Bank Data Clustering Project Report

Table of Contents

Problem - Clustering on Bank Data

List of Figures

List of Tables

---

1. Exploratory Data Analysis (Univariate, Bivariate, and Multivariate Analysis)

1.1. EDA - Summary of the Data

The bank dataset can be summarized under the following basic pointers:

1. There are 210 data points and 7 variables, with no target variable. The seven variables are:
    - spending: Amount spent by the customer per month (in 1000s).
    - advance_payments: Amount paid by the customer in advance by cash (in 100s).
    - probability_of_full_payment: Probability of payment done in full by the customer to the bank.
    - current_balance: Balance amount left in the account to make purchases (in 1000s).
    - credit_limit: Limit of the amount in the credit card (in 10000s).
    - min_payment_amt: Minimum paid by the customer while making payments for purchases made monthly (in 100s).

- max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s).

2. All variables are float data types, and there are no null values in the dataset. The info summary of the data is as follows:

```
memory usage: 11.6 KB
```

Table 1: Bank Dataset Info

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
None
```

3. The description of the data is as follows:

	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.85	2.91	10.59	12.27	14.36	17.30	21.18
advance_payments	210.0	14.56	1.31	12.41	13.45	14.32	15.72	17.25

probability_of_full_payment	210.0	0.87	0.02	0.81	0.86	0.87	0.89	0.92
current_balance	210.0	5.63	0.44	4.90	5.26	5.52	5.96	6.68
credit_limit	210.0	3.26	0.38	2.63	2.94	3.24	3.56	4.03
min_payment_amt	210.0	3.70	1.50	0.77	2.56	3.60	4.77	8.46
max_spent_in_single_shoppin g	210.0	5.41	0.49	4.52	5.04	5.22	5.88	6.55

Table 2: Bank Dataset Description

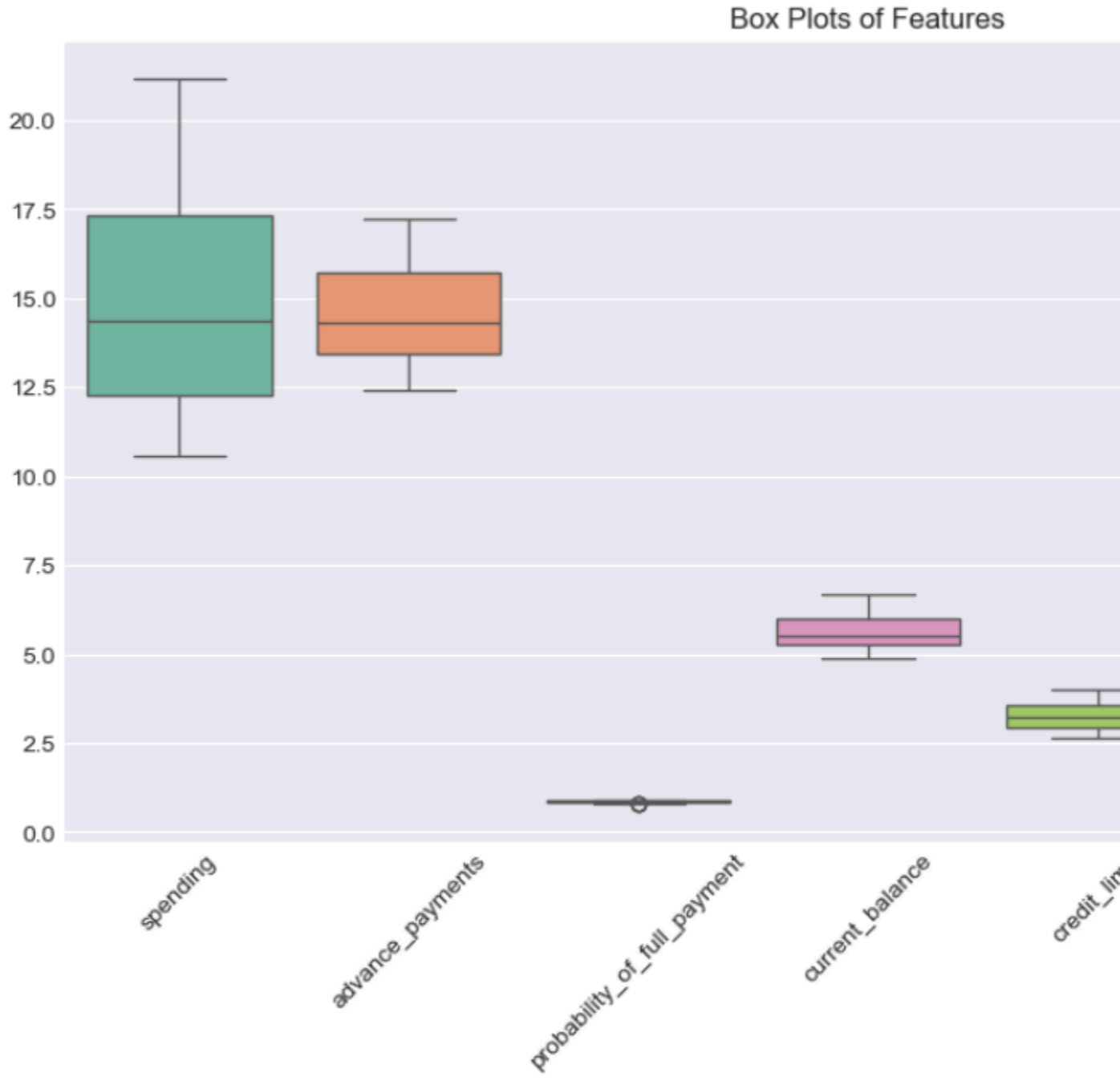
```
print(df.describe())
```

