greatlearning
Power Ahead

Banking Customer Segmentation

Name Bhushan Rai
PGP-DSBA Online
January' 21
Date: 27/06/2021

greatlearning
Power Ahead

Table of Contents

Contents

List of Figures

List of Tables

Executive Summary

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

Introduction

The purpose is to explore the data set and find the spending areas of the customers as accordance to their credit profile, so promotional offers can be provided based on their transaction history.

Data Description

1: spending: Amount spent by the customer per month (in 1000s)

2: advance_payments: Amount paid by the customer in advance by cash (in 100s)

3: probability_of_full_payment: Probability of payment done in full by the customer to the bank
4:current_balance: Balance amount left in the account to make purchases (in 1000s)

5:credit_limit: Limit of the amount in credit card (10000s)

6:min_payment_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)

7:max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s) Import the necessary libraries and load the dataset.

Sample of the dataset:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

Table 1. Dataset Sample

Exploratory Data Analysis

Let us check the types of variables in the data frame

```
 #   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
```

The dataset has 210 observations with dtyes (7)as float

Check for missing values in the dataset:

```
spending                      0
advance_payments              0
probability_of_full_payment   0
current_balance               0
credit_limit                  0
min_payment_amt               0
max_spent_in_single_shopping  0
dtype: int64
```

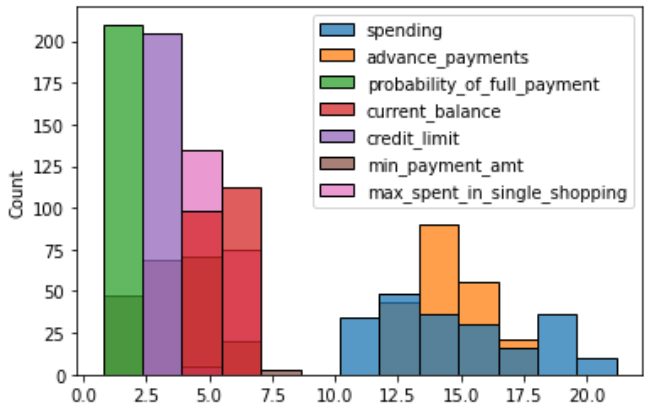
No missing values present in dataset

```
The number of rows in dataset is 210 .
The number of columns in dataset is 10 .
```

Descriptive Statistics:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

```
<AxesSubplot:ylabel='Count'>
```

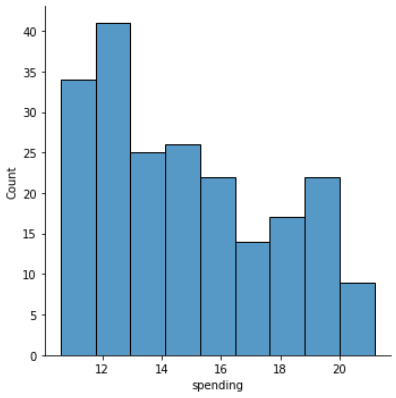


The data seems good and evenly distributed , mean and medium are almost equal. Standard deviation is high for spending varibale

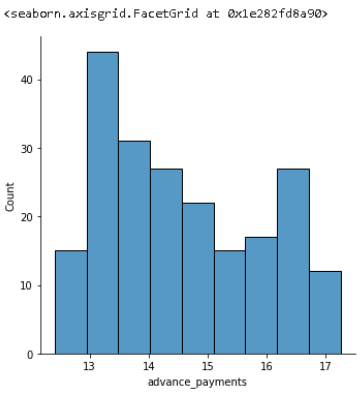
1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis.

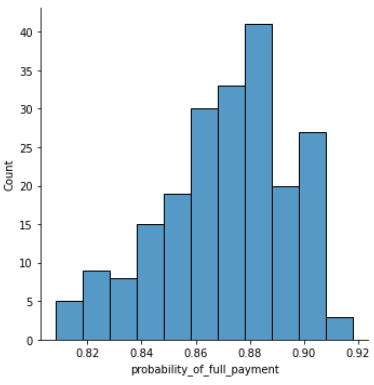
Univariate Analysis:

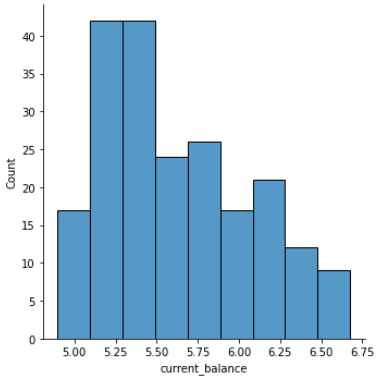
```
<Figure size 648x648 with 0 Axes>
```

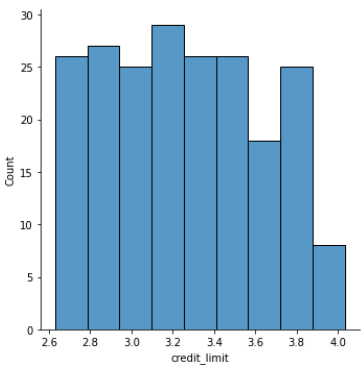


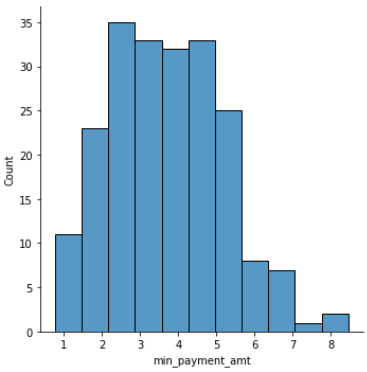

```
<seaborn.axisgrid.FacetGrid at 0x1e282fd8a90>
```

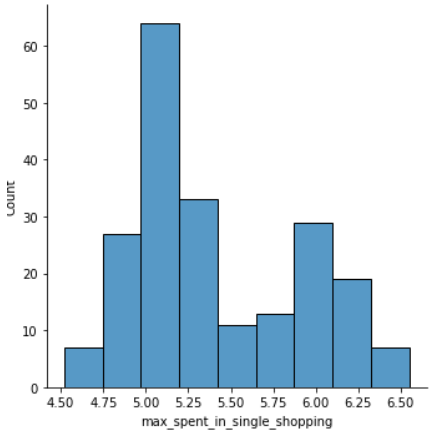






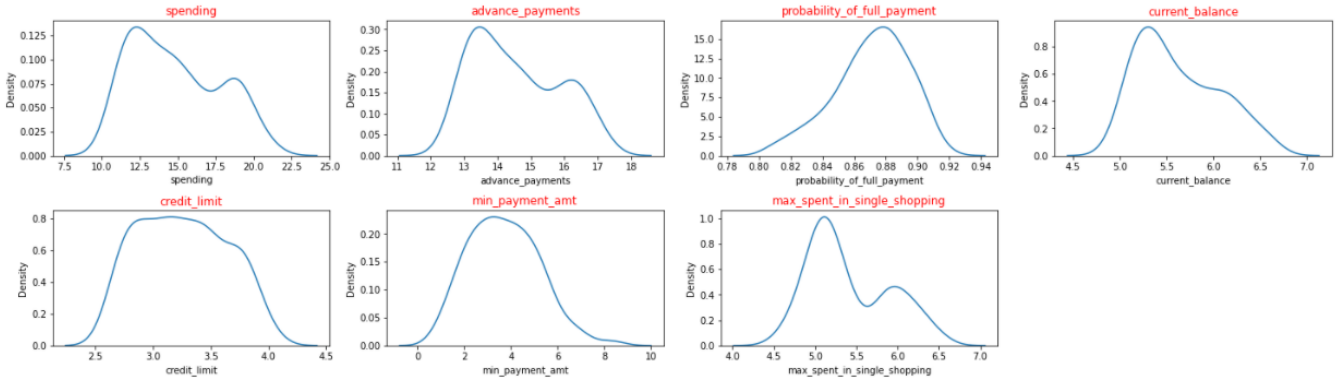






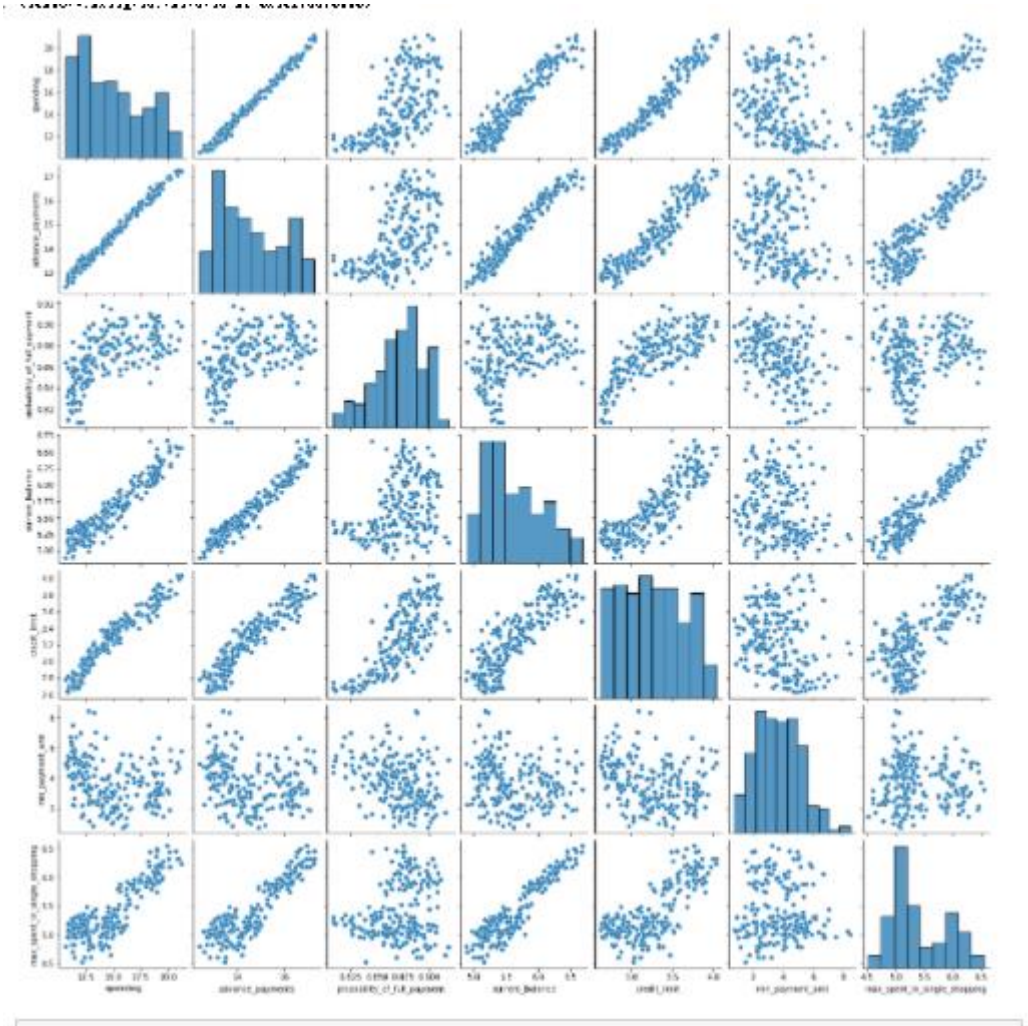
Calculate the skewness in the dataset:

```
max_spent_in_single_shopping    0.561897
current_balance                 0.525482
min_payment_amt                 0.401667
spending                        0.399889
advance_payments                0.386573
credit_limit                    0.134378
probability_of_full_payment    -0.537954
dtype: float64
```



Data is rightly skewed for all variable, except for probability_of_full_payment which is left skewed

Bivariate Analysis



All variables are highly correlated to each other

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
spending	1.000000	0.994341	0.608288	0.949985	0.970771	-0.229572	0.863693
advance_payments	0.994341	1.000000	0.529244	0.972422	0.944829	-0.217340	0.890784
ability_of_full_payment	0.608288	0.529244	1.000000	0.367915	0.761635	-0.331471	0.226825
current_balance	0.949985	0.972422	0.367915	1.000000	0.860415	-0.171562	0.932806
credit_limit	0.970771	0.944829	0.761635	0.860415	1.000000	-0.258037	0.749131
min_payment_amt	-0.229572	-0.217340	-0.331471	-0.171562	-0.258037	1.000000	-0.011079
ent_in_single_shopping	0.863693	0.890784	0.226825	0.932806	0.749131	-0.011079	1.000000

Correlation Plot

