Estatística Aplicada
EAD Ao Vivo

Tema da aula

Estudo de Caso

Segmentação – Cartão de Crédito

09/12/2020

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseada em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

ESTATÍSTICA APLICADA EAD AO VIVO

Professora:
Drª Karin Ayumi Tamura

Coordenadores:
Profª Drª Alessandra de Ávila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Corpo Diretivo

COORDENADORES DO LABDATA | ATUAÇÃO ACADÊMCIA E PROFISSIONAL

Profª Dra.
Alessandra Montini

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Têm muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em estatística aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Membro do Conselho Curador da FIA, Coordenadora de Grupos de Pesquisa no CNPQ, Parecerista da FAPESP e Colunista de grandes Portais de Tecnologia.

linkedin.com/in/alessandramontini/

Prof. Dr.
Adolpho Walter Canton

Diretor do LABDATA-FIA. Consultor em Projetos de *Analytics*, *Big Data* e Inteligência Artificial. Professor FEA – USP. PhD em Estatística Aplicada pela *University of North Carolina at Chapel Hill*, Estados Unidos.

Currículo - Profª Drª Karin Ayumi Tamura

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

Profª Dra.
Karin Ayumi Tamura

Contato: karin.tamura@fia.com.br

- **FORMAÇÃO ACADÊMICA:** Pós-doutora (2015), Doutora (2012), mestre (2007) e bacharel (2003) em Estatística pelo Instituto de Matemática e Estatística da USP, tendo como área de pesquisa modelos de regressão, análise multivariada de dados e algoritmos de *machine learning*.
- **ATUAÇÃO PROFISSIONAL:** Foi *Head* de *Analytics* por 14 anos, e atualmente é Conselheira Executiva e *Head* de Inovação na *Marketdata Solutions*, uma empresa do grupo WPP, e Professora Doutora no LABDATA FIA.
- **HISTÓRICO:** Atuação no mercado por 17 anos, com experiência profissional no segmento bancário (Bradesco) e consultoria *(Marketdata Solutions)*. Atuou como docente em cursos de pós-graduação (2010-15) no LABDATA FIA e ABEMD. Especialista em Estatística e *Advanced Analytics* trabalhando em projetos de diversos segmentos do mercado. Participante de congressos nacionais e internacionais voltados a área de Estatística, Dados e Algoritmos de *Machine Learning*.

"Tenho duas paixões no meu trabalho: dados e pessoas. Voltar a lecionar no LABDATA FIA está sendo a realização de um sonho planejado desde a minha época de aluna de pós-graduação. Meu objetivo como professora é integrar a visão do mercado com as técnicas e tecnologias de análise de dados, por meio de uma atuação humanista no ensino aos alunos"

Projetos atendidos

Conteúdo Programático do Curso

21 AULAS AO VIVO COM PROFA. KARIN | 27 PLANTÕES AO VIVO COM PROF. STEPHAN, 7 LISTAS DE EXERCÍCIOS E EAD VIDEO AULA EM PYTHON

Dia	Mês	Aula	EAD Ao Vivo	Plantão Prof. Stephan
5	Agosto	Introdução ao Curso e Análise Exploratória de Dados	Aula Prof. Karin	06/ago
12	Agosto	Análise Exploratória de Dados	Aula Prof. Karin	13/ago
19	Agosto	Análise Exploratória de Dados - Introdução ao R	Aula Prof. Karin	20/ago
26	Agosto	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	27/ago
2	Setembro	Regressão Linear Simples	Aula Prof. Karin	03/set
9	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	10/set
16	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	17/set
23	Setembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	24/set
30	Setembro	Análise de Cluster	Aula Prof. Karin	01/out
7	Outubro	Análise de Cluster	Aula Prof. Karin	08/out
14	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	15/out
21	Outubro	Arvore de Decisão	Aula Prof. Karin	22/out
28	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	29/out
4	Novembro	Regressão Logística	Aula Prof. Karin	05/nov
11	Novembro	Regressão Logística	Aula Prof. Karin	11/nov
18	Novembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	19/nov
25	Novembro	estudo de caso	Aula Prof. Karin	26/nov
2	Dezembro	estudo de caso	Aula Prof. Karin	30/dez
9	Dezembro	estudo de caso	Aula Prof. Karin	10/dez
16	Dezembro	Análise de Série Temporal - modelo auto regressivo	Aula Prof. Karin	17/dez
23	Dezembro	Lista de Exercícios em Sala de Aula (Frequência Liberada - véspera Natal)	-	-
Recesso Escolar		EAD - INTRODUÇÃO AO PYTHON EAD - INTRODUÇÃO AO PYTHON	EAD Video Aula (8 horas)	- -
6	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	07/jan
13	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	14/jan
20	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	20/jan
27	Janeiro	Introdução a Big Data - Aplicações de Machine Learning e Deep Learning	Aula Prof. Karin	28/jan
3	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	04/fev
10	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	11/fev
17	Fevereiro	Lista de Exercícios (Frequência Liberada - quarta de cinzas)	-	18/fev
24	Fevereiro	EXERCICIOS DE REVISÃO - EAD (19hs e 23hs - com presença obrigatória)	-	24/fev
3	Março	Prova (Plataforma On Line: 19hs e 23hs)	-	

