MBA
USP
ESALQ

Detecção de anomalias em gastos dos deputados estaduais com K-Means

Rodolfo Orlando Viana

Ana Julia Righetto

INTRODUÇÃO A verba de gabinete

Criada em 1997, a “verba de gabinete”, nome informal para Auxílio-Encargos Gerais de Gabinete de Deputado e Auxílio-Hospedagem, garante aos 94 parlamentares da Assembleia Legislativa de São Paulo o ressarcimento mensal de despesas inerentes ao mandato até o limite de 1.250 unidades fiscais do estado [Ufesp].

VALORES EM 2022

- Ufesp
  R$ 31,97
- Limite mensal por deputado
  R$ 39.962,50
- Total empenhado
  R$ 26.652.243,51

MBA USP ESALQ

INTRODUÇÃO Controle de gastos

Tendo origem nos cofres públicos, órgãos de controle como o Ministério Público do Estado não raro abrem procedimentos investigatórios para investigar eventual malversação no uso da verba de gabinete por parte de parlamentares.

EXEMPLOS

- 29.0001.0246360.2021-54
  Apura locação de imóveis de aliados políticos e nunca utilizados

- 0037174-14.2021.8.26.0000
  Apura ressarcimento de despesas nunca efetuadas

MBA USP ESALQ

INTRODUÇÃO Papel de ciência de dados

Técnicas de aprendizado de máquina podem auxiliar os órgãos de controle a detectar quais das despesas efetuadas são anomalias e devem ser objetos de escrutínio pormenorizado.

Neste trabalho, foi utilizado um algoritmo autoral de K-Means nos dados de alimentação e hospedagem de 2018 a 2022, com valores corrigidos pela inflação.

K-MEANS Definição

Em linhas gerais, K-Means é um algoritmo para clusterização e classificação. A técnica particiona um conjunto de dados X em k agrupamentos (clusters) não sobrepostos, sendo k um número pré-determinado.

Cada ponto de dado pertence ao agrupamento em que haja menor distância em relação ao centro do cluster (centroide).

O algoritmo busca minimizar a soma dos quadrados da distância dentro do cluster.

NOTAÇÃO

$$\sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2$$

onde,

k: número de clusters

S_i: cluster i

x: ponto de dado

μ_i: média da distância dos pontos em S_i

K-MEANS Visualização

Dado um conjunto de dados univariado, os pontos são distribuídos conforme seus valores (Figura 1).

Com a quantidade de clusters pré-determinada, são calculados os centroides a partir da minimização do quadrado das distâncias (Figura 2). Os pontos próximos aos centroides foram clusters (Figura 3).

Os pontos que não se encontram nos clusters são considerados anomalias (Figura 4).

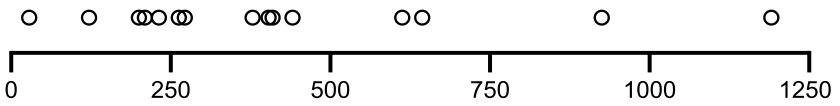


Figura 1

Figura 2

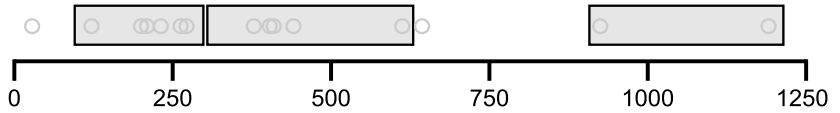


Figura 3

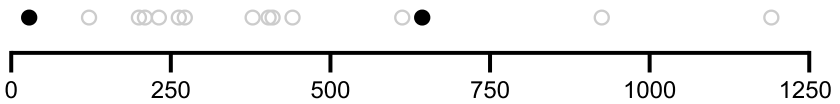


Figura 4

K-MEANS Desafios e soluções

DESAFIOS

1. Determinação da quantidade de clusters
2. Inicialização de centroides considerando mínimo global em vez de mínimo local
3. Critério para convergência ideal dos centroides
4. Validação dos resultados

SOLUÇÕES

1. Método do cotovelo
2. Método K-Means++
3. Comparação do movimento de centroides entre iterações
4. Método da silhueta; índice de Davies-Bouldin

K-MEANS Método do cotovelo

O método do cotovelo executa K-Means múltiplas vezes, iterando sobre valores para k e calculando a soma dos quadrados das distâncias entre pontos e centroide. Quanto maior o valor de k, menor a soma.

