



Aprendizado de Máquina II

# *Aplicação de Algoritmos de Clusterização em uma Base de Dados de Reservas de Hotéis*


Autores: Pedro Alexandre de Araújo Aguiar, Clodomir Joaquim de  
Santana Junior, Carmelo José Albanez Bastos Filho

Grupo: Anna Luiza Alves e Gustavo Costa



## Aplicação de Algoritmos de Clusterização em uma Base de Dados de Reservas de Hotéis

*Application of Clustering Algorithms in Hotels Reservation Datasets*

**Pedro Alexandre de Araújo Aguiar<sup>1</sup>**  [orcid.org/0000-0002-9973-763X](https://orcid.org/0000-0002-9973-763X)

**Clodomir Joaquim de Santana Junior<sup>1</sup>**  [orcid.org/0000-0001-7869-7184](https://orcid.org/0000-0001-7869-7184)

**Carmelo José Albanez Bastos Filho<sup>1</sup>**  [orcid.org/0000-0002-0924-5341](https://orcid.org/0000-0002-0924-5341)

<sup>1</sup>Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Pernambuco, Brasil.

E-mail do autor principal: [paaa@ecomp.poli.br](mailto:paaa@ecomp.poli.br)

### Resumo

---

Este artigo faz uma análise da aplicação dos algoritmos de clusterização K-Means e Fuzzy C-Means. O estudo de caso visa identificar perfis de clientes de uma agência de viagens online, com o objetivo de melhorar a eficácia do envio de ofertas através de e-mail marketing, possibilitando o envio de anúncios personalizados para cada perfil. O processo de clusterização foi feito baseado na similaridade entre os usuários, levando em conta 13 características extraídas das vendas dos clientes. O resultado mostra que, apesar de chegarem a grupos parecidos, o K-Means teve desempenho levemente superior ao Fuzzy C-Means, no que diz respeito a avaliação através da métrica de estatística Gap.

**Palavras-Chave:** Clusterização; K-Means; Fuzzy C-Means;

# *Introdução e Objetivos do Trabalho*

---

- Introdução sobre clusterização e seu objetivo de segmentar dados com base em similaridades.
  - **Problema Abordado no Artigo:** Custos elevados com campanhas de e-mail marketing sem segmentação para clientes de uma agência de viagens online.
  - **Propósito do Artigo:** Identificar perfis de clientes de uma agência de viagens online para melhorar a eficácia do envio de ofertas.
  - Comparativo dos Algoritmos K-Means e Fuzzy C-Means
-

# Metodologia

## Base de Dados

- **Base de Dados:** 2.959 vendas aprovadas de reservas de hotéis de uma agência de viagens online no Brasil, realizadas entre 2016 e 2017.
- **Características Utilizadas:** 13 características extraídas das vendas dos clientes.

Característica	Descrição
Uf	Estado onde o cliente reside
Idade	Idade do clientes
Total Comprado	Soma total de todas as compras do cliente
Ticket Médio	Valor médio de compra do cliente
Quantidade de Compras	Quantidade de compras do cliente
Tipo de Pagamento	Forma que o cliente pagou a compra (Cartão de crédito, boleto)
Quantidade de Parcelas	Quantidade de parcelas que o cliente escolheu dividir
Cidade	Cidade do cliente
Valor da venda	Valor da venda específica
Hotel	Hotel comprado pelo cliente
Destino	Cidade de destino da viagem
Mês de estadia	Mês que o cliente escolheu se hospedar
Quantidade de diárias	Total de dias que o cliente irá ficar hospedado



# Metodologia

## Pré-processamento dos Dados

- Utilização da ferramenta Open Refine.
- Idade: Extraída da data de nascimento.
- Cidade: Padronização dos nomes
- Destino: Unificação de destinos que se referiam ao mesmo local (Ex: Ipojuca/PE e Porto de Galinhas unificados para Porto de Galinhas).
- Mês de Estadia: Derivada do dia de entrada no hotel.