Conteúdo da Aula

1. Método de Partição: K-médias
   i. Definição da quantidade ótima de grupos

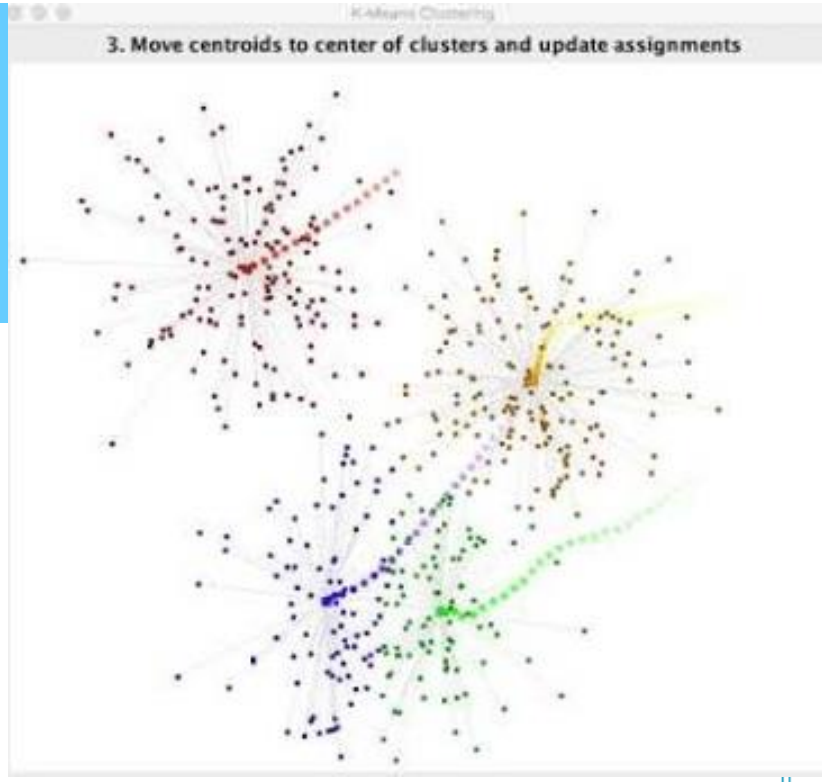
2. *Business Case:* Perfil de uso de cartão de crédito
   i. Validação dos dados na visão de negócio
   ii. Discussão de *missings* e *outliers*
   iii. Análise da relação entre as variáveis
   iv. Discussão e escolha das variáveis da Análise de Cluster
   v. Definição de quantidade de grupos
   vi. Discussão da escolha das técnicas de agrupamento
   vii. Definição dos grupos da visão de negócios
   viii. Descrição das "personas"

1. Método de Partição: K-médias

Método de Partição K-médias

É procedimento de aproximação, pois calcula-se a distância de cada elemento apenas para cada um dos k centroides (e não entre todas as observações do banco de dados), e por isso pode ser usado em grandes bancos de dados.

O número de *clusters* (k) precisa ser previamente definido.