Em determinado momento, a diferença se tornará marginal. Graficamente, forma-se um "cotovelo“ (Figura 5). O ponto em que essa estabilização se torna perceptível representa uma estimativa do número ideal de clusters.

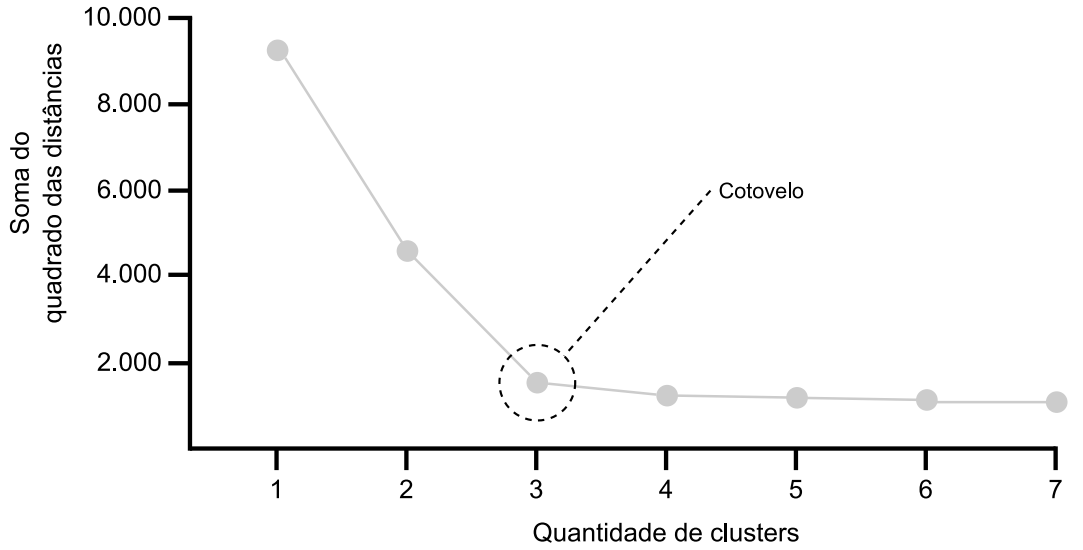


Figura 5

K-MEANS Método K-Means++

Após a determinação do número ideal de clusters, utilizou-se o método de inicialização K-Means++.

Nele, o centroide de cada cluster passa por iterações para definição de onde ele deve se posicionar. O ponto escolhido decorre da probabilidade de determinado ponto ser o melhor centroide com base na sua distância.

ETAPAS

1. Escolha aleatória de um centroide.
2. Cálculo das distâncias de cada ponto em relação ao centroide escolhido.
3. Seleção de um ponto para ser o próximo centroide a partir da probabilidade proporcional ao quadrado da distância em relação ao centroide anterior.
4. Repetição das etapas 2 e 3 até que k centroides sejam escolhidos.