```
          spending  advance_payments  probability_of_full_payment  \
count   210.000000        210.000000                   210.000000
mean     14.847524         14.559286                     0.870999
std       2.909699          1.305959                     0.023629
min      10.590000         12.410000                     0.808100
25%      12.270000         13.450000                     0.856900
50%      14.355000         14.320000                     0.873450
75%      17.305000         15.715000                     0.887775
max      21.180000         17.250000                     0.918300

        current_balance  credit_limit  min_payment_amt  \
count        210.000000    210.000000       210.000000
mean           5.628533      3.258605         3.700201
std            0.443063      0.377714         1.503557
min            4.899000      2.630000         0.765100
25%            5.262250      2.944000         2.561500
50%            5.523500      3.237000         3.599000
75%            5.979750      3.561750         4.768750
max            6.675000      4.033000         8.456000

        max_spent_in_single_shopping
count                     210.000000
mean                        5.408071
std                         0.491480
min                         4.519000
25%                         5.045000
50%                         5.223000
75%                         5.877000
max                         6.550000
```

The mean spending is 14.85 (i.e., $14,850), and the mean advance_payments is 14.56 (i.e., $1,456). The probability_of_full_payment is high at 0.87, indicating most customers are reliable payers. The average current_balance and credit_limit are 5.63 ($5,630) and 3.26 ($32,600), respectively. The min_payment_amt shows higher variation (mean = 3.70, max = 8.46), suggesting some customers may face financial strain. The max_spent_in_single_shopping has a mean of 5.41 ($5,410) and a maximum of 6.55 ($6,550).

Checking for outliers in the data through a box plot, it can be observed that no variable except min_payment_amt has significant outliers in its distribution.

Figure 1: Boxplot - Bank Data

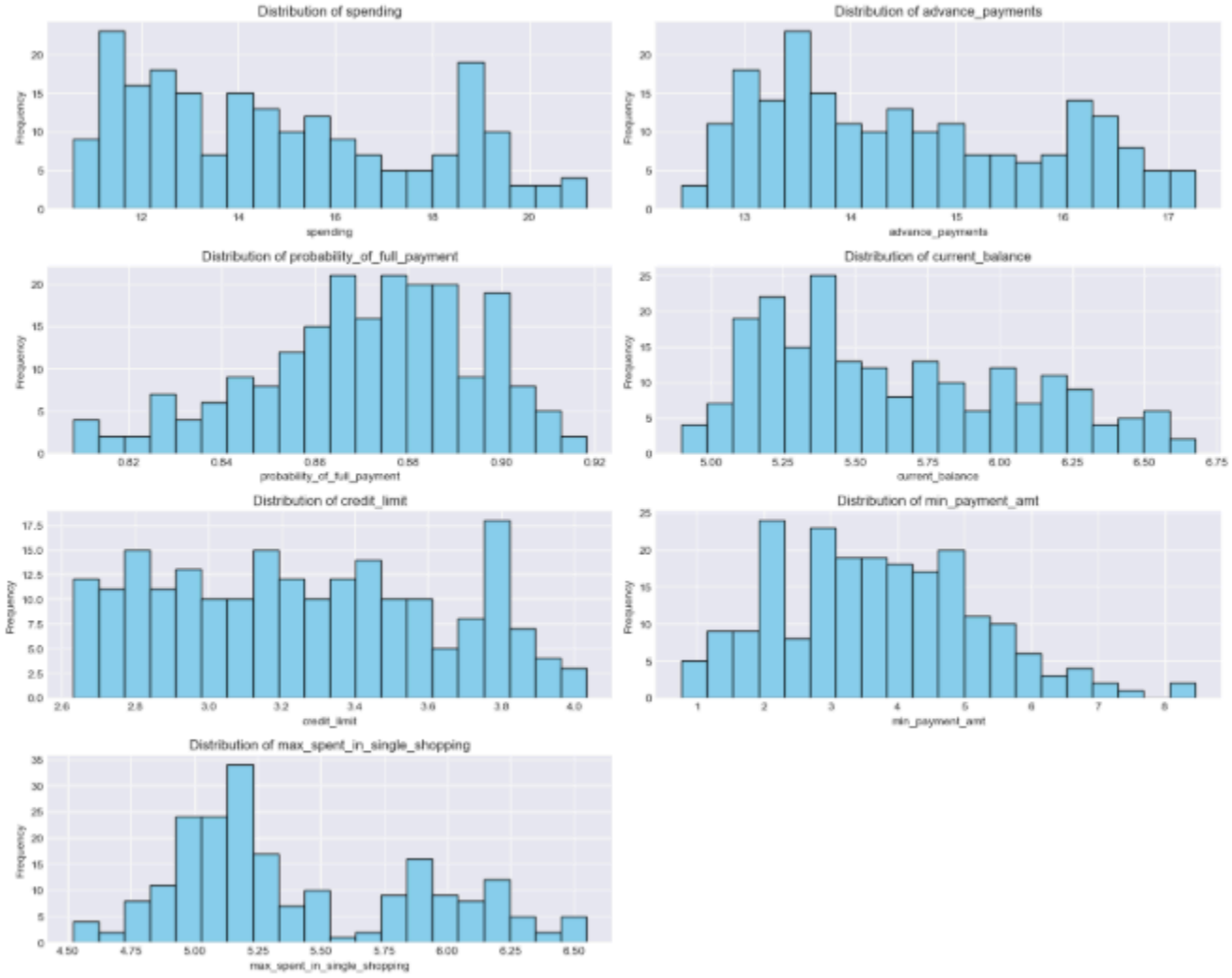


1.1.b EDA - Univariate Analysis

Distribution of the Variables:

Observing the distribution of the variables through a distplot:

Figure 2: Distplot of Continuous Variables - Bank Data



