Trabalho final individual – MBA Big Data e Business Analytics – Métodos Matriciais e Cluster Analysis

**Nome: PATRICIA YOSHIE KISHI BUENO**

**Ao entregar este trabalho final no e-class Você estará declarando:**

**"Estou ciente das razões pelas quais estamos fazendo este trabalho de forma remota. Assim, declaro que utilizei somente formas éticas de resolver as questões e não consultei ninguém, de nenhuma forma, para respondê-las ”**

Escrevi para que Você leia, por favor!!!

**Instruções**:

* Este trabalho tem **5 páginas**, incluindo esta.
* Consulta a notas de aula, apontamentos, net, scripts etc. é permitida.
* **A consulta a colegas ou outras pessoas, qualquer que seja o meio, é proibida**.
* Responda as perguntas digitando as respostas neste arquivo, nos locais indicados. Se não for incômodo, coloque as respostas em vermelho, azul ou verde.
* Ao terminar, **não esqueça de colocar seu nome**. Verifique cuidadosamente se não esqueceu nenhuma resposta nos espaços indicados.
* Por favor, não cole os scripts do R, os outputs do R (a não ser os gráficos, caso solicitados) nem as tabelas de dados originais.
* **Salve como .pdf** e faça upload no ECLASS na caixa de **entrega de atividades**, na pasta correspondente. Trabalhos entregues em outro formato não serão aceitos
* Se tiver dificuldades ou dúvidas com o ECLASS, por favor, envie-me um mail [alsicsu@gmail.com](mailto:alsicsu@gmail.com). Veja que a data de entrega é **até as 18 horas do dia 23/02/2021** (dentro do prazo regulamentar fixado pela FGV).
* Não envie as respostas por e mail ou WhatsApp, por favor (não receberei).

**Questão 1:** Considere a planilha CARS32. (sugestão: salve o arquivo fornecido pelo professor salve-o como **cc**, cc= CARS32).

* Trabalhe somente com as variáveis quantitativas fornecidas pelo Professor.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Weight | Length | Braking | Cylinders | Displacement | City | Highway | GHG |

* Não elimine variáveis nem observações mesmo que Você ache que seria melhor
* Não transforme as variáveis (exceto a padronização solicitada)

Determine as componentes principais das variáveis quantitativas após padronizá-las. Selecione o **menor** número de componentes que expliquem **pelo menos 95%** da variância total.

a) Quantas componentes principais Você selecionou? Resp: **4**

b) Qual a correlação da variável *weight* com a segunda componente principal? Resp: **0.111**\_\_\_\_\_\_ (3 casas decimais)

c) Padronize as duas primeiras componentes principais utilizando o comando scale. Utilize-as para agrupar as observações em dois clusters utilizando o método kmeans (slides) com set.seed(11) e nstart=25. Dê o tamanho de cada cluster.

Resp: **17** , **15**

d) Calcule a matriz de correlações **R** das oito variáveis quantitativas da planilha CARS32. Determine os valores dos dois maiores autovalores de **R** (sugestão: para evitar notação científica utilize antes o comando options(scipen = 999).

Resp: **6.010** **1.067** (3 casas decimais)

e) Determine o primeiro elemento da diagonal principal da matriz inversa de R.

Resp: **9.777** (3 casas decimais)

continua

f) Considere o vetor x:

x=c(-0.17413241, -0.36753180, 0.01375612, -0.68214295, -0.52446332, -0.57510432,

-0.79494180, 0.32461643)

*(evite digitar os valores ; melhor copiar as duas linhas juntas e colar diretamente no script do R)*

*(para conferir veja se sum(x) = -2.779944 )*

x é um auto vetor de R? por que (explique de forma objetiva e sucinta, dê os valores que justificam sua resposta: (se utilizar mais de 3 linhas provavelmente Você errou!)

(sim/não)\_**Sim**\_\_

justificativa: **Pois, R.x = λ, onde os valores de x e λ são paralelos. Ou seja, R %\*% x = kx, sendo k o auto valor.**

Em caso positivo, qual o auto valor correspondente?