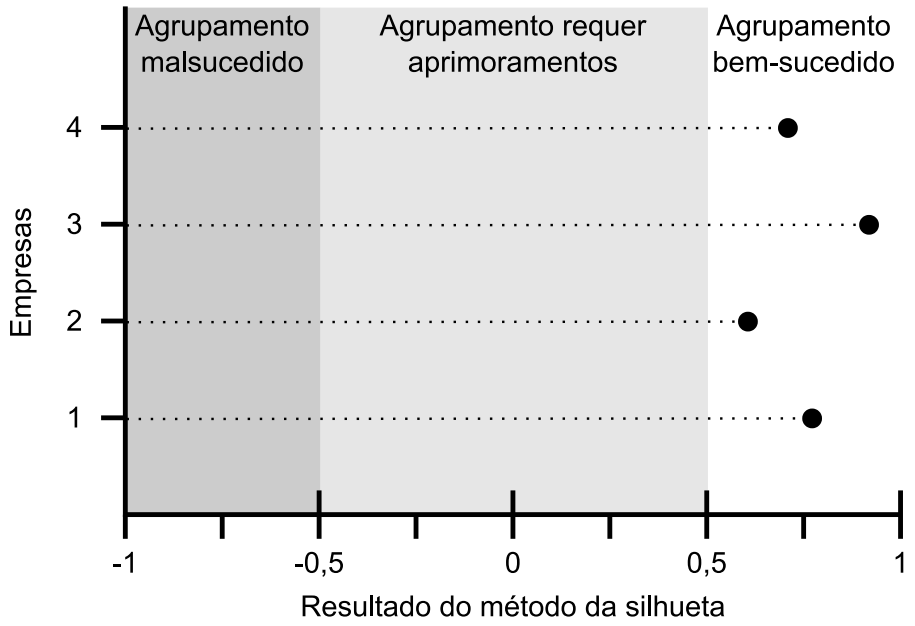
K-MEANS Comparação de movimento de centroides

Entre as iterações em K-Means++, o algoritmo compara a movimentação dos centroides, e converge apenas quando a diferença entre iterações é inferior ao limite estabelecido para inércia, de 0,0001.

VALIDAÇÃO Método da silhueta

Para validar a escolha dos centroides e a clusterização, uma das ferramentas adotadas foi o método da silhueta, que observa a similaridade de um ponto com seu cluster em comparação com outros clusters, e retorna resultados no intervalo de -1 a 1 (Figura 6).

Figura 6



NOTAÇÃO

$$\frac{b_i - a_i}{\max(a_i, b_i)}$$

onde,

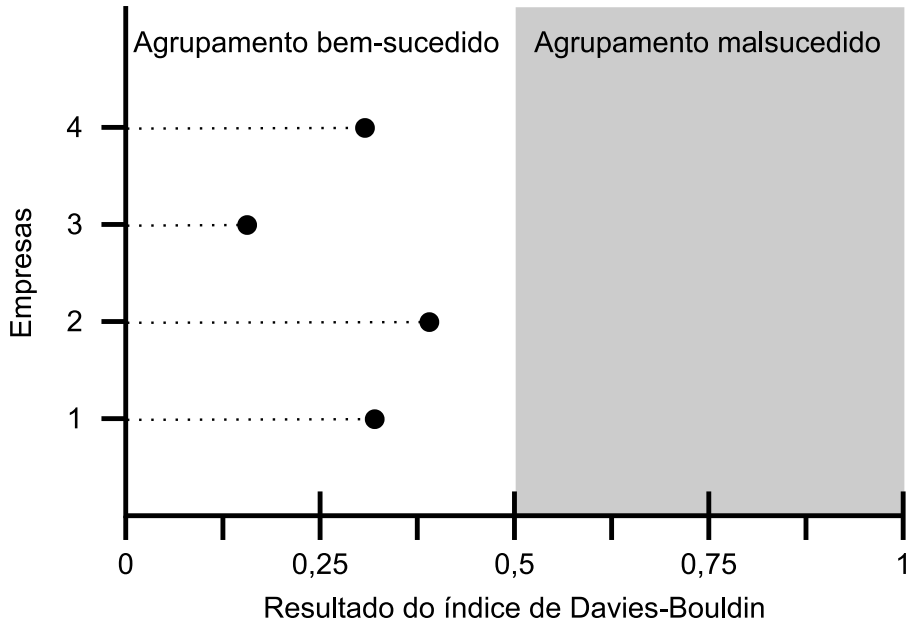
a_i: distância média de i para pontos intracluster

b_i: distância média de i para pontos extracluster

VALIDAÇÃO Índice de Davies-Bouldin

A segunda ferramenta adotada para validar os resultados foi o índice Davies-Bouldin, que observa a coesão do cluster, dada a lógica de que um agrupamento adequado é denso em si, ao passo que distante dos demais. Seus resultados vão de 0 a 1 (Figura 7).