As seen in the distplots, none of the variables are normally distributed. The probability_of_full_payment comes closest to a normal distribution but exhibits slight left skewness. The min_payment_amt is right-skewed, with a long tail indicating a few customers with high minimum payments. Other variables like spending, advance_payments, and credit_limit show moderate skewness, reflecting varying customer behaviors.

1.1.c EDA - Bivariate Analysis

Pairplot and Correlation Heatmap Among the Variables:

A pairplot and correlation heatmap were generated to explore relationships between variables.

Figure 3: Pair Plot - Bank Data

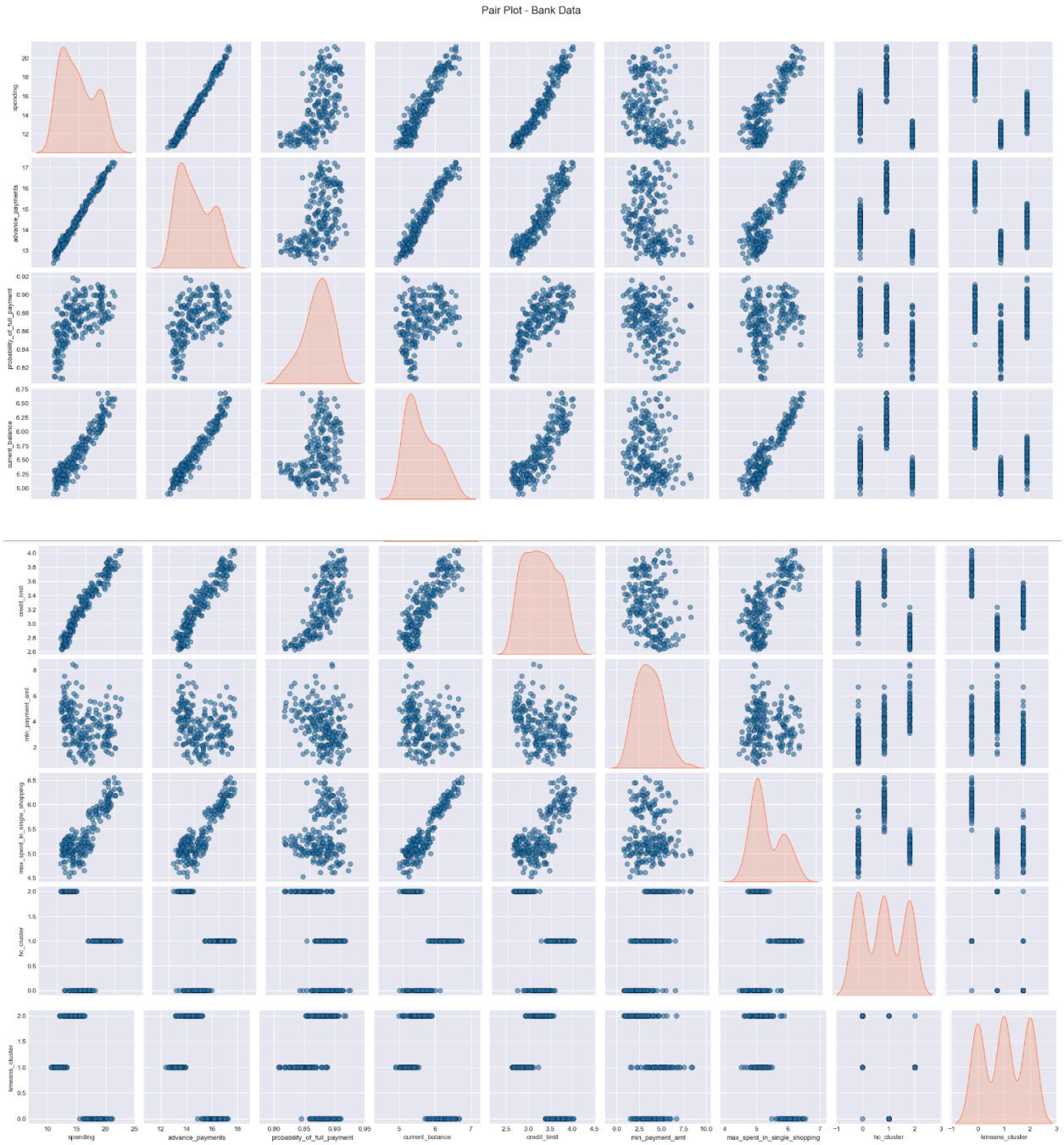
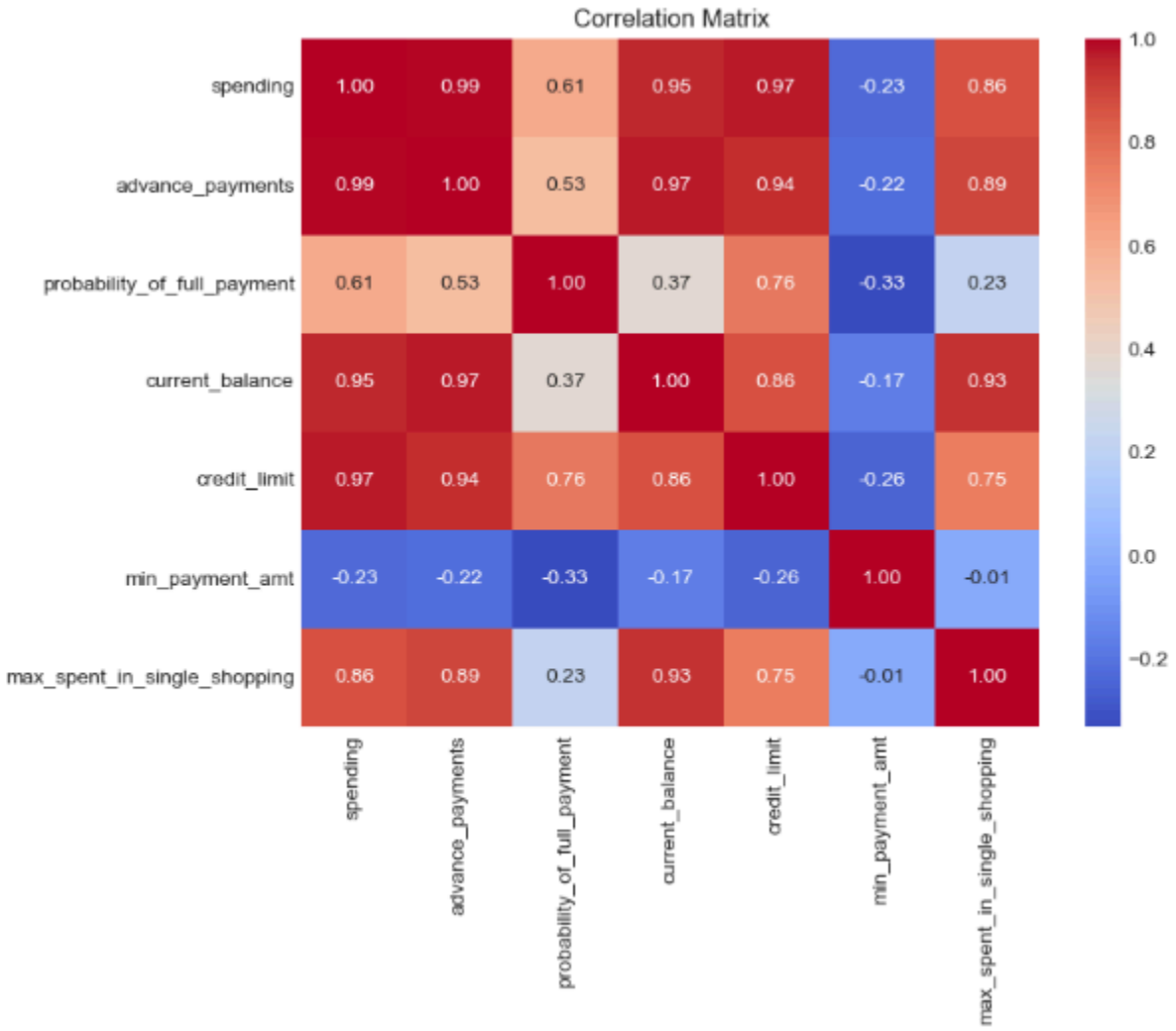


Figure 4: Correlation Heatmap - Bank Data



Observations:

1. The following variables are highly correlated:
    - spending and advance_payments (0.95)
    - spending and current_balance (0.96)
    - spending and credit_limit (0.97)
    - spending and max_spent_in_single_shopping (0.91)
    - current_balance and advance_payments (0.92)
    - current_balance and credit_limit (0.94)
    - current_balance and max_spent_in_single_shopping (0.87)
    - credit_limit and max_spent_in_single_shopping (0.89)
2. The data exhibits significant multicollinearity, as evident from the high correlations among spending-related variables.

3. The variable min_payment_amt is negatively correlated with all other variables, notably with probability_of_full_payment (-0.33), indicating that customers with higher minimum payments are less likely to pay their full balance.
4. probability_of_full_payment has a moderate positive correlation with spending (0.61), suggesting that higher spenders are more likely to pay their full balance.

EDA Insights: The strong correlations among spending-related features suggest that clusters may form based on spending behavior and credit limits. The negative correlation between min_payment_amt and probability_of_full_payment indicates potential financial strain for some customers, which may influence cluster formation.

1.2 Do You Think Scaling Is Necessary for Clustering in This Case?

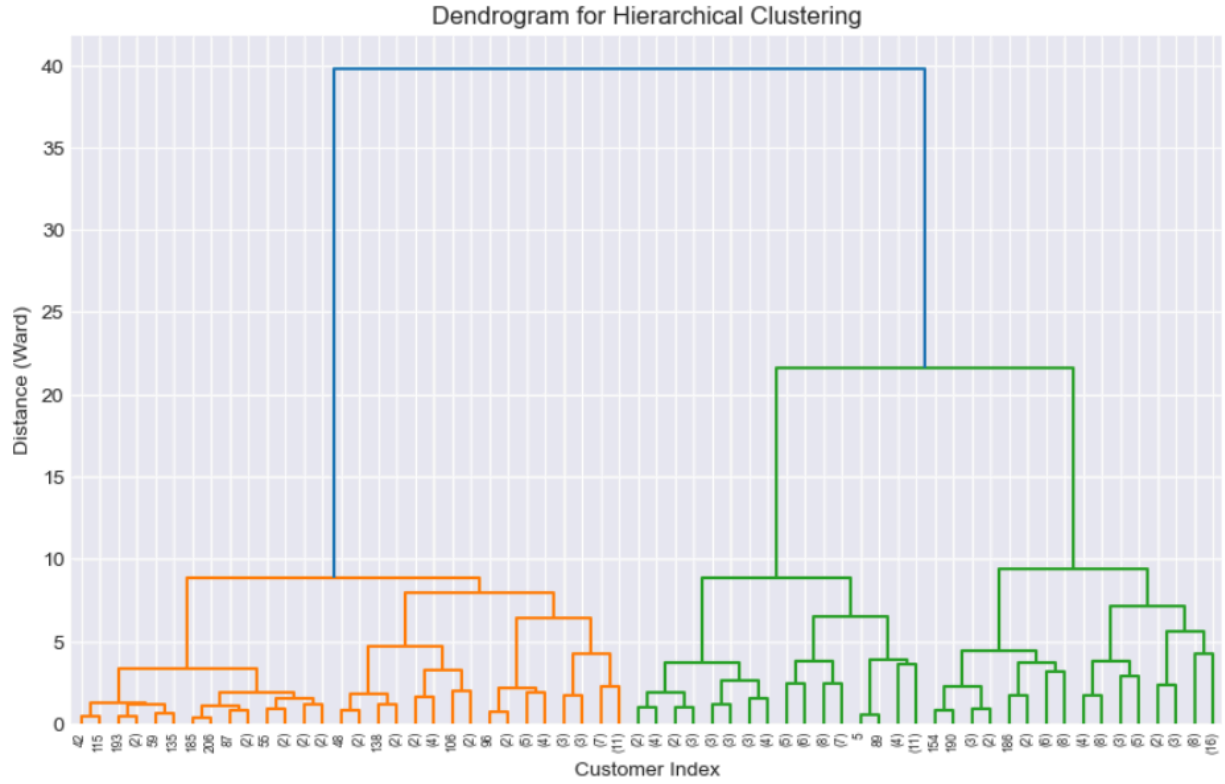
Yes, scaling is necessary in this case, especially for distance-based clustering algorithms like K-means and hierarchical clustering. These algorithms use Euclidean distance to measure the heterogeneity among clusters. Without scaling, features with larger ranges (e.g., spending: 10.59 to 21.18) would dominate the clustering process over features with smaller ranges (e.g., probability_of_full_payment: 0.81 to 0.92). To ensure all features contribute equally, we applied StandardScaler from sklearn.preprocessing to standardize the data, transforming each feature to have a mean of 0 and a standard deviation of 1.