Resp: **0.453** (3 casas decimais)

**Questão 2:** Considere apenas as variáveis as seguintes variáveis quantitativas da planilha cars32 (arquivo enviado em anexo).

"weight" "Length" "Braking" "Cylinders" "City"

Padronize-as utilizando a função **scale** e utilize-as como drivers para agrupar as observações.

* Não elimine observações mesmo que lhe pareçam outliers.
* Não transforme as variáveis (exceto a padronização solicitada)
* Não elimine variáveis mesmo que correlacionadas com outras

a) determine a distância de Manhattan entre as duas primeiras observações (CAR\_1 e CAR\_2) no arquivo dado, considerando os dados padronizados.

Resp: **5.625** (3 casas decimais)

b) Utilize a matriz de distâncias calculada com a função dist. (Não utilize daisy). Utilizando o método hierárquico Ward.D2 determine 2 clusters. Dê o número de elementos de cada cluster . Cluster 1 elementos= **20** Cluster 2 elementos= **12**

c) dê os códigos (em ordem alfabética) de 4 carros que pertencem ao **menor** cluster

Resp: **CAR\_2 CAR\_8 CAR\_11 CAR\_13**

continua

d) Com auxílio de boxplots compare os clusters. Qual apresenta maiores valores de *weight*? (não cole a figura). Dê o valor da mediana de weight neste cluster (valor aproximado via figura)

cluster **1** (na ordem respondida no item b)

mediana de weight (valor não padronizado) neste cluster **4000**

e) Agrupe as observações com os valores padronizados em 3 clusters, utilizando pamk. Utilize a matriz de distâncias calculada com a função dist. (Não utilize daisy). Rode utilizando o set.seed(11) .

Tamanhos dos clusters: **11 13 8**

f) Descreva **verbalmente** as diferenças entre os clusters obtidos no item anterior considerando apenas a variável *Displacement*. (sugestão: faça o boxplot e analise). Resposta **sucinta** e **clara** (digite a seguir). No máximo 3 linhas, utilizando texto tamanho 11. Não cole os boxplots, por favor

**A variável *Displacement* não foi considerada como driver.**

(não lerei mais que 3 linhas)

**Questão 3**: considere a matriz binária de transações na planilha PRODUTOS (consta do arquivo em Excel da prova).

a) Gere uma tabela de dupla entrada para**{ A,B}  {F}** e calcule (utilizando as fórmulas) o suporte, a confiança e o lift

|  |  |  |  |
| --- | --- | --- | --- |
|  | **{F}** | Não **{F}** |  |
| **{ A,B}** | **5** | **3** | **8** |
| **Não { A,B}** | **3** | **1** | **4** |
|  | **8** | **4** | **12** |

Suporte= **0.41** confiança= **0.62**  lift= **0.63** φ= **0.11**

(dê as respostas com pelo menos 2 casas decimais)

b) Utilize o pacote **arules** para determinar as regras de associação com suporte maior ou igual a 20% e confiança maior ou igual a 80%, contendo no mínimo dois produtos e no máximo quatro produtos Não esqueça de eliminar a coluna 1 (TID). Quantas regras foram geradas?

Resp: **104**

c) verifique no output o suporte, a confiança e o lift para a regra de associação seguinte

**{B,H,L} => {J}**

Suporte= **0.20** confiança= **0.80** lift= **2.28**

**(dê as respostas com 2 casas decimais)**

d) Eliminando as regras redundantes, quantas regras não redundantes sobraram?   
Resp: **49**.

e) Gere regras de associação em que o rhs (consequente) seja apenas o produto B, com suporte maior ou igual a 20% e confiança maior ou igual a 60% (cuidado: diferente do item b), contendo no mínimo dois produtos e no máximo quatro produtos.

Quantas regras foram geradas? **14**.

Qual a regra com **maior** confiança? **{J,L} => {B}**.

Qual seu lift? **1.38** (2 casas decimais)