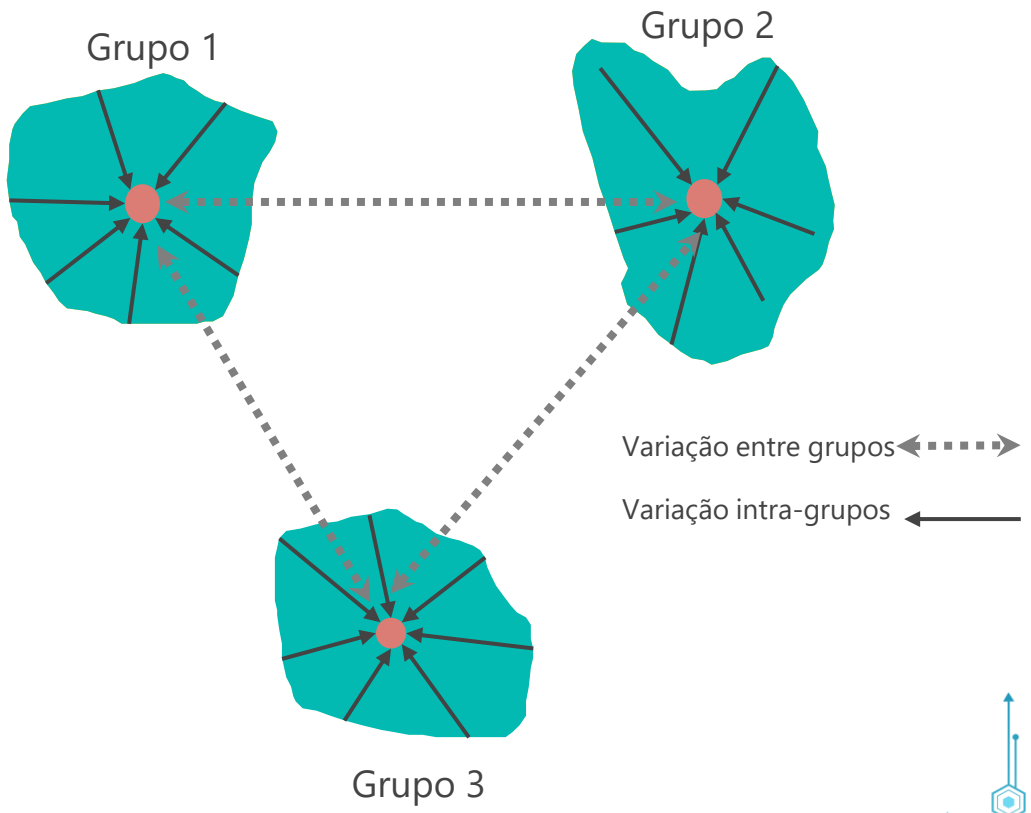
https://www.youtube.com/watch?v=nXY6PxAaOk0

Método de Partição: ***K-means* (K-médias)**

- O centroide de um grupo é definido como a média das distâncias de seus elementos.
- É um processo iterativo no qual, a cada passo, os elementos são agrupados no *cluster* com o centroide mais próximo, com subsequentes recálculos dos centros.
- As observações são agrupadas nos centroides até que as partições encontradas satisfaçam o critério de qualidade adotado.

Método de Partição K-médias

- O método K-médias utiliza um critério de qualidade do agrupamento baseado **Soma de Quadrados Dentro** do grupo (**WSS**, em inglês *Within Sum of Squares*), que mede a homogeneidade interna de cada um dos grupos e soma-se esse indicador de homogeidade para um determinado agrupamento formado por k grupos.
- Um bom agrupamento é que aquele que minimiza o WSS.
- Matematicamente, estamos em busca da quantidade de grupos em que a soma dos quadrados dentro dos clusters (WSS) seja mais próxima de 0.



Escolha da quantidade ótima de grupos

Para encontrar a quantidade ótima de grupos, podemos realizar simulações para vários tamanhos de grupos.

Na prática, quanto maior a quantidade de grupos, menor será o WSS, pois quanto mais grupos existirem, mas semelhantes os indivíduos serão entre eles (mais homogêneos), até se chegar o limite de cada observação analisada ser um grupo.