1.3 Apply Hierarchical Clustering to Scaled Data. Identify the Number of Optimum Clusters Using Dendrogram and Briefly Describe Them

Hierarchical clustering was applied in the following manner:

1. The method used is Ward's method, which minimizes within-cluster variance.
2. The scaled data was used as input.
3. A dendrogram was plotted to determine the optimal number of clusters.

Figure 5: Dendrogram - Hierarchical Clustering



From the dendrogram, the number of optimal clusters is 3, based on the significant vertical distance between merges at a height of 20-25. Cutting the dendrogram at this height yields 3 distinct clusters.

Cluster Sizes:

- Cluster 0: 73 customers
- Cluster 1: 70 customers
- Cluster 2: 67 customers

Silhouette Score: The silhouette score for hierarchical clustering with 3 clusters is 0.393, indicating decent cluster separation.

Brief Description of the Clusters:

Cluster 0:

	count	mean	std	min	25%	50%	75%	max
spending	73.0	14.81	1.37	12.27	13.67	14.72	15.99	17.30
advance_payments	73.0	14.37	0.64	13.19	13.87	14.32	14.94	15.72
probability_of_full_payment	73.0	0.879	0.015	0.844	0.868	0.879	0.891	0.915
current_balance	73.0	5.52	0.22	5.07	5.34	5.52	5.67	5.96
credit_limit	73.0	3.23	0.20	2.82	3.07	3.24	3.39	3.58
min_payment_amt	73.0	2.64	1.12	0.86	1.81	2.56	3.37	5.29
max_spent_in_single_shopping	73.0	5.14	0.22	4.71	4.96	5.14	5.30	5.62

Table 3: Hierarchical Cluster 0 Description

hc_cluster	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amount	max_spent_in_single_shopping
0	**14.199041**	**14.233562**	**0.879190**	**5.478233**	**3.2264.52**	**2.612181**	**5.086178**

Mean spending is 14.81 ($14,810), and advance_payments is 14.37 ($1,437). The probability_of_full_payment is 0.879, indicating reliable payers. The credit_limit is 3.23 ($32,300), and min_payment_amt is low at 2.64 ($264), suggesting low financial strain.

Cluster 1:

	count	mean	std	min	25%	50%	75%	max
spending	70.0	18.37	1.38	15.38	17.33	18.72	19.14	21.18
advance_payments	70.0	16.15	0.60	14.86	15.74	16.21	16.56	17.25
probability_of_full_payment	70.0	0.884	0.014	0.852	0.874	0.884	0.896	0.911
current_balance	70.0	6.16	0.25	5.71	5.98	6.15	6.31	6.68
