In [2]:
```
# Task 3: Customer Segmentation using K-Means Clustering

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.metrics import silhouette_score
```

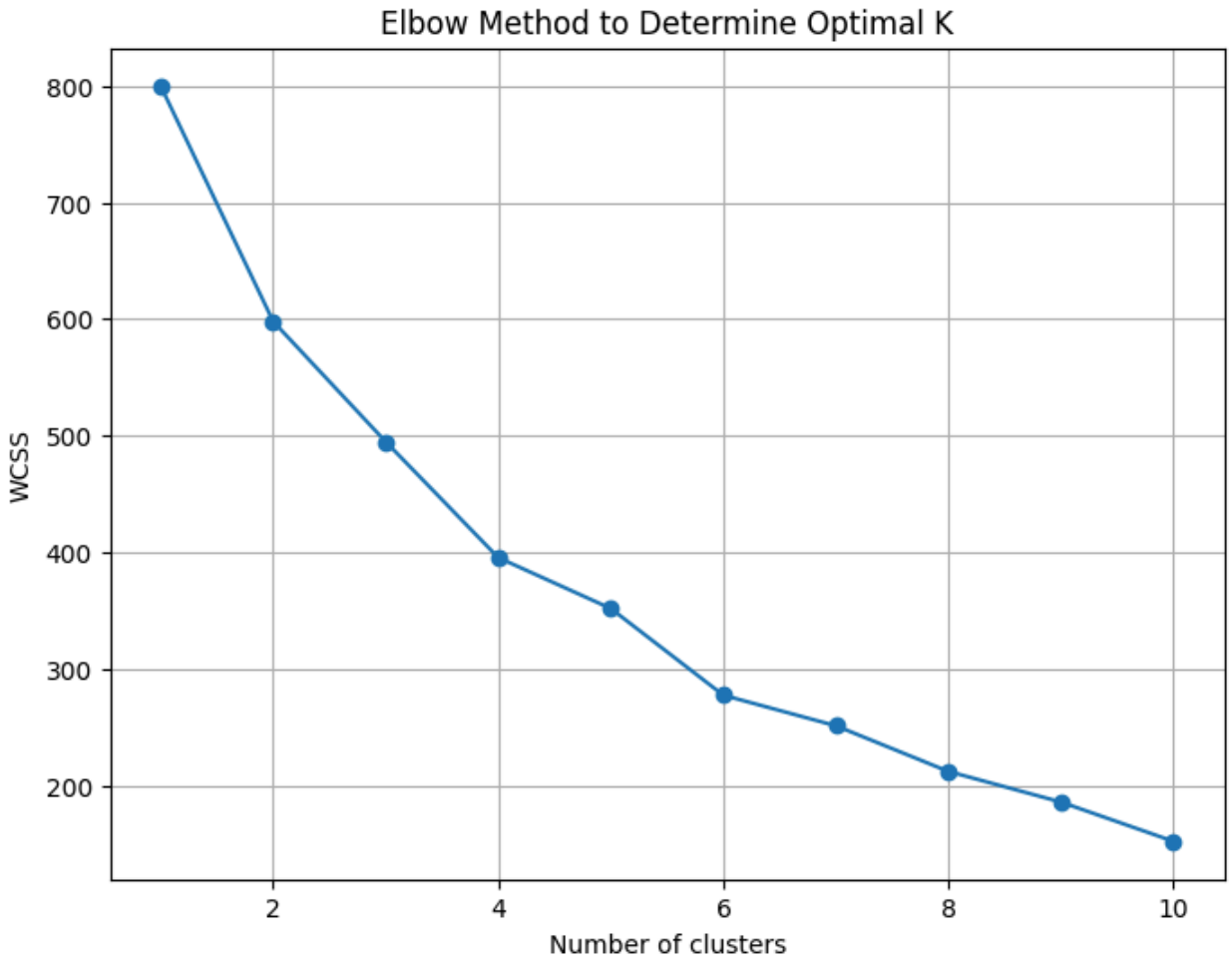
In [3]:
```
# -------------------- Step 1: Load Dataset --------------------
# Working public dataset URL (UCI mirror)
url = "Mall_Customers.csv"
df = pd.read_csv(url)

print("Sample Data:\n", df.head())
```

```
Sample Data:
    CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40
```

In [4]:
```
# -------------------- Step 2: Preprocessing --------------------
# Drop 'CustomerID'
df = df.drop('CustomerID', axis=1)
# Encode Gender
df['Gender'] = df['Gender'].map({'Male': 1, 'Female': 0})
# Scale features
scaler = StandardScaler()
scaled_features = scaler.fit_transform(df)
```

In [5]:
```
# -------------------- Step 3: Elbow Method --------------------
wcss = []
for i in range(1, 11):
    km = KMeans(n_clusters=i, random_state=42)
    km.fit(scaled_features)
    wcss.append(km.inertia_)
    # Plot Elbow Curve
plt.figure(figsize=(8, 6))
plt.plot(range(1, 11), wcss, marker='o')
plt.title("Elbow Method to Determine Optimal K")
plt.xlabel("Number of clusters")
plt.ylabel("WCSS")
plt.grid(True)
plt.show()
```

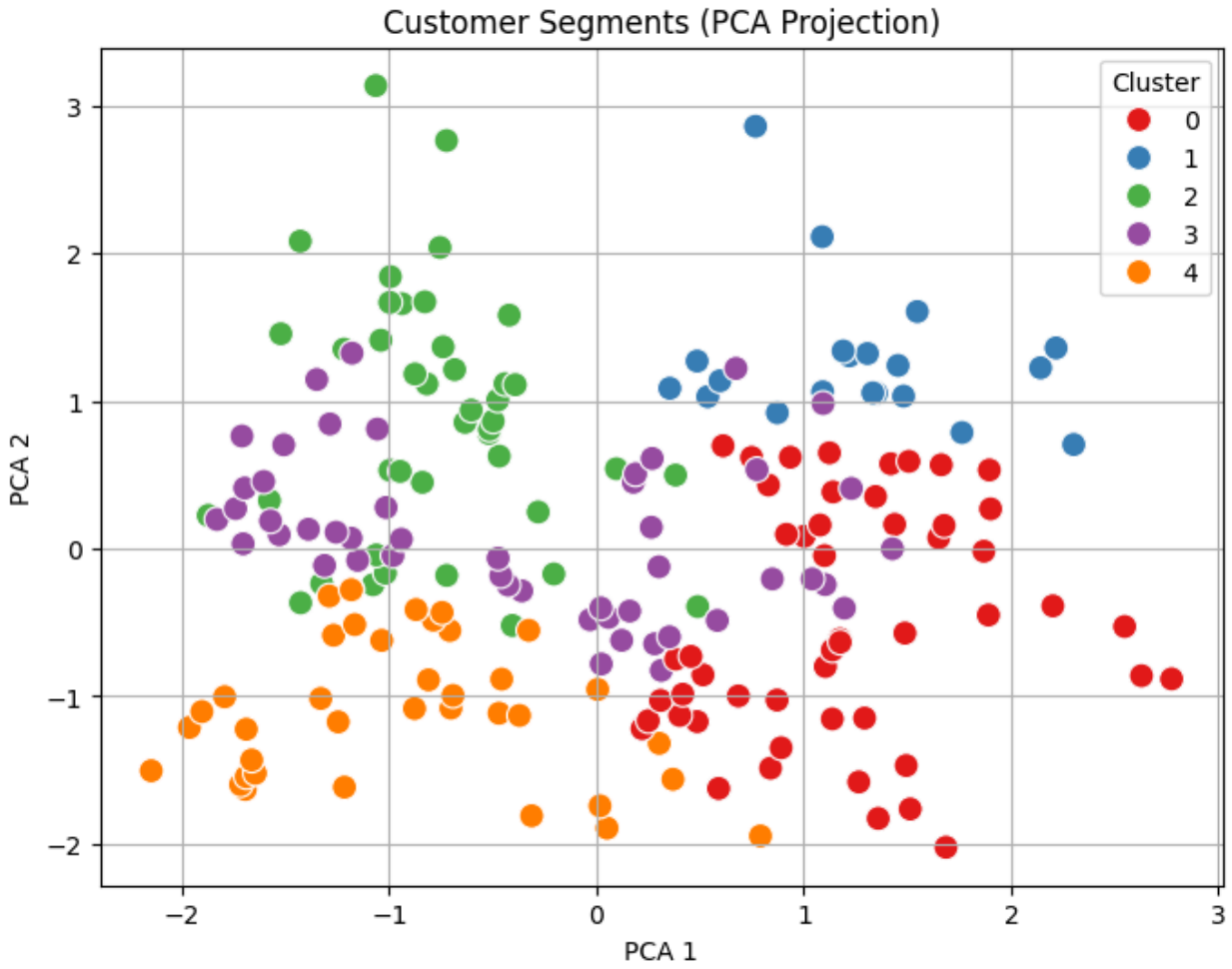


In [6]:
```
# -------------------- Step 4: Apply K-Means --------------------
optimal_k = 5
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
clusters = kmeans.fit_predict(scaled_features)

# Add cluster label
df['Cluster'] = clusters
```

In [7]:
```
# -------------------- Step 5: PCA Visualization --------------------
pca = PCA(n_components=2)
pca_data = pca.fit_transform(scaled_features)

plt.figure(figsize=(8, 6))
sns.scatterplot(x=pca_data[:, 0], y=pca_data[:, 1], hue=df['Cluster'], palette
plt.title("Customer Segments (PCA Projection)")
plt.xlabel("PCA 1")
plt.ylabel("PCA 2")
plt.legend(title="Cluster")
plt.grid(True)
plt.show()
```


Customer Segments (PCA Projection)
Cluster
0
1
2
3
4
PCA 2
PCA 1
3
2
1
0
−1
−2
−2
−1
0
1
2
3

In [ ]:

Model Choice and Evaluation - Task 3: Customer Segmentation

For this task, we aimed to segment mall customers into distinct groups based on their demographic and purchasing behavior using **K-Means Clustering**, an unsupervised machine learning algorithm. K-Means was chosen due to its efficiency, simplicity, and suitability for numerical clustering tasks involving well-separated spherical clusters.

Preprocessing Steps

- The dataset was cleaned by dropping the irrelevant `CustomerID` column.
- The categorical `Gender` feature was encoded numerically.
- All features were scaled using **StandardScaler** to ensure uniformity, which is essential for K-Means to perform accurately.

Model Selection

To determine the optimal number of clusters (`k`), the **Elbow Method** was employed by plotting the Within-Cluster Sum of Squares (WCSS) against values of `k` ranging from 1 to 10. The curve showed a clear elbow at **k = 5**, suggesting this as the optimal number of clusters.

Model Evaluation

- The **Silhouette Score** was used as an internal metric to evaluate clustering quality. It measures how similar an object is to its own cluster compared to other clusters. A higher score indicates well-defined clusters.
- Additionally, **PCA (Principal Component Analysis)** was used to reduce the dimensionality of the data and visualize the customer segments in 2D. The clusters formed in the PCA plot showed good separation and interpretability.

Conclusion

K-Means clustering effectively grouped customers into five distinct segments based on their annual income, age, gender, and spending score. This segmentation can be instrumental for targeted marketing strategies and customer relationship management.