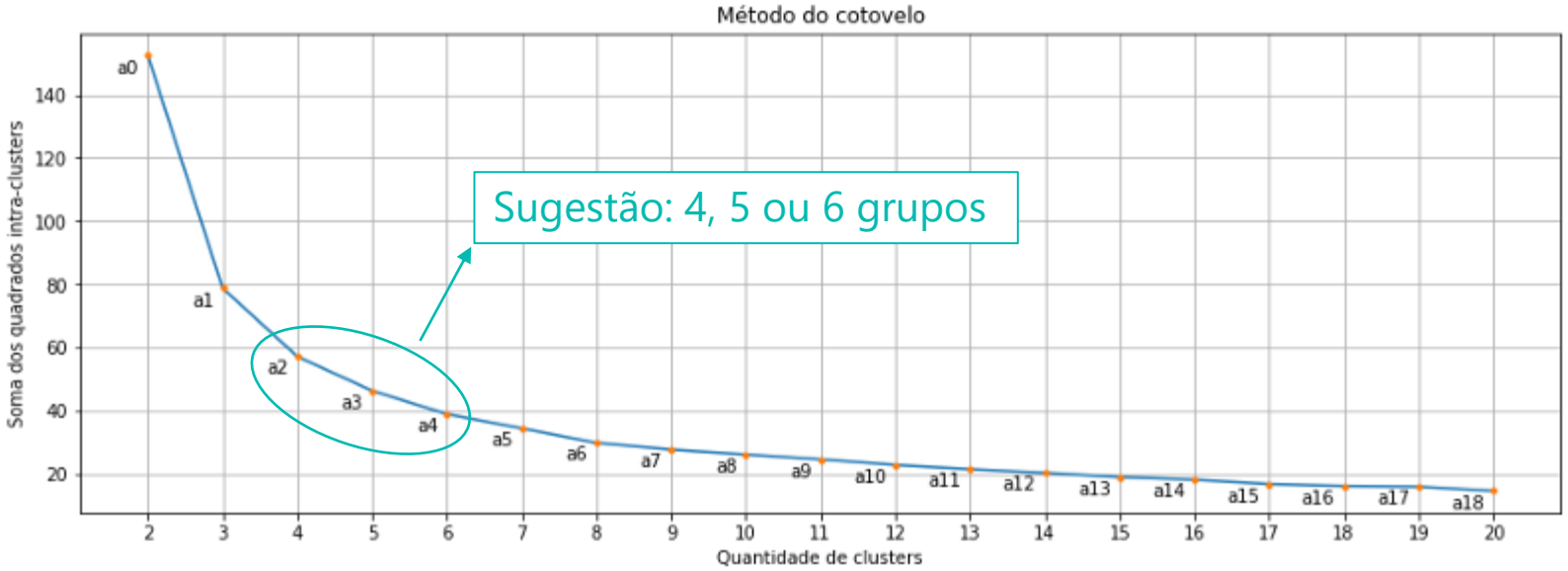
Mas como definimos o corte ideal? Qual a quantidade ótima de grupos?

Escolha da quantidade ótima de grupos

A partir de várias simulações de tamanhos de agrupamentos, à medida que aumentamos o tamanho dos grupos, a Soma de Quadrados Dentro (WSS) vai diminuindo.

A escolha do k ótimo será aquela em que o WSS começar a decair mais "lentamente", ou seja, não apresentando uma queda expressiva no WSS para um número maior de clusters.

Este procedimento realizado pela análise gráfica também é conhecido como **Método do Cotovelo (ou Elbow)**, pelo formato do gráfico sempre apresentar um forte decaimento do WSS para poucas quantidades de grupos e uma queda "lenta" para agrupamentos com maior quantidades de grupos.



https://jtemporal.com/kmeans-and-elbow-method/

Escolha da quantidade ótima de grupos – No R

Biblioteca necessária para realizar o método WSS

```
library(factoextra)

fviz_nbclust(dados, kmeans, method = "wss")
```

Método k-médias

Critério: Soma de quadrados dentro (WSS: *Within Sum of Squares*)

Onde a base de dados está armazenada (considerar apenas as variáveis que serão utilizadas na Análise de Cluster)

Case: Serviço de Entregas

Uma empresa de serviço de delivery de comida pronta tem o objetivo de fazer ações de relacionamento e reconhecimento com seus clientes. Para isso gostaria de identificar os perfis de clientes referentes à frequência de pedidos, o valor desses pedidos, a distância entre o estabelecimento-residência do cliente, e o tempo de entrega.

Fonte da Base: https://www.kaggle.com/asaumya/k-means-clustering-food-delivery-case-study/data

ID_Cliente	N_pedidos	Valor	Dist_m_restaurante	Tempo_m_entrega
1269647	212	138808	1,6	51
167631	211	56404	2,2	42
301524	189	36020	2,5	57
1268254	184	32489	3,1	55
357161	182	85150	2,4	36
1294857	171	55597	1,6	31
387095	168	19055	2,1	48
785080	160	39588	1,8	16

Vamos fazer juntos?