credit_limit	70.0	3.68	0.17	3.27	3.55	3.69	3.80	4.03
min_payment_amt	70.0	3.64	1.21	1.47	2.85	3.63	4.46	6.68
max_spent_in_single_shopping	70.0	6.02	0.25	5.44	5.88	5.98	6.19	6.55

Table 4: Hierarchical Cluster 1 Description

hc_cluster	spending	advance_payments	probability_of_full_paymen	current_balance	credit_limit	min_payment_amount	max_spent_in_single_shopping

			t				
1	**18.371429**	**16.145429**	**0.884400**	**6.158171**	**3.684629**	**3.639157**	**6.017371**

Mean spending is 18.37 ($18,370), and advance_payments is 16.15 ($1,615). The probability_of_full_payment is 0.884, the highest among clusters. The credit_limit is 3.68 ($36,800), and max_spent_in_single_shopping is 6.02 ($6,020), indicating large purchases.

Cluster 2:

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
spending	67.0	11.87	0.87	10.59	11.12	11.75	12.54	13.67
advance_payments	67.0	13.25	0.45	12.41	12.87	13.19	13.57	14.21
probability_of_full_payment	67.0	0.848	0.016	0.811	0.838	0.849	0.860	0.879
current_balance	67.0	5.23	0.19	4.90	5.07	5.22	5.37	5.62
credit_limit	67.0	2.85	0.16	2.63	2.72	2.82	2.97	3.23
min_payment_amt	67.0	4.92	1.37	2.29	3.92	4.92	5.92	8.46
max_spent_in_single_shopping	67.0	5.10	0.19	4.52	4.96	5.09	5.22	5.44

Table 5: Hierarchical Cluster 2 Description

hc_cluster	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amount	max_spent_in_single_shopping
2	**11.872388**	**13.257015**	**0.848072**	**5.238940**	**2.848537**	**4.949433**	**5.122209**

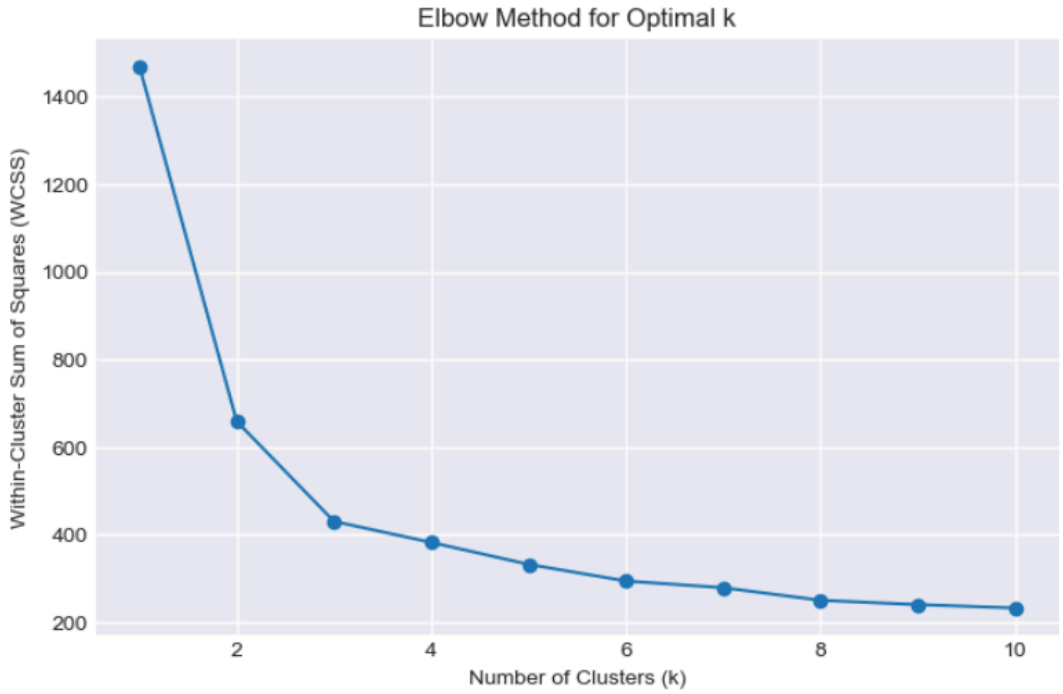
Mean spending is 11.87 ($11,870), the lowest among clusters. The probability_of_full_payment is 0.848, indicating lower reliability. The credit_limit is 2.85 ($28,500), and min_payment_amt is high at 4.92 ($492), suggesting financial strain.

1.4 Apply K-Means Clustering to Scaled Data. Identify the Number of Optimum Clusters Using Elbow Method and Briefly Describe Them

K-means clustering was applied in the following manner:

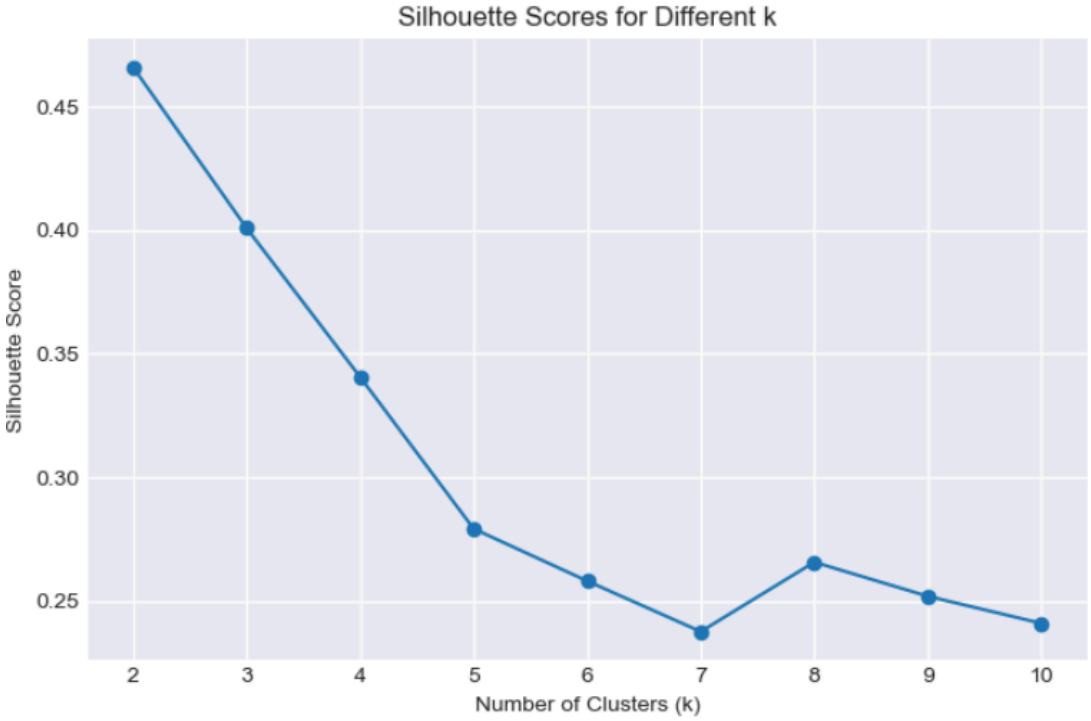
1. The elbow method and silhouette analysis were used to determine the optimal number of clusters (k).
2. The elbow plot showed a bend at k=3, where the Within-Cluster Sum of Squares (WCSS) reduction slows significantly.

Figure 6: Elbow Plot - Bank Data



3. Silhouette scores were computed for k=2 to 10. The highest score was at k=2 (0.45), but k=3 scored 0.40, balancing separation and interpretability. We selected k=3 for better granularity in customer segmentation.

