**Summary:**

This notebook explores customer data to identify different customer segments using clustering and then builds a classifier to predict these segments.

**1. Data Exploration**

I started by loadin the Customer\_Data dataset, checked its shape (how many rows and columns), looked at the first few rows to understand the data, and looked at the data types and a summary of the statistics.

* **Key Findings**:
  + The dataset has 8950 rows and 18 columns.
  + Columns like CREDIT\_LIMIT and MINIMUM\_PAYMENTS have some missin' values.
  + The data includes a mix of numerical and categorical (the CUST\_ID) features.

**2. Data Cleaning**

Next up was cleaning the data. I handled the missing values in MINIMUM\_PAYMENTS and CREDIT\_LIMIT by flling them with the mean of their respective columns then i also checked for duplicate rows, but there were none.

* **Key Findings**:
  + Missing values in MINIMUM\_PAYMENTS (313) and CREDIT\_LIMIT (1) were imputed using the mean.
  + No duplicate rows were found.

**3. Exploratory Data Analysis (EDA)**

Then i did some EDA to understand the data better, looked at the distribution of features using histograms, the relationships between features using a correlation heatmap, and checked for outliers using boxplots.

* **Key Findings**:
  + Most features have skewed distributions, indicating the presence of outliers.
  + The correlation heatmap showed some relationships between features
  + Boxplots showed the presence of outliers in several numerical columns.

**4. Outlier Handling**

Based on the boxplots, I decided to handle the outliers, i used the (IQR) method to identify outliers and then capped them at the lower and upper bounds which I calculated, then verified this with new boxplots and statistics.

* **Key Findings**:
  + Outliers in numerical features (except TENURE) were capped using the IQR method.
  + The boxplots after capping showed a reduction in extreme values.

**5. Feature Engineering and Preprocessing**

This stage was about getting the data ready for clustering.

* **Feature Scaling**: I scaled the numerical features using StandardScaler to make sure they had a similar range of values.
* **Dimensionality Reduction**: I used PCA to reduce the number of features while keeping most of the information.
* **New Features**: I created some new features that might be helpful, like ratios of different purchase types and credit limit utilization.
* **Data for Clustering**: Finally I prepared the PCA-transformed data as the input for the clustering algorithm.
* **Key Findings**:
  + Numerical features were scaled.
  + Data dimensionality was reduced to 10 principal components using PCA, retaining 95% variance.
  + Created Seven new ratio and average-based features.
  + The PCA-transformed data is ready for clustering.

**6. Clustering**

I applied K-Means clustering to the prepared data.

* **Optimal Clusters**: Using the elbow method, I determined that 5 was a good number of clusters.
* **Applying K-Means**: trained the K-Means model with 5 clusters and added the cluster labels back to the original DataFrame.
* **Evaluating Clustering**: checked the clustering quality using the silhouette score (about 0.20) and inertia.
* **Visualizing Clusters**: then I visualized the clusters using a scatter plot of the first two PCA components.
* **Key Findings**:
  + 5 distinct customer segments were identified using K-Means.
  + Each cluster has unique characteristics related to spending habits, balance, and credit usage.
  + The clusters are kinda separated in the PCA visualization.

**7. Classification**

Finally, I used the cluster labels as labels to train a classifier.

* **Data Preparation**: I split the PCA-transformed features and the cluster labels into training and testing sets.
* **Training a Classifier**: chose and trained a RandomForestClassifier on the training data.
* **Evaluating Classifier**: then I evaluated the classifier's performance on the testing data using a confusion matrix, classification report, and accuracy score.
* **Key Findings**:
  + The RandomForestClassifier achieved a high accuracy of around 97% in predicting the cluster labels based on the PCA-transformed features.
  + The confusion matrix and classification report showed strong performance across most clusters.

**Overall**

The notebook successfully loaded, cleaned, and preprocessed the customer data. I identified 5 distinct customer segments using K-Means clustering and built a highly accurate classifier to predict these segments based on the transformed features.