RStudio

Arquivo: Servico_entregas-Complemento.xls

Case: Serviço de Entregas

Uma empresa de serviço de delivery de comida pronta tem o objetivo de fazer ações de relacionamento e reconhecimento com seus clientes. Para isso gostaria de identificar os perfis de clientes referentes à frequência de pedidos, o valor desses pedidos, a distância entre o estabelecimento-residência do cliente, e o tempo de entrega.

Fonte da Base: https://www.kaggle.com/asaumya/k-means-clustering-food-delivery-case-study/data

Variável	Descrição
ID_Cliente	ID do Cliente
N_pedidos	Número de pedidos totais
Valor	Valor total dos pedidos (R$)
Dist_m_restaurante	Distância média entre o(s) restaurante(s) e o cliente (km)
Tempo_m_entrega	Tempo médio dos pedidos

Vamos fazer juntos?

RStudio

Arquivo: Servico_entregas-Complemento.xls

Exercícios: Serviço de Entregas

Uma empresa de serviço de delivery de comida pronta tem o objetivo de fazer ações de relacionamento e reconhecimento com seus clientes. Para isso gostaria de identificar os perfis de clientes referentes à frequência de pedidos, o valor desses pedidos, a distância entre o estabelecimento-residência do cliente, e o tempo de entrega.

Fonte da Base: https://www.kaggle.com/asaumya/k-means-clustering-food-delivery-case-study/data

(a) Realize a padronização e calcule a matriz de distâncias euclidianas entre os países.
(b) Considere o método Elbow, e decida a quantidade de grupos ótimo para rodar o K-médias.
(c) Rode o método K-médias com a quantidade de grupos indicado nos itens (b).

Vamos fazer juntos?

RStudio

Arquivo: Servico_entregas-Complemento.xls

2. Business Case

O objetivo do Business Case é utilizar todos os conceitos de **Análise de Cluster**, testando os diferentes métodos de agrupamentos e sugerindo um agrupamento que faça sentido para o negócio.

Na prática, não é dito quais variáveis serão testadas, e a escolha vai depender dos **objetivos de negócios** e das **análises preliminares dos dados**.

Durante a solução do case será realizada a **Análise Exploratória** com intuito de entender os campos não visão do negócio, tratar ***missings*** e ***outliers***.

Avaliaremos a **correlação entre as variáveis** para verificar se estamos escolhendo variáveis muito similares para calcular o critério de semelhança entre os indivíduos, pois mesmo fazendo a padronização se inserirmos variáveis "muito correlacionadas", é como se tivéssemos dando um "peso maior" para aquele grupo de características correlacionadas.

Também discutiremos as indicações do uso do **Método Hierárquico** vs **Método de Partição**.

Vamos discutir tudo isso neste *Business Case*. =)

Case: Marketing Cartão

Uma instuição financeira, emissora de Cartão de Crédito, deseja segmentar seus clientes para implementar estratégias de **atendimento** e **relacionamento diferenciados** de acordo com o uso do produto cartão de crédito. Dentre as variáveis transacionais disponíveis, quais poderiam ser escolhidas parar representar o "uso do cartão de crédito"? Quantos segmentos existem na carteira de clientes? A base disponibilizadas é uma amostra de aproximadamente 1% dos clientes da emissora de cartão.

Fonte Adaptada: https://www.kaggle.com/mirichoi0218/insurance

Cod_Cliente	LIMITE_DISP_T0	LIMITE_TOTAL_T0	PERC_USO_LIMITE_T0	PERC_FAT_CARTAO_12M	QTDE_TRANSACAO_T0	VALOR_FATURA_T0
1	40,900749	1000	0,959099251	0,833333333	2	201,802084
2	3202,467416	7000	0,542504655	0,916666667		4103,032597
3	2495,148862	7500	0,667313485	1	12	622,066742
5	817,714335	1200	0,318571388	1	1	678,334763
6	1809,828751	1800	-0,005460417	1	8	1400,05777
7	627,260806	13500	0,953536237	1	64	6354,314328
8	1823,652743	2300	0,207107503	1	12	679,065082
9	1014,926473	7000	0,855010504	1	5	688,278568

Cartao_Credito_BusinessCase.xlsx

Case: Marketing Cartão

Uma instuição financeira, emissora de Cartão de Crédito, deseja segmentar seus clientes para implementar estratégias de **atendimento** e **relacionamento diferenciados** de acordo com o uso do produto cartão de crédito. Dentre as variáveis transacionais disponíveis, quais poderiam ser escolhidas parar representar o "uso do cartão de crédito"? Quantos segmentos existem na carteira de clientes? A base disponibilizadas é uma amostra de aproximadamente 1% dos clientes da emissora de cartão.