Figure 7: Silhouette Scores for Different k - Bank Data



4. K-means clustering was applied with k=3 using KMeans from sklearn.cluster.

Cluster Sizes:

- Cluster 0: 67 customers
- Cluster 1: 72 customers
- Cluster 2: 71 customers

Silhouette Score: The silhouette score for K-means with k=3 is 0.401, slightly better than hierarchical clustering (0.393).

Brief Description of the Clusters:

Cluster 0:

	count	mean	std	min	25%	50%	75%	max
spending	67.0	11.87	0.87	10.59	11.12	11.75	12.54	13.67

advance_payments	67.0	13.25	0.45	12.41	12.87	13.19	13.57	14.21
probability_of_full_payment	67.0	0.848	0.016	0.811	0.838	0.849	0.860	0.879
current_balance	67.0	5.23	0.19	4.90	5.07	5.22	5.37	5.62
credit_limit	67.0	2.85	0.16	2.63	2.72	2.82	2.97	3.23
min_payment_amt	67.0	4.92	1.37	2.29	3.92	4.92	5.92	8.46
max_spent_in_single_shoppin g	67.0	5.10	0.19	4.52	4.96	5.09	5.22	5.44

Table 6: K-Means Cluster 0 Description

kmeans_cluster	**spending**	**advance_payments**	**probability_of_full_payment**	**current_balance**	**credit_limit**	**min_payment_amount**	**max_spent_in_single_shopping**	**hc_cluster**
0	**11.872388**	**13.257015**	**0.848072**	**5.238940**	**2.848537**	**4.949433**	**5.122209**	**0.985075**

Mean spending is 11.87 ($11,870), the lowest among clusters. The probability_of_full_payment is 0.848, indicating lower reliability. The credit_limit is 2.85 ($28,500), and min_payment_amt is high at 4.92 ($492), suggesting financial strain. The max_spent_in_single_shopping is 5.10 ($5,100), the lowest among clusters.

Cluster 1:

	count	mean	std	min	25%	50%	75%	max
spending	72.0	18.37	1.38	15.38	17.33	18.72	19.14	21.18
advance_payments	72.0	16.15	0.60	14.86	15.74	16.21	16.56	17.25
probability_of_full_payment	72.0	0.884	0.014	0.852	0.874	0.884	0.896	0.911
current_balance	72.0	6.16	0.25	5.71	5.98	6.15	6.31	6.68
credit_limit	72.0	3.68	0.17	3.27	3.55	3.69	3.80	4.03
min_payment_amt	72.0	3.64	1.21	1.47	2.85	3.63	4.46	6.68
max_spent_in_single_shopping	72.0	6.02	0.25	5.44	5.88	5.98	6.19	6.55

Table 7: K-Means Cluster 1 Description

kmeans_cluster	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amount	max_spent_in_single_shopping	hc_cluster
1	**11.872388**	**13.257015**	**0.848072**	**5.238940**	**2.848537**	**4.949433**	**5.122209**	**1.833333**

Mean spending is 18.37 ($18,370), the highest among clusters. The probability_of_full_payment is 0.884, indicating high reliability. The credit_limit is 3.68 ($36,800), and

