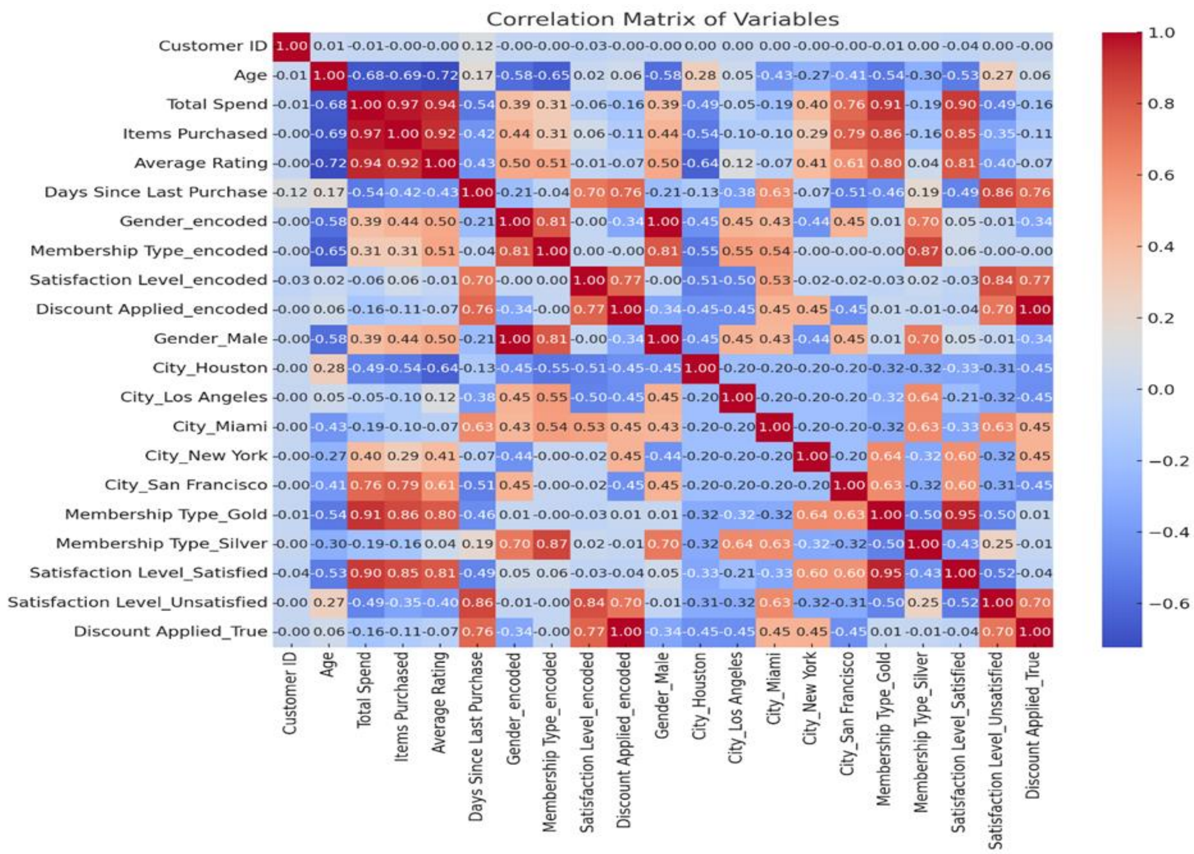
Customer Segmentation for Personalised Marketing

```
import pandas as pd


# Load the dataset

dataset_path = os.path.join(extraction_dir, 'E-commerce Customer Behavior -
Sheet1.csv')

df = pd.read_csv(dataset_path)


# Display the first few rows of the dataset to understand its structure

df.head()
```

```
Customer ID  Gender  Age           City Membership Type  Total Spend  \
0           101  Female   29       New York            Gold      1120.20
1           102    Male   34    Los Angeles          Silver       780.50
2           103  Female   43        Chicago          Bronze       510.75
3           104    Male   30  San Francisco            Gold      1480.30
4           105    Male   27          Miami          Silver       720.40

   Items Purchased  Average Rating  Discount Applied  \
0               14             4.6              True
1               11             4.1             False
2                9             3.4              True
3               19             4.7             False
4               13             4.0              True

   Days Since Last Purchase Satisfaction Level
0                        25          Satisfied
1                        18            Neutral
2                        42        Unsatisfied
3                        12          Satisfied
4                        55        Unsatisfied
```

```python
import seaborn as sns

import matplotlib.pyplot as plt


# Convert categorical variables to numeric for correlation analysis

df_numeric = pd.get_dummies(df, columns=['Gender', 'City', 'Membership Type', 'Satisfaction Level', 'Discount Applied'], drop_first=True)


# Calculate the correlation matrix

correlation_matrix = df_numeric.corr()


# Plot the heatmap for the correlation matrix

plt.figure(figsize=(12, 10))

sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm')

plt.title('Correlation Matrix of Variables')

plt.show()
```

