Documentação e Conclusão do Código K-Means

Introdução

Este notebook implementa o algoritmo de agrupamento K-means para identificar grupos de centros logísticos baseados em geolocalização (latitude e longitude). O objetivo é agrupar locais de entrega e calcular o valor total das entregas em cada cluster para otimizar a localização dos centros de distribuição.

1. Importação de Bibliotecas

O código começa com a importação das bibliotecas necessárias:

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
```

- `pandas`: usado para manipulação de dados.
- `numpy`: utilizado para operações matemáticas.
- `KMeans`: algoritmo de clusterização do `sklearn`.
- `StandardScaler`: usado para normalizar os dados.
- `matplotlib`: biblioteca para visualização dos clusters.

2. Carregamento e Preparação dos Dados

O dataset de geolocalização é carregado e processado para garantir que os valores de latitude, longitude e preço estejam no formato correto (float):

```
data = pd.read_csv(file_path)
data['latitude'] = data['latitude'].str.replace(',', '.').astype
data['longitude'] = data['longitude'].str.replace(',', '.').asty
data['price'] = data['price'].str.replace(',', '.').astype(float
```

Aqui, o dataset está sendo lido de um arquivo CSV, e os valores de latitude, longitude e preço, que originalmente estavam no formato de string com vírgulas, são convertidos para o formato de ponto flutuante.

3. Normalização dos Dados

Os valores de latitude e longitude são normalizados para garantir que todas as características estejam na mesma escala antes de aplicar o K-means:

```
scaler = StandardScaler()
scaled_features = scaler.fit_transform(data[['latitude', 'longit
```

O `StandardScaler` ajusta os valores para que tenham uma média de 0 e desvio padrão de 1, o que é necessário para algoritmos baseados em distância, como o K-means.

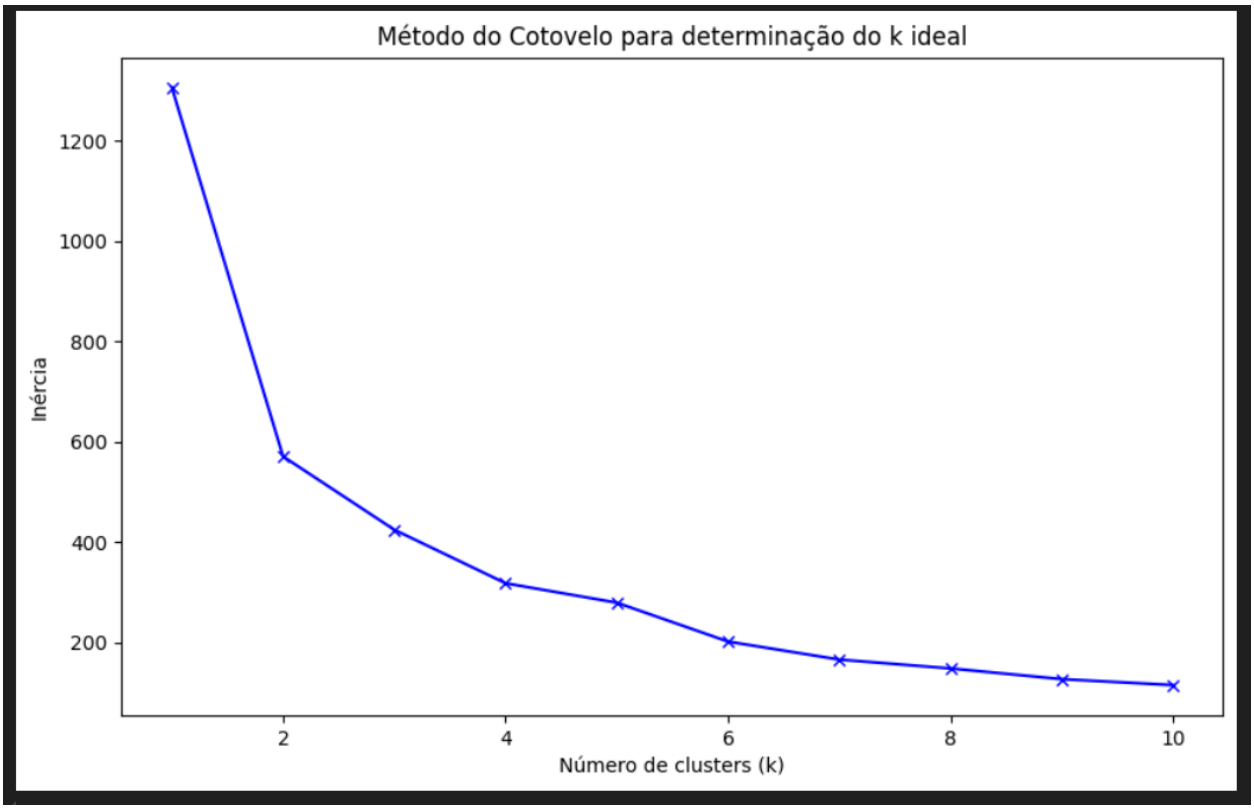
4. Determinação do Número de Clusters (Método do Cotovelo)

O método do cotovelo é usado para determinar o número ideal de clusters. Aqui, o K-means é executado para diferentes valores de `k` (número de clusters), e a inércia (soma das distâncias quadráticas dentro dos clusters) é calculada:

```
inertias = []
k_range = range(1, 11)
for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_features)
    inertias.append(kmeans.inertia_)
```

O gráfico do cotovelo é gerado para visualização:

```
plt.figure(figsize=(10, 6))
plt.plot(k_range, inertias, 'bx-')
plt.xlabel('Número de clusters (k)')
plt.ylabel('Inércia')
plt.title('Método do Cotovelo para determinação do k ideal')
plt.show()
```



5. Aplicação do K-means

Após observar o gráfico do cotovelo, foi decidido utilizar `k=4` clusters para agrupar os centros de entrega:

```
k = 4
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(scaled_features)
data['Cluster'] = kmeans.labels_
```

Cada ponto de entrega é atribuído a um dos 4 clusters, e os rótulos dos clusters são adicionados ao dataframe.

6. Cálculo do Valor Total das Entregas por Cluster

Aqui, o valor total das entregas para cada cluster é calculado e impresso:

```
cluster_totals = data.groupby('Cluster')['price'].sum()
print("Valor total das entregas para cada cluster:")
print(cluster_totals)
```

Além disso, são calculadas estatísticas adicionais, como o valor total, valor médio e o número de entregas em cada cluster, junto com a média de latitude e longitude:

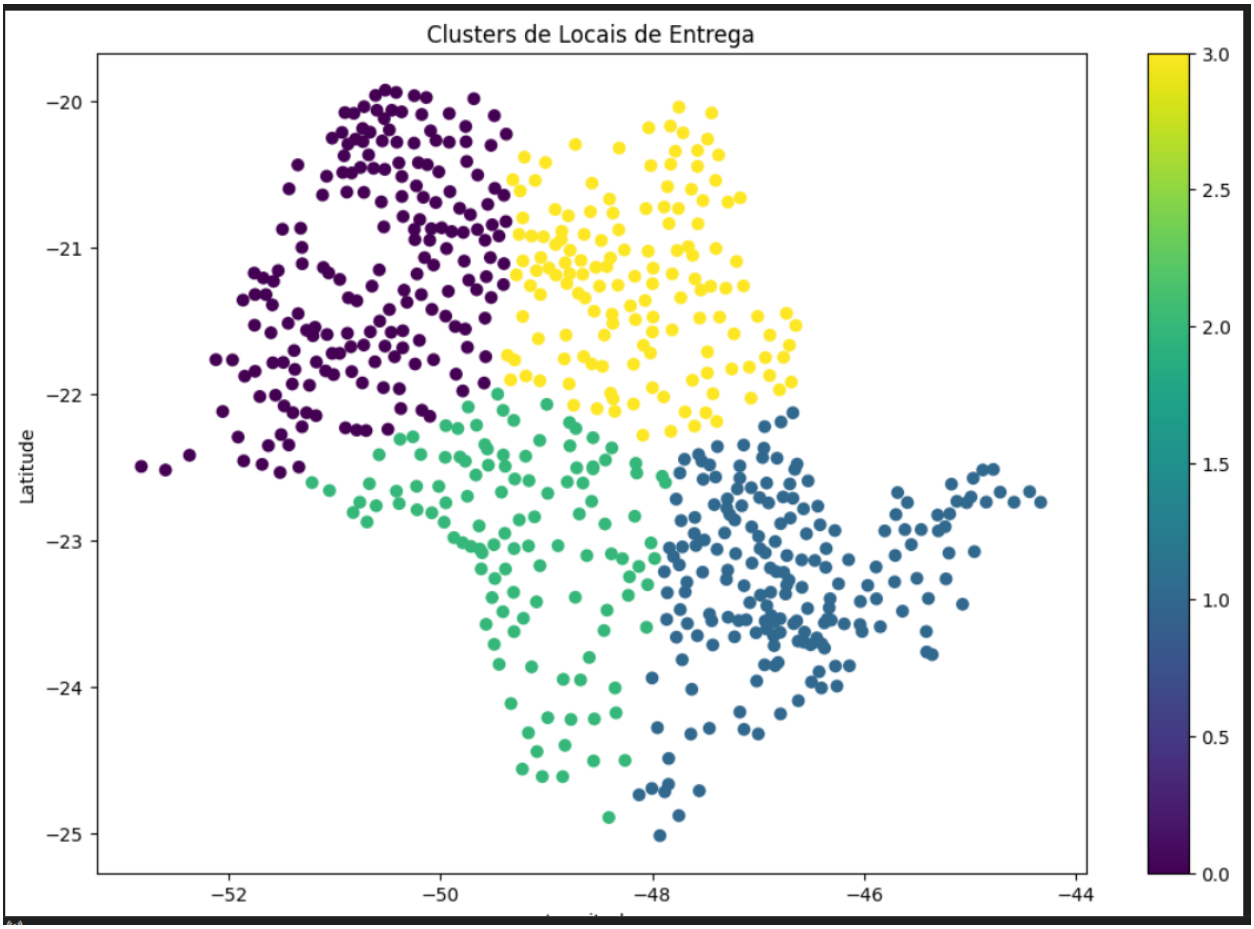
```
cluster_stats = data.groupby('Cluster').agg({
    'price': ['sum', 'mean', 'count'],
    'latitude': 'mean',
    'longitude': 'mean'
}).round(2)
```

7. Visualização dos Clusters

Um gráfico é gerado para visualizar os clusters formados com base nas coordenadas de latitude e longitude:

```
plt.figure(figsize=(12, 8))
scatter = plt.scatter(data['longitude'], data['latitude'], c=dat
plt.colorbar(scatter)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
```

```
plt.title('Clusters de Locais de Entrega')
plt.show()
```



Explicação Teórica

O algoritmo de cluster escolhido foi o **K-means**. Este método é popular por sua simplicidade e eficiência, especialmente ao lidar com dados geoespaciais.

- **Método do Cotovelo** foi usado para determinar o número de clusters. O gráfico do cotovelo é uma técnica visual para identificar o ponto em que o acréscimo de clusters não reduz significativamente a inércia, o que indica o número ideal de clusters.
- **K-means** foi escolhido porque:

- É eficiente para agrupamento de dados geoespaciais.
- Produz resultados interpretáveis.
- É bem escalável para grandes volumes de dados.

Avaliação e Conclusão

- **Positivos**:
  - Identificou grupos distintos de centros de entrega.
  - Forneceu insights valiosos sobre o valor total das entregas em diferentes áreas.
- **Limitações**:
  - Assume que os clusters são circulares.
  - Requer um número predefinido de clusters.
  - Sensível a outliers.

O K-means oferece uma base sólida para a otimização da localização dos centros de distribuição, mas pode ser complementado com outras análises para decisões mais estratégicas.

Conclusão com Base no Gráfico e Estatísticas

Gráfico dos Clusters

O gráfico mostra a distribuição dos pontos de entrega, que foram agrupados em quatro clusters com base nas coordenadas de latitude e longitude. Os diferentes clusters são representados por diferentes cores, indicando as regiões geográficas onde os pontos de entrega estão concentrados. A separação clara entre os clusters mostra que o algoritmo K-means conseguiu identificar áreas distintas de entrega.

- O cluster mais ao sul (verde) cobre uma área com maior dispersão em latitude.
- O cluster roxo e o cluster amarelo mostram uma divisão clara de áreas geográficas adjacentes, com a transição de um cluster para o outro acontecendo em áreas de fronteira próximas.

- O cluster azul cobre a região mais a leste, com uma alta concentração de pontos.

Análise das Estatísticas

Os resultados estatísticos fornecem insights sobre o valor total, a média de valores de entrega, o número de entregas, e as localizações médias de cada cluster:

1. **Cluster 1 (Roxo)**:
    - Valor total: R$ 12.425,34.
    - Maior número de entregas (196), com uma média de R$ 63,39 por entrega.
    - Localizado mais ao sul e leste, com uma latitude média de -23,26 e longitude média de -46,69.
2. **Cluster 0 (Amarelo)**:
    - Valor total: R$ 11.971,48.
    - Segundo maior número de entregas (193), com uma média de R$ 62,03 por entrega.
    - Latitude média de -21,16 e longitude média de -50,62, indicando uma região mais ao norte.
3. **Cluster 3 (Azul)**:
    - Valor total: R$ 9.354,57.
    - 144 entregas, com uma média de R$ 64,96 por entrega (a maior média entre os clusters).
    - Localizado ao leste, com latitude média de -21,24 e longitude média de -48,14.
4. **Cluster 2 (Verde)**:
    - Valor total: R$ 7.432,42.
    - Menor número de entregas (120) e o menor valor total.
    - Localizado ao sul, com uma latitude média de -23,02 e longitude média de -49,22.

Conclusão:

- O **Cluster 1** é o mais ativo em termos de valor total e número de entregas, localizado em uma área com alta densidade de entregas.
- O **Cluster 2** tem o menor número de entregas e o menor valor total, o que pode indicar uma área geograficamente menos ativa em termos de logística.
- O **Cluster 3**, embora tenha um número menor de entregas, apresenta a maior média de valor por entrega, indicando que, nesta área, as entregas podem ter valores mais altos.