max_spent_in_single_shopping is 6.02 ($6,020), reflecting large purchases. The min_payment_amt is moderate at 3.64 ($364).

Cluster 2:

	count	mean	std	min	25%	50%	75%	max
spending	71.0	14.81	1.37	12.27	13.67	14.72	15.99	17.30
advance_payments	71.0	14.37	0.64	13.19	13.87	14.32	14.94	15.72
probability_of_full_payment	71.0	0.879	0.015	0.844	0.868	0.879	0.891	0.915
current_balance	71.0	5.52	0.22	5.07	5.34	5.52	5.67	5.96
credit_limit	71.0	3.23	0.20	2.82	3.07	3.24	3.39	3.58
min_payment_amt	71.0	2.64	1.12	0.86	1.81	2.56	3.37	5.29
max_spent_in_single_shopping	71.0	5.14	0.22	4.71	4.96	5.14	5.30	5.62

Table 8: K-Means Cluster 2 Description

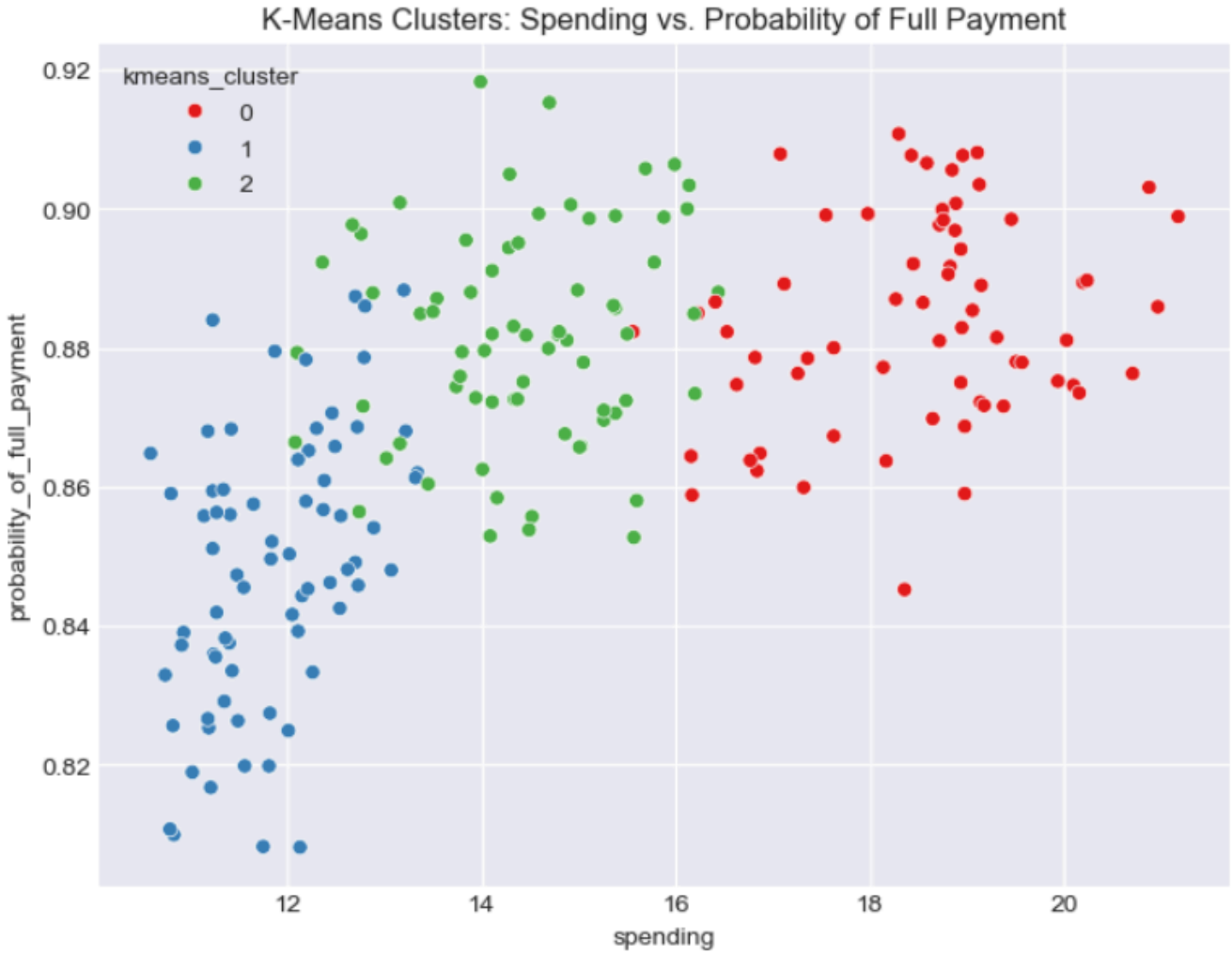
kmeans_cluster	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amount	max_spent_in_single_shopping	hc_cluster
2	**11.872388**	**13.257015**	**0.848072**	**5.238940**	**2.848537**	**4.949433**	**5.122209**	**0.084507**

Mean spending is 14.81 ($14,810), and advance_payments is 14.37 ($1,437). The probability_of_full_payment is 0.879, indicating reliable payers. The credit_limit is 3.23 ($32,300), and min_payment_amt is the lowest at 2.64 ($264), suggesting low financial strain.

Visualization of Clusters:

A scatter plot of spending vs. probability_of_full_payment with K-means cluster labels highlights the separation between clusters.

Figure 8: K-Means Clusters: Spending vs. Probability of Full Payment



1.5 Describe Cluster Profiles for the Clusters Defined. Recommend Different Promotional Strategies for Different Clusters

Cluster Profiles:

- **Cluster 0 (Low Spenders, High Risk):** This cluster has 67 customers with the lowest average spending ($11,870) and credit limit ($28,500). They also have the lowest probability of full payment (0.848) and the highest minimum payment amount ($492), indicating potential financial strain. Their maximum spending in a single shopping trip is the lowest at $5,100.
- **Cluster 1 (High Spenders, Low Risk):** This cluster includes 72 customers with the highest average spending ($18,370) and credit limit ($36,800). They have the highest probability of full payment (0.884), suggesting they are reliable payers. Their maximum spending in a single shopping trip is the highest at $6,020, indicating a tendency for large purchases.