Fonte Adaptada: https://www.kaggle.com/mirichoi0218/insurance

Variável	Descrição
Cod_Cliente	Código do cliente
LIMITE_DISP_T0	Valor do limite restante em seu cartão para fazer compras no último mês
LIMITE_TOTAL_T0	Valor do limite do cartão de crédito
PERC_USO_LIMITE_T0	Percentual do uso limite do cartão
PERC_FAT_CARTAO_12M	Percentual de meses que o houve faturamento do cartão de crédito
QTDE_TRANSACAO_3M	Quantidade de transações nos últimos 3 meses
VALOR_FATURA_T0	Valor da fatura no último mês

Cartao_Credito_BusinessCase.xlsx

Case: Marketing Cartão

Uma instuição financeira, emissora de Cartão de Crédito, deseja segmentar seus clientes para implementar estratégias de **atendimento** e **relacionamento diferenciados** de acordo com o uso do produto cartão de crédito. Dentre as variáveis transacionais disponíveis, quais poderiam ser escolhidas parar representar o "uso do cartão de crédito"? Quantos segmentos existem na carteira de clientes? A base disponibilizadas é uma amostra de aproximadamente 1% dos clientes da emissora de cartão.

Fonte Adaptada: https://www.kaggle.com/mirichoi0218/insurance

(a) Realize a análise exploratória univariada. Calcule as medidas resumos e construa boxplots e histogramas para todas as variáveis. Analise os resultados.
(b) Avalie a presença e quantifique os *missings* para cada variável. Seria possível tratar os *missings*? Se sim, trate-os segundo o contexto do negócio.
(c) Avalie a presença e quantifique os *outliers* para cada variável. Seria possível tratar os *outliers*? Se sim, trate-os segundo o contexto do negócio.
(d) Para a base "tratada" nos itens (b) e (c), realize a análise exploratória univariada novamente. Calcule as medidas resumos e construa boxplots e histogramas para todas as variáveis. Analise os resultados.
(e) Avalie a correlação entre as variáveis por meio Correlação de Pearson. Discuta a relação entre as variáveis e decida quais variáveis serão utilizadas para agrupar os clientes.
(f) Realize a padronização das variáveis.
(g) Dado as características da base de dados, qual método de agrupamento você adotaria? Discuta com a sala.
(h) Realize a análise de agrupamento com os 2 métodos hierárquicos, selecionando aleatoriamente 1.000 observações, pela análise do dendrograma, escolha um dos métodos e defina a quantidade de grupos.
(i) Realize a análise de agrupamento pelo método hierárquico K-médias e defina a quantidade de grupos.

Case: Marketing Cartão

Uma instuição financeira, emissora de Cartão de Crédito, deseja segmentar seus clientes para implementar estratégias de **atendimento** e **relacionamento diferenciados** de acordo com o uso do produto cartão de crédito. Dentre as variáveis transacionais disponíveis, quais poderiam ser escolhidas parar representar o "uso do cartão de crédito"? Quantos segmentos existem na carteira de clientes? Considere uma amostra de aproximadamente 1% dos clientes da emissora de cartão para o estudo

Fonte Adaptada: https://www.kaggle.com/mirichoi0218/insurance

(j) Compare a quantidade de grupos encontrados pelos métodos hierárquicos e K-médias.
(k) Considerando que a base de dados é "grande", realize o agrupamento dos clientes pelo K-médias utilizando k definido no item (j).
(l) Descreva as personas e justifique para área de negócios porque o agrupamento formado é adequado para implementar estratégias de **atendimento** e **relacionamento diferenciados**.

Referências

- Johnson, R. A. e Wichern, D. W. *Applied Multivariate Statistical Analysis*. Prentice-Hall Inc., 6th ed. 2007
- Timm, N.H. *Applied Multivariate Analysis*. Springer-Verlang, 2002