Característica	Descrição
Uf	Estado onde o cliente reside
Idade	Idade do clientes
Total Comprado	Soma total de todas as compras do cliente
Ticket Médio	Valor médio de compra do cliente
Quantidade de Compras	Quantidade de compras do cliente
Tipo de Pagamento	Forma que o cliente pagou a compra (Cartão de crédito, boleto)
Quantidade de Parcelas	Quantidade de parcelas que o cliente escolheu dividir
Cidade	Cidade do cliente
Valor da venda	Valor da venda específica
Hotel	Hotel comprado pelo cliente
Destino	Cidade de destino da viagem
Mês de estadia	Mês que o cliente escolheu se hospedar
Quantidade de diárias	Total de dias que o cliente irá ficar hospedado



## **Algoritmos de Clusterização**

### **K-MEANS:**

- Algoritmo particional
- Popular por simplicidade e eficiência.
- Medida de similaridade: Distância Euclidiana.
- Atualização dos centroides através da média de cada característica.

### **Fuzzy C-Means (FCM):**

- Extensão do C-Means particional, utilizando lógica fuzzy.
- Permite que um elemento pertença a mais de um cluster, com graus de pertinência.
- Cálculo de centroides, distância euclidiana e atualização da matriz de pertencimento.

# *Metodologia*

## **Experimentos**



### **Estratégia**

- 30 execuções para cada algoritmo e para cada valor de K (número de clusters)
- K variando de 2 a 10 grupos
- Média de 4 métricas calculadas para cada valor de K
  - Estatística GAP, Distância Inter-Cluster, Erro Quantizado e Distância Intra-Cluster



# *Resultados K-Means*

---

- **Distância Intra-Cluster:** Diminuiu com o aumento de K, indicando grupos mais compactos.
  - **Distância Inter-Cluster:** Aumentou com o aumento de K, indicando grupos mais separados.
  - **Métrica Principal (Gap):** A Estatística Gap apontou K=3 como a quantidade ideal de grupos.
  - **Perfil Identificado:** Os 3 grupos foram definidos principalmente pelo "Mês de Estadia" (início, meio e fim de ano).
  - **Comportamento Esperado:** O algoritmo demonstrou o comportamento esperado nas métricas.
-