Figura 7



NOTAÇÃO

$$\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

onde,

i, j: clusters distintos

S: dispersão interna

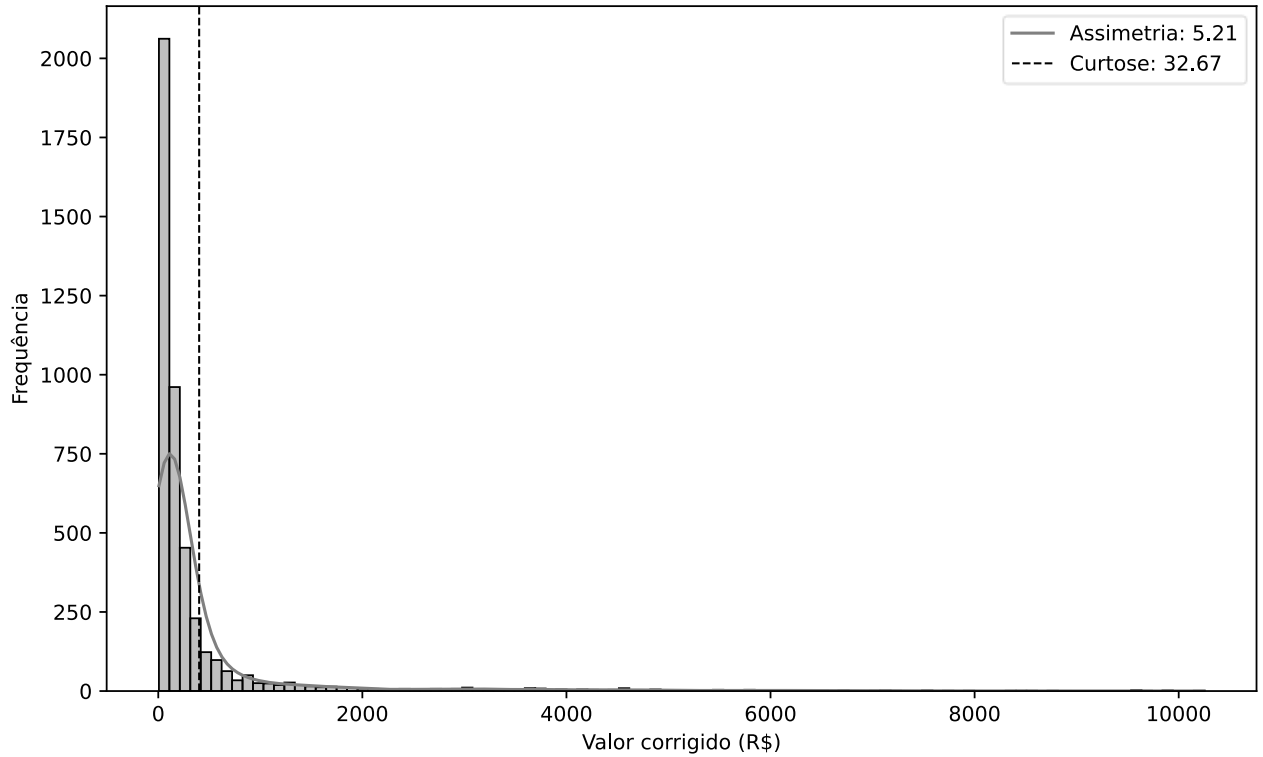
M_{ij}: distância entre clusters

RESULTADOS Estatísticas dos dados

No quinquênio observado, foram 4.453 registros de despesas em 86 números únicos de CNPJ, com valor médio de R$ 400,76; porém, com desvio-padrão elevado — coeficiente de variação de 241,41%.

