anterior, realize uma análise fatorial exploratória, utilizando a matriz de correlações e método de rotação promax.

1. Compare as cargas fatoriais obtidas por este método e pelo método anterior. Qual método lhe parece mais adequado? Justifique.
2. Calcule a matriz de correlações entre os fatores obtidos com este método e com o método do exercício anterior. Considerando o objetivo de entender os efeitos destes fatores sobre um conjunto de variáveis resposta, qual dos dois resultados você escolheria? Por que? (responda suscintamente)