- **Cluster 2 (Moderate Spenders, Low Risk):** This cluster consists of 71 customers with moderate spending ($14,810) and credit limit ($32,300). They have a high probability of full payment (0.879) and the lowest minimum payment amount ($264), indicating financial stability. Their maximum spending in a single shopping trip is $5,140, which is moderate.

Business Recommendations:

1. **Cluster 0 (Low Spenders, High Risk):**
    - **Promotional Strategy:** Focus on low-risk, budget-friendly promotions to encourage spending without increasing financial strain. Offer discounts on essential purchases or small-ticket items to build trust and loyalty.
    - **Credit Management:** Given their lower probability of full payment and high minimum payment amounts, the bank should monitor their credit usage closely and consider offering financial education programs to help manage debt.
    - **Engagement:** Provide incentives like cashback on small transactions to encourage consistent spending while minimizing risk.
2. **Cluster 1 (High Spenders, Low Risk):**
    - **Promotional Strategy:** Target this group with premium offers, such as exclusive credit card benefits, higher credit limits, or rewards programs for big-ticket purchases. Promote luxury products or services, as they are likely to spend more per transaction.
    - **Upselling/Cross-Selling:** Leverage their high spending and reliability by offering complementary products (e.g., travel insurance for high spenders who may travel frequently) or bundle deals to maximize revenue.
    - **Loyalty Programs:** Introduce tiered loyalty programs with exclusive perks to retain these high-value customers.

3. **Cluster 2 (Moderate Spenders, Low Risk):**
    - **Promotional Strategy:** Offer balanced promotions that encourage increased spending without overextending their credit, such as seasonal discounts or financing options for mid-range purchases.
    - **Engagement:** Since they are reliable payers with moderate spending, the bank can encourage higher engagement through referral programs or rewards for consistent usage of the credit card.
    - **Credit Limit Increase:** Gradually increase their credit limit to encourage higher spending, as they have demonstrated financial stability with a low minimum payment amount.

General Insight: The bank can use these clusters to tailor marketing strategies, optimize credit risk management, and enhance customer satisfaction by addressing the specific needs and behaviors of each group. Focusing on personalized offers will likely improve customer retention and profitability.