O conjunto apresenta cauda à direita mais longa e pico acentuado em comparação à distribuição normal (Figura 8).

Figura 8



RESULTADOS Descobertas do algoritmo

Com a aplicação do algoritmo, foram obtidas 262 anomalias que somaram R$ 197.697,24 — 11,08% do valor total de despesas.

Do conjunto de 86 empresas, todas apresentam resultados ideais para o método da silhueta (valores entre 0,577 e 0,918); 79 apresentaram resultados ideais para o índice de Davies-Bouldin (valores entre 0,166 e 0,489), enquanto sete apresentaram resultados abaixo do ideal (valores entre 0,508 e 0,573).

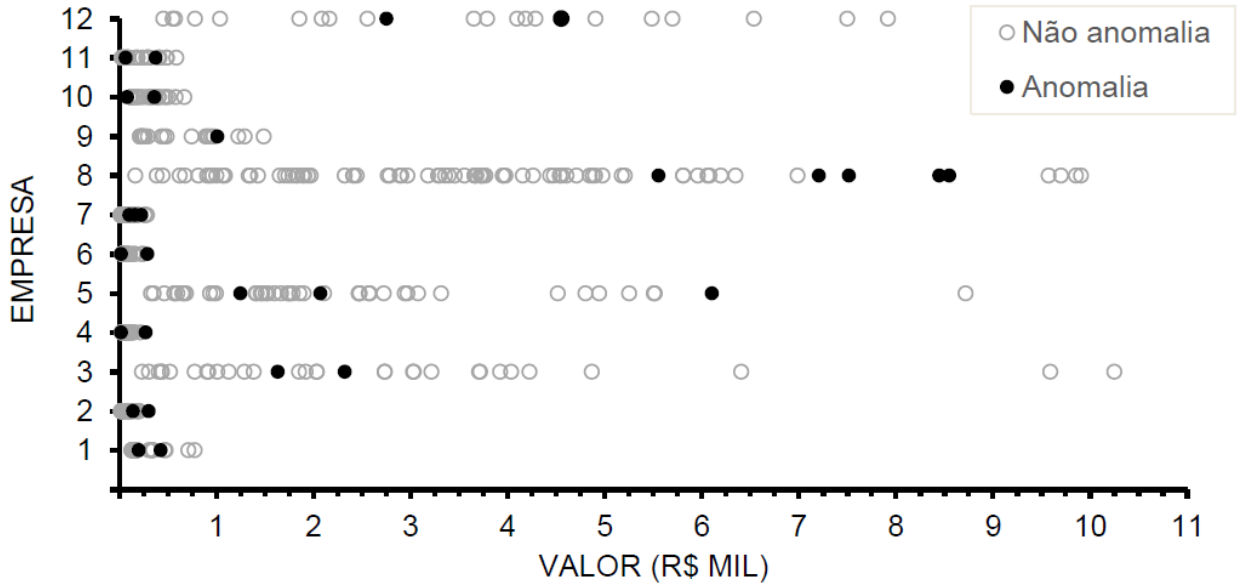
Em suma, a clusterização foi bem executada.

RESULTADOS Discussão sobre anomalia

Anomalias, no contexto deste trabalho, são valores de despesas que não se enquadram nos agrupamentos criados pelo algoritmo. Algumas se posicionam no meio de todas as despesas de determinada empresa, não sendo os maiores valores no conjunto de despesas (Figura 9).

São, portanto, falsos positivos.

Figura 9



RESULTADOS Números finais

Para descartar falsos positivos, foram consideradas anomalias passíveis de inquirição dos órgãos de controle somente aquelas cujos valores são maiores que o maior valor de não anomalia do último cluster.

Tal critério levou ao resultado de 46 anomalias em 32 empresas, com valor total de R$ 44.348,88.