K-means Clustering for Customer Segmentation

```
import pandas as pd

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler


# Load the dataset

file_path = '/mnt/data/Mall_Customers.csv'

data = pd.read_csv(file_path)


# Preprocess the data

# Assuming the relevant features for clustering are 'Annual Income (k$)' and 'Spending Score (1-100)'

X = data[['Annual Income (k$)', 'Spending Score (1-100)']]


# Standardize the data

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)


# Determine the optimal number of clusters using the Elbow method

inertia = []

for n in range(1, 11):

    kmeans = KMeans(n_clusters=n, random_state=42)

    kmeans.fit(X_scaled)

    inertia.append(kmeans.inertia_)
```

```
# Plot the Elbow curve
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), inertia, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal Number of Clusters')
plt.grid(True)
plt.savefig('/mnt/data/elbow_method.png')
plt.show()

# From the Elbow curve, let's assume the optimal number of clusters is 5
optimal_clusters = 5
kmeans = KMeans(n_clusters=optimal_clusters, random_state=42)
clusters = kmeans.fit_predict(X_scaled)

# Add the cluster labels to the original dataframe
data['Cluster'] = clusters

# Visualize the clusters
plt.figure(figsize=(10, 6))
for cluster in range(optimal_clusters):
    plt.scatter(X_scaled[clusters == cluster, 0], X_scaled[clusters == cluster, 1], label=f'Cluster {cluster}')
plt.xlabel('Annual Income (k$) - Scaled')
plt.ylabel('Spending Score (1-100) - Scaled')
plt.title('K-means Clustering of Customers')
plt.legend()
```

```
plt.grid(True)

plt.savefig('/mnt/data/kmeans_clusters.png')

plt.show()



# Save the clustered data to a new CSV file

clustered_data_path = '/mnt/data/clustered_customers.csv'

data.to_csv(clustered_data_path, index=False)
```

Model Interpretation

K-means Clustering Model Evaluation:

The K-means clustering algorithm was used to group customers based on their annual income and spending score.

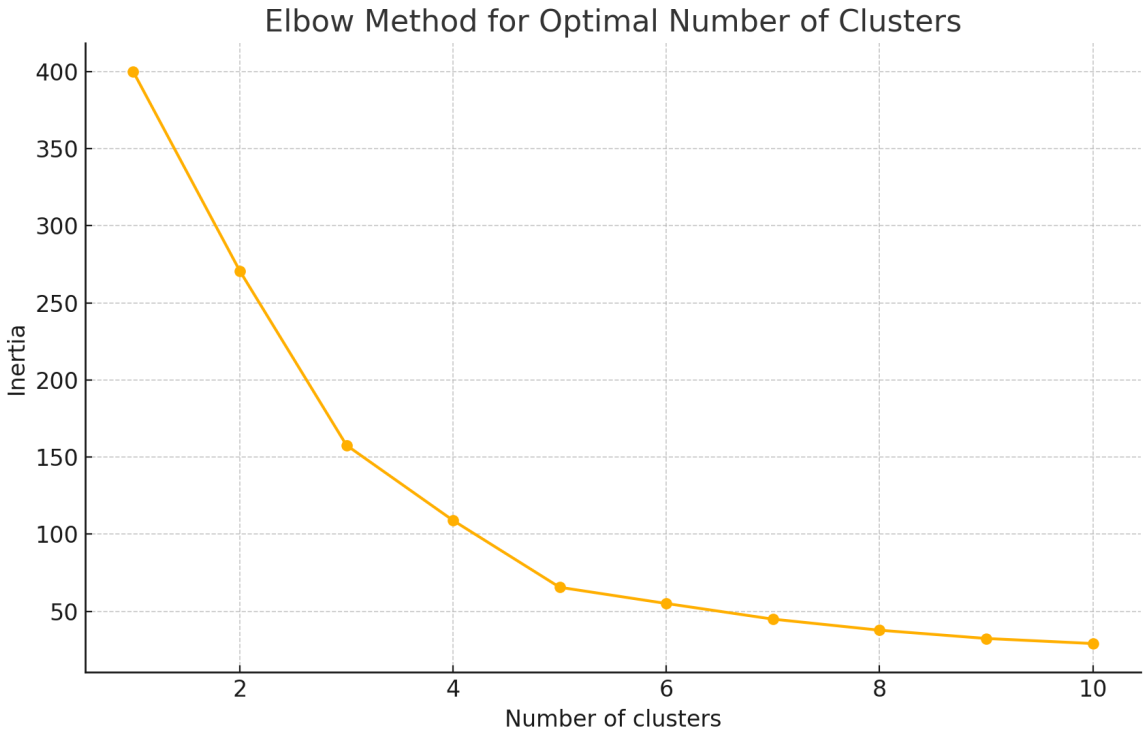
1. **Elbow Method**:

- The elbow method was employed to determine the optimal number of clusters. The plot shows that the optimal number of clusters is around 5, as the inertia starts to decrease at a slower rate after this point.

2. **Clustering Visualization**:

- The scatter plot visualizes the clusters of customers based on the standardized annual income and spending score. Each cluster is represented by a different color, showing how the customers are segmented into distinct groups.

These clusters can be used to identify different customer segments, such as high-income high-spending customers, low-income low-spending customers, and so on, enabling more targeted marketing strategies.


Elbow Method for Optimal Number of Clusters
Inertia
Number of clusters
400
350
300
250
200
150
100
50
2
4
6
8
10


K-means Clustering of Customers
Spending Score (1-100) - Scaled
Annual Income (k$) - Scaled
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3