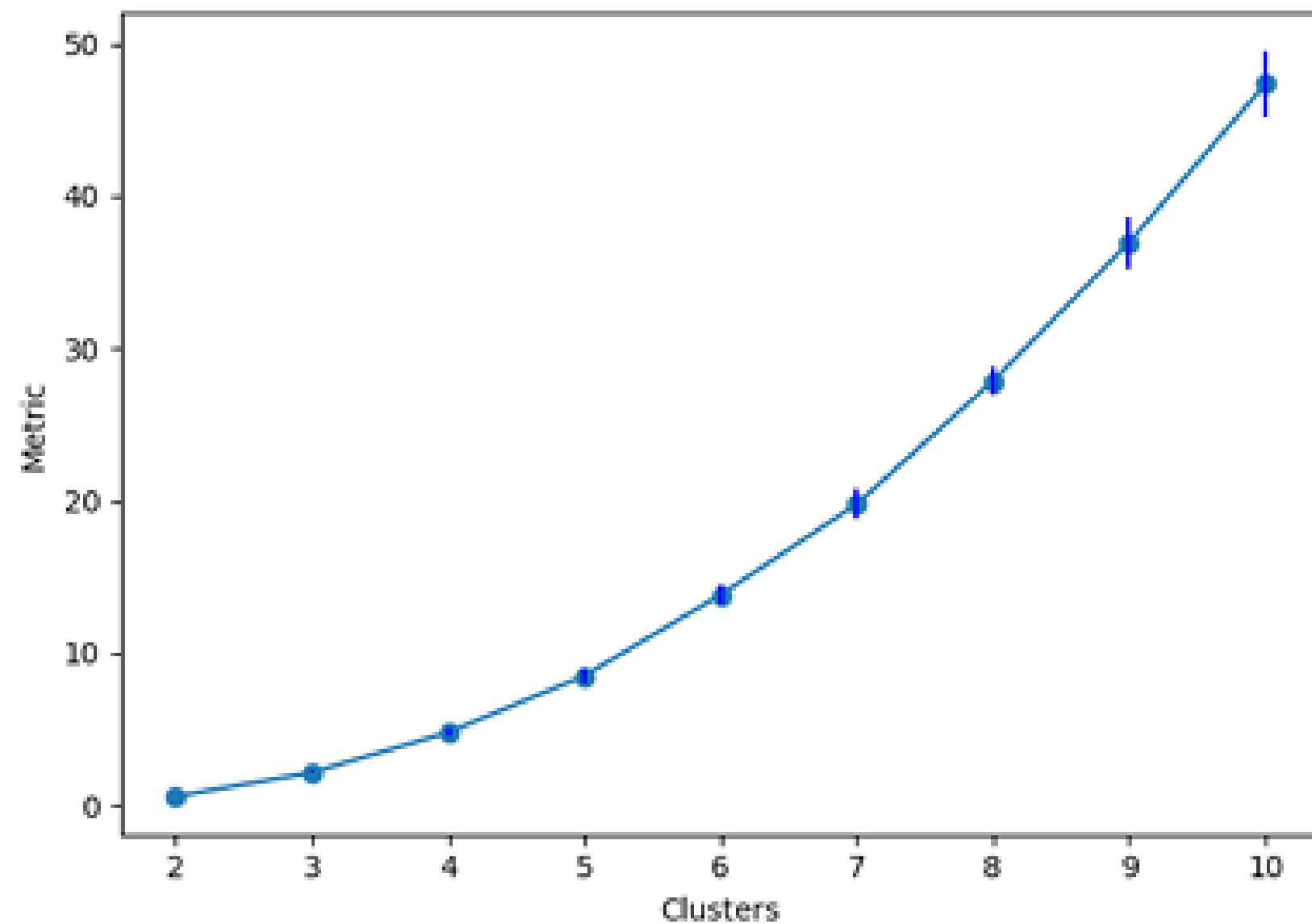


# Gráficos *K-Means*

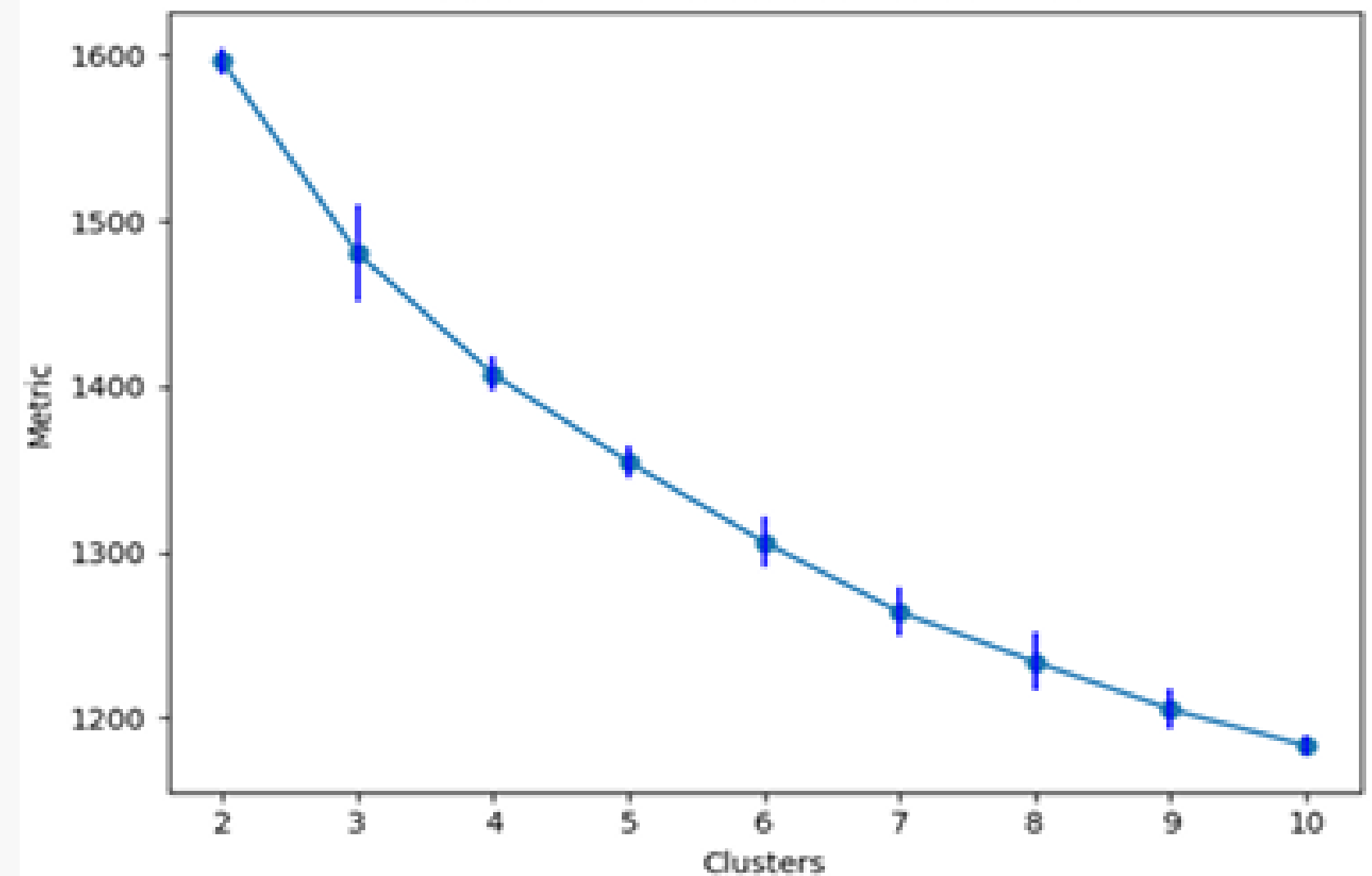
---

Métricas como as distâncias intra e inter-cluster, isoladamente, quase sempre indicarão que um número maior de  $K$  é melhor, tornando difícil encontrar o "número ideal" de grupos.