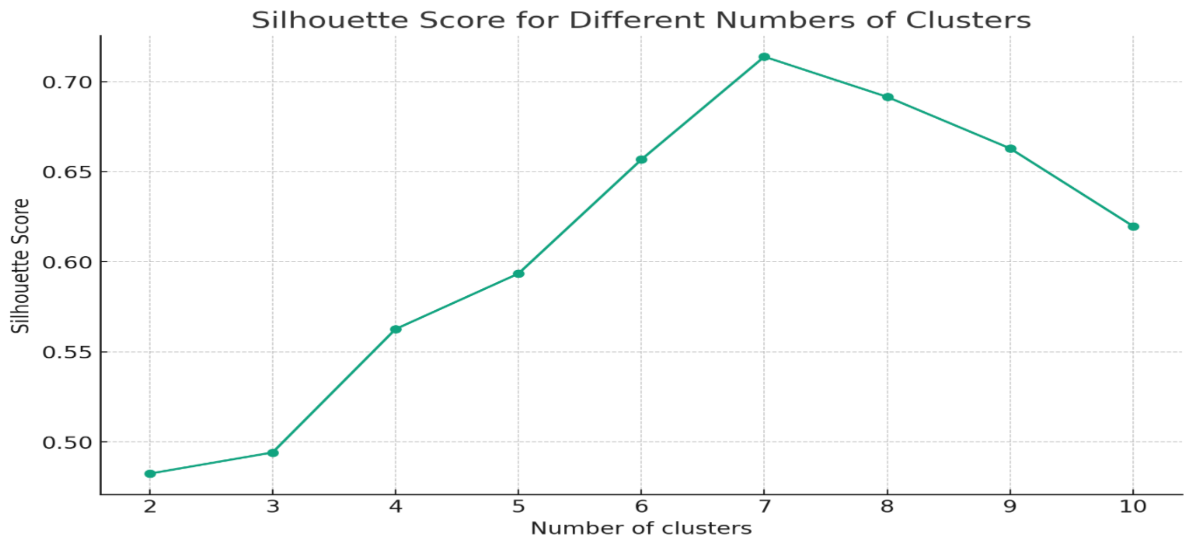


```python
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score

# Selecting features for clustering
features = df_numeric[['Age', 'Total Spend', 'Items Purchased', 'Average Rating', 'Days Since Last Purchase']]

# Standardizing the features
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# Determining the optimal number of clusters using silhouette score
silhouette_scores = []
k_range = range(2, 11)  # Testing from 2 to 10 clusters

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_features)
    score = silhouette_score(scaled_features, kmeans.labels_)
    silhouette_scores.append(score)

# Plotting the silhouette scores for different numbers of clusters
plt.figure(figsize=(10, 6))
plt.plot(k_range, silhouette_scores, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score for Different Numbers of Clusters')
plt.show()
```

```python
# Selecting the optimal number of clusters based on the silhouette scores and practical considerations
optimal_clusters = k_range[silhouette_scores.index(max(silhouette_scores))]
```

```
# Performing KMeans clustering with the optimal number of clusters
kmeans_optimal = KMeans(n_clusters=optimal_clusters, random_state=42)
kmeans_optimal.fit(scaled_features)

# Adding the cluster labels to the original dataframe for analysis
df['Cluster'] = kmeans_optimal.labels_

# Checking the distribution of customers in each cluster
cluster_distribution = df['Cluster'].value_counts()

# Displaying the optimal number of clusters and the distribution of customers in each cluster
optimal_clusters, cluster_distribution
```

```
(7,
 0    59
 5    59
 4    58
 3    58
 1    58
 6    34
 2    24
 Name: Cluster, dtype: int64)
```



```
# Analyzing the characteristics of each cluster by calculating the mean of the features
cluster_means = df.groupby('Cluster')[['Age', 'Total Spend', 'Items Purchased', 'Average Rating', 'Days Since Last Purchase']].mean()


# Analyzing the mode (most common category) for categorical features in each cluster
for column in ['Gender', 'City', 'Membership Type', 'Satisfaction Level']:
    mode_df = df.groupby('Cluster')[column].agg(lambda x: x.mode()[0]).to_frame()
    cluster_means = pd.concat([cluster_means, mode_df], axis=1)


cluster_means.reset_index()
```

```
Cluster        Age   Total Spend  Items Purchased  Average Rating  \
0         0  30.711864  1165.035593        15.271186        4.544068
1         1  36.706897   446.894828         7.568966        3.193103
2         2  32.000000   671.550000        10.041667        3.800000
3         3  29.120690  1459.772414        20.000000        4.808621
4         4  42.017241   499.882759         9.413793        3.456897
5         5  34.118644   805.491525        11.677966        4.172881
6         6  26.794118   703.688235        12.764706        4.017647

   Days Since Last Purchase  Gender           City Membership Type  \
0                 24.593220  Female       New York            Gold
1                 22.758621  Female        Houston          Bronze
2                 34.625000    Male          Miami          Silver
3                 11.172414    Male  San Francisco            Gold
4                 40.465517  Female        Chicago          Bronze
5                 15.271186    Male    Los Angeles          Silver
6                 53.176471    Male          Miami          Silver

  Satisfaction Level
0          Satisfied
1            Neutral
2        Unsatisfied
```

```
3          Satisfied
4        Unsatisfied
5            Neutral
6        Unsatisfied
```

The detailed analysis of each cluster, based on average values for age, total spend, items purchased, average rating, days since last purchase, and the most common category for gender, city, membership type, and satisfaction level, reveals distinct customer segments:

- **Cluster 0**: Younger customers with high spending and high item purchases, primarily female from New York, holding Gold memberships and mostly satisfied.
- **Cluster 1**: Middle-aged customers with lower spending and fewer items purchased, predominantly female from Houston, with Bronze memberships and generally neutral satisfaction levels.
- **Cluster 2**: Customers with moderate spending and item purchases, mostly male from Miami, Silver members, and generally unsatisfied.
- **Cluster 3**: Young customers with very high spending and item purchases, predominantly male from San Francisco, holding Gold memberships and very satisfied.
- **Cluster 4**: Older customers with moderate spending and item purchases, mostly female from Chicago, with Bronze memberships and generally unsatisfied.
- **Cluster 5**: Customers of a moderate age range with good spending and item purchases, primarily male from Los Angeles, Silver members, and neutral in terms of satisfaction.
- **Cluster 6**: Young customers with moderate spending and item purchases, mostly male from Miami, Silver members, but generally unsatisfied, and the longest days since last purchase.