Untitled17.ipynb

File Edit View Insert Runtime Tools Help

Share

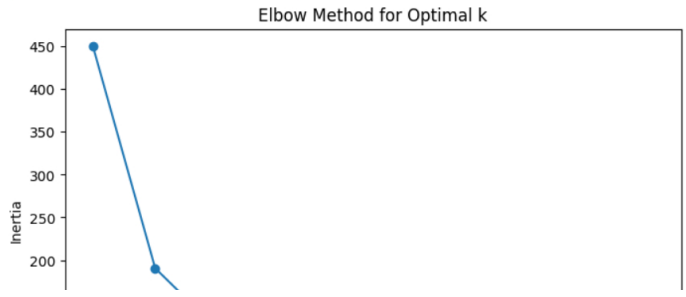
+ Code + Text

RAM Disk

Gemini


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
dataset_url = "https://raw.githubusercontent.com/mwaskom/seaborn-data/master/iris.csv"
df = pd.read_csv(dataset_url)
print(df.head())
print("\nDataset info: ")
print(df.info())
print("\nMissing values in Dataset:")
print(df.isnull().sum())
features = df[['sepal_length','sepal_width','petal_length']]
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)
print("\n First 5 rows of scaed features:")
print(scaled_features[:5])
inertia=[]
k_range = range(1,11)
for k in k_range:
    kmeans = KMeans(n_clusters=k,random_state=42)
    kmeans.fit(scaled_features)
    inertia.append(kmeans.inertia_)
plt.figure(figsize=(8,5))
plt.plot(k_range,inertia,marker='o')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal k')
plt.show()

optimal_k = 3
kmeans = KMeans(n_clusters=optimal_k,random_state=42)
cluster_labels = kmeans.fit_predict(scaled_features)
df['Cluster']=cluster_labels
print("\nFirst 5 rows of clustered data is:")
print(df.head())

plt.figure(figsize=(8,6))
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s=300,c='red',label='Centroids')
sns.scatterplot(x=scaled_features[:,0],y = scaled_features[:,1],hue = cluster_labels,palette='viridis')
plt.xlabel('Featured 1')
plt.ylabel('Featured 2')
plt.title('Costumer segment')
plt.legend()
plt.show()
```

```
   sepal_length  sepal_width  petal_length  petal_width species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa

Dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
None

Missing values in Dataset:
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64

 First 5 rows of scaed features:
[[-0.90068117  1.01900435 -1.34022653]
 [-1.14301691 -0.13197948 -1.34022653]
 [-1.38535265  0.32841405 -1.39706395]
 [-1.50652052  0.09821729 -1.2833891 ]
 [-1.02184904  1.24920112 -1.34022653]]
```



✓ 0s completed at 4:21 PM