KMeans - Metric sumInterClusterDistance

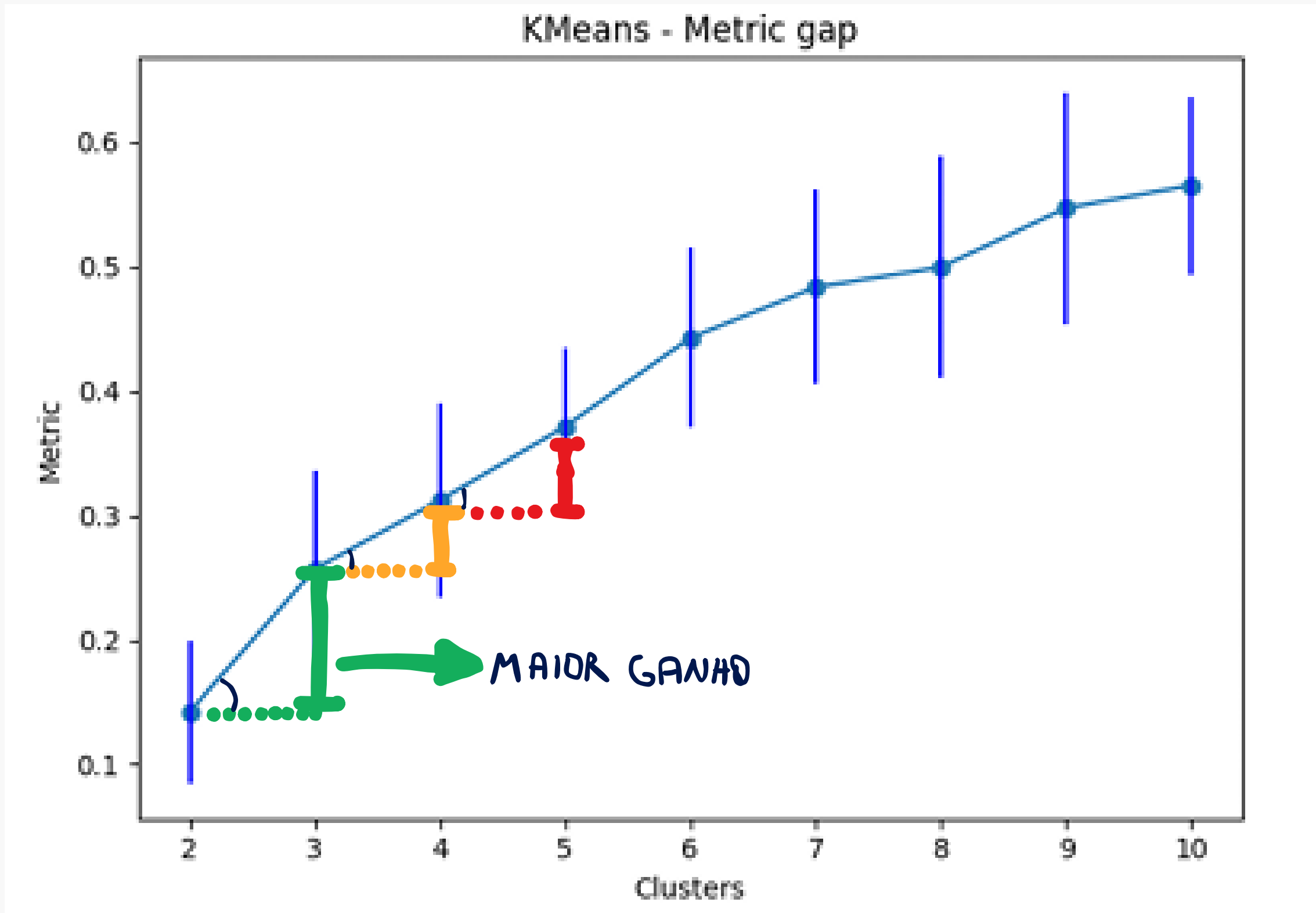


KMeans - Metric intraClusterStatistic



# Gráfico GAP K-Means

---



# *Resultados Fuzzy C-Means*

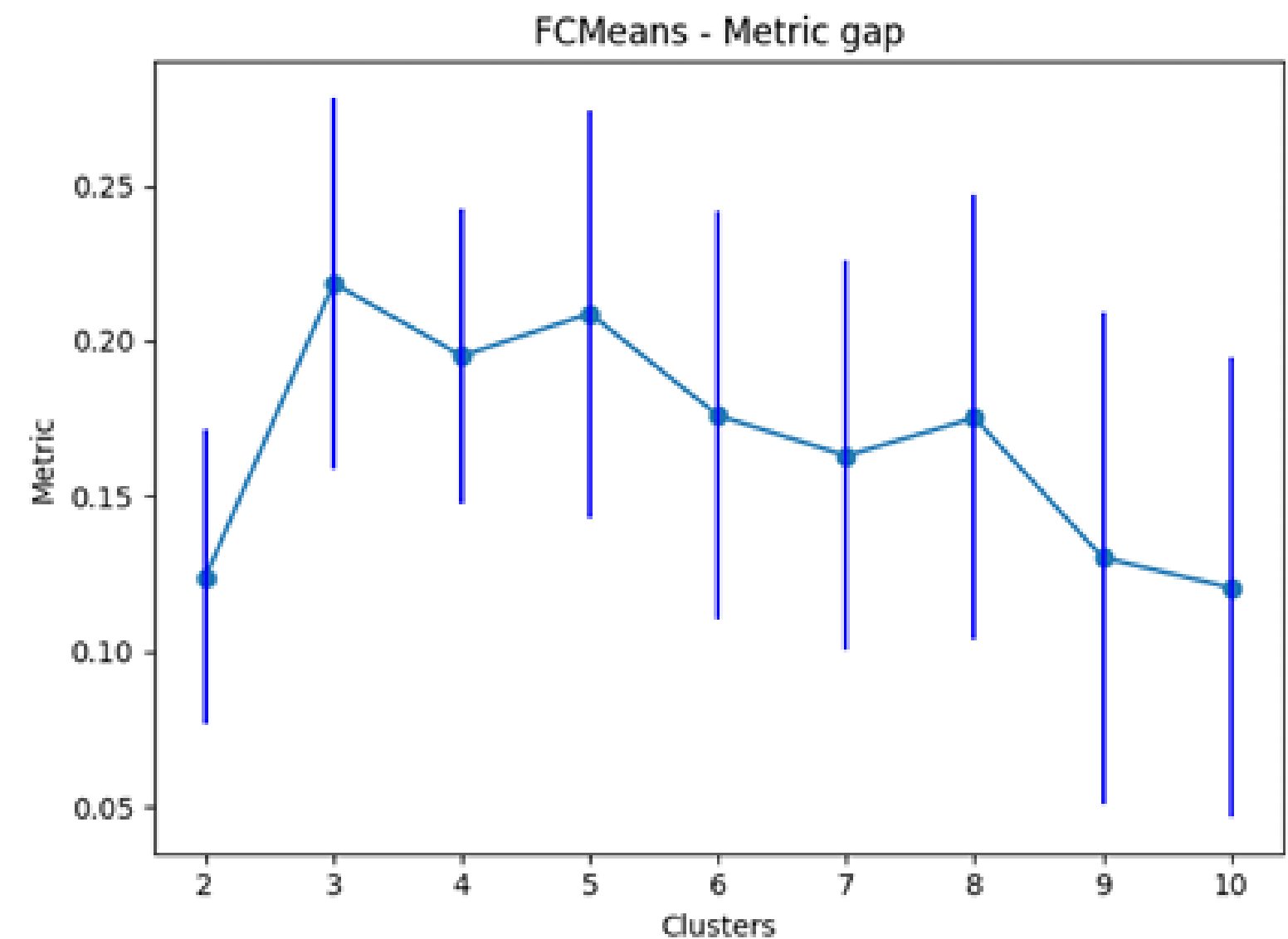
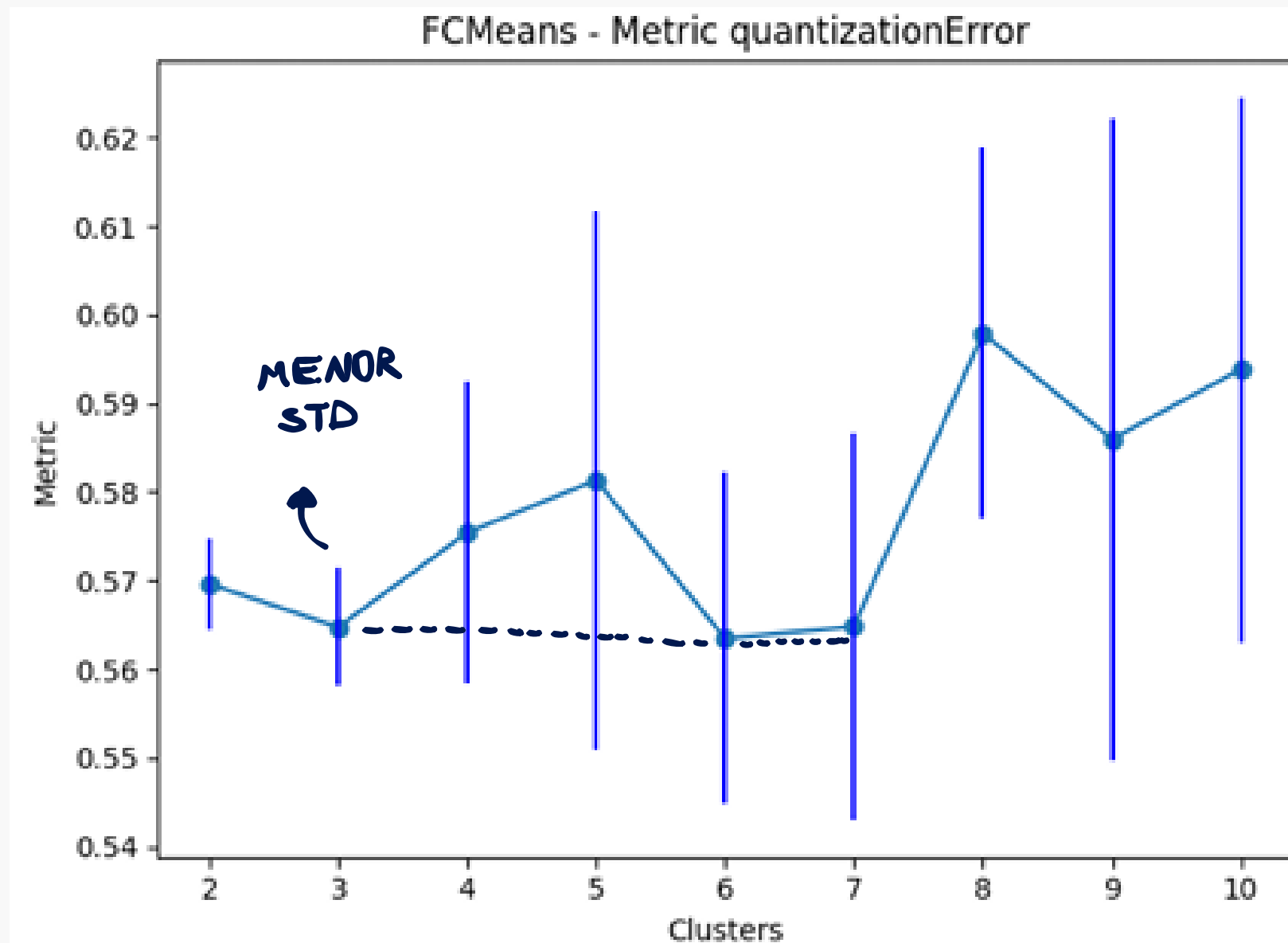
---

- **Comportamento Instável:** As métricas de qualidade dos grupos oscilaram bastante.
  - **Erro Quantizado:** A métrica não melhorou de forma consistente e ficou oscilando para cada valor de K.
  - **Métrica Principal (Gap):** A Estatística Gap atingiu seu valor máximo em  $K=3$ , indicando este como o número ideal.
  - **Perfil Identificado:** Assim como no K-Means, os perfis foram criados com base no "Mês da Estadia".
-

# Gráficos Fuzzy C-Means

---

Mais uma vez, a estatística GAP foi a responsável por definir o número de K ideal.



# *Resultados Comparativos*

---

- **Número de Clusters:** Ambos os algoritmos, pela métrica Gap, sugeriram  $K=3$  como o número ideal de perfis.
  - **Perfis Encontrados:** Os dois métodos criaram grupos muito semelhantes, baseados na época do ano em que os clientes viajam.
  - **Desempenho Geral:** O K-Means teve um desempenho levemente superior.
  - **Justificativa:** O valor da Estatística Gap do K-Means foi maior, o que demonstra grupos mais consistentes e bem definidos.
-

# *Limitações e Críticas*

---

- **Pré-processamento Questionável:** A transformação de "Destino" e "Hotel" em números por ordem alfabética foi pouco eficaz, algo admitido pelos próprios autores. Isso criou distâncias artificiais pelo pré-processamento equivocado (LabelEncoder).
  - **Falta de Clareza na Metodologia:** Detalhes cruciais do pré-processamento são explicados apenas na seção de discussão, e não na metodologia, não houve verificação de outliers por exemplo ou outras etapas essenciais de pré-processamento.
  - Conhecimento óbvio (viagens em sazonalidades diferentes)
-

.....



*Obrigado*



.....