Assignment: Application and Challenges of K-Means Clustering

Objective:

The objective of this assignment is to explore the real-world applications and challenges of the K-Means Clustering algorithm. Students will examine how K-Means can be applied to solve practical problems and evaluate its limitations in comparison to other clustering algorithms.

Instructions:

Complete the following tasks that focus on the applications and challenges of K-Means Clustering. The assignment consists of 2 parts.

Part 1: Real-World Applications of K-Means

Task 1: Select a Real-World Scenario

- Choose one real-world application where K-Means clustering can be used (e.g.,customer segmentation, image compression, anomaly detection, or market segmentation).
- Provide an explanation of how K-Means clustering works in this scenario and why it is useful.

Task 2: Benefits of Using K-Means

- Discuss two main benefits of using K-Means in your chosen scenario. For example, how it improves decision-making, reduces complexity, or enhances predictions.

Expected Output:

- Around 150-200 words for Task 1 and Task 2 combined.

Part 1: Real-World Applications of K-Means

Task 1: Customer segmentation

I'll choose **customer segmentation** as the real-world scenario. K-Means clustering is widely used in customer segmentation to group customers based on similar purchasing behaviors, demographics, or browsing patterns. In this scenario, businesses collect data such as age, income, spending habits, and product preferences. K-Means groups these customers into K clusters, each representing a distinct customer profile. For example, one cluster may contain high-income frequent buyers, while another may represent budget-conscious occasional shoppers. This segmentation helps businesses target specific groups with personalized marketing strategies, product recommendations, or loyalty programs.

Task 2: Benefits of Using K-Means

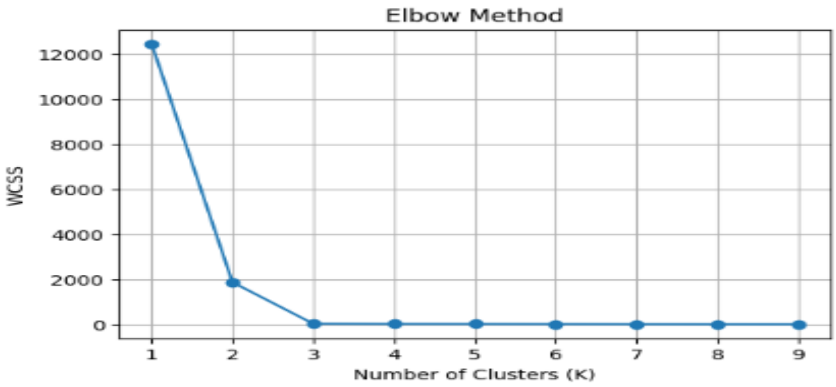
One major benefit of using K-Means in customer segmentation is improved decision-making. By understanding different customer types, businesses can tailor marketing efforts more effectively, increasing engagement and conversion rates. A second benefit is reduced complexity. Instead of analyzing thousands of individual customer profiles, companies can focus on a few well-defined clusters, making strategic planning more efficient and data-driven.

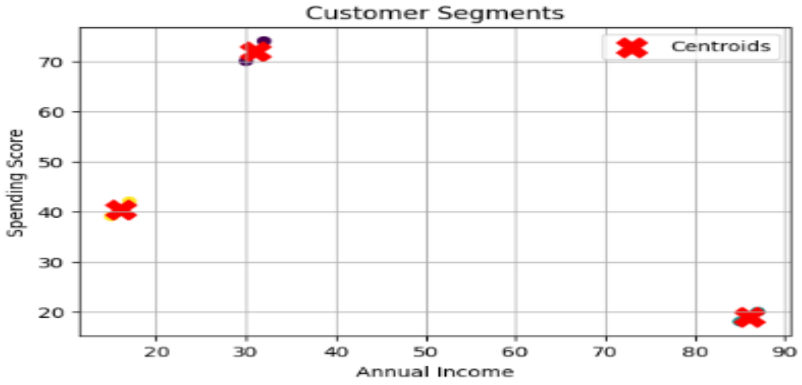
Code:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Sample customer data
data = {
   'Annual Income': [15, 16, 17, 30, 31, 32, 85, 86, 87],
   'Spending Score': [39, 40, 42, 70, 72, 74, 18, 19, 20]
}
df = pd.DataFrame(data)

# Elbow Method to find optimal K
wcss = []
for i in range(1, 10):
   kmeans = KMeans(n_clusters=i, random_state=0)
   kmeans.fit(df)
   wcss.append(kmeans.inertia_)

# Plot the elbow graph
plt.figure(figsize=(6, 4))
plt.plot(range(1, 10), wcss, marker='o')
plt.title('Elbow Method')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('WCSS')
plt.grid(True)
plt.show()

# Apply KMeans with chosen K (e.g., 3 from elbow)
kmeans = KMeans(n_clusters=3, random_state=0)
df['Cluster'] = kmeans.fit_predict(df)

# Plot the clusters
plt.figure(figsize=(6, 4))
plt.scatter(df['Annual Income'], df['Spending Score'], c=df['Cluster'], cmap='viridis')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
        s=200, c='red', marker='X', label='Centroids')
plt.title('Customer Segments')
```

```
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.legend()
plt.grid(True)
plt.show()
```

Graph:





Part 2: Challenges and Alternatives

Task 1: Limitations of K-Means Clustering

- List and explain two limitations of K-Means clustering (e.g., sensitivity to initial centroids, difficulty handling non-spherical clusters, or issues with clusters of varying sizes).

Task 2: When Not to Use K-Means

- Describe a situation where K-Means clustering is not the best choice and explain why. Suggest a more suitable algorithm for that scenario.

Expected Output:

- Around 150-200 words total for all tasks.

Part 2:

Task 1: Limitations of K-Means Clustering

One limitation of K-Means is its **sensitivity to initial centroids**. Since K-Means starts with randomly chosen cluster centers, different initializations can lead to different results. This can affect the stability and accuracy of the final clusters. Another limitation is that K-Means **struggles with non-spherical**

clusters or clusters of varying densities and sizes. The algorithm assumes clusters are roughly circular and of similar size, so it may group distinct clusters incorrectly if the actual data structure is more complex.

Task 2: When Not to Use K-Means

K-Means is not ideal when dealing with data that has a natural hierarchy or nested structure, such as in gene expression analysis or customer subgroups. In these cases, forcing the data into flat clusters may ignore meaningful relationships. A better alternative is Hierarchical Clustering, which builds a tree-like structure (dendrogram) showing how data points are grouped step by step. This approach doesn't require choosing the number of clusters in advance and can reveal deeper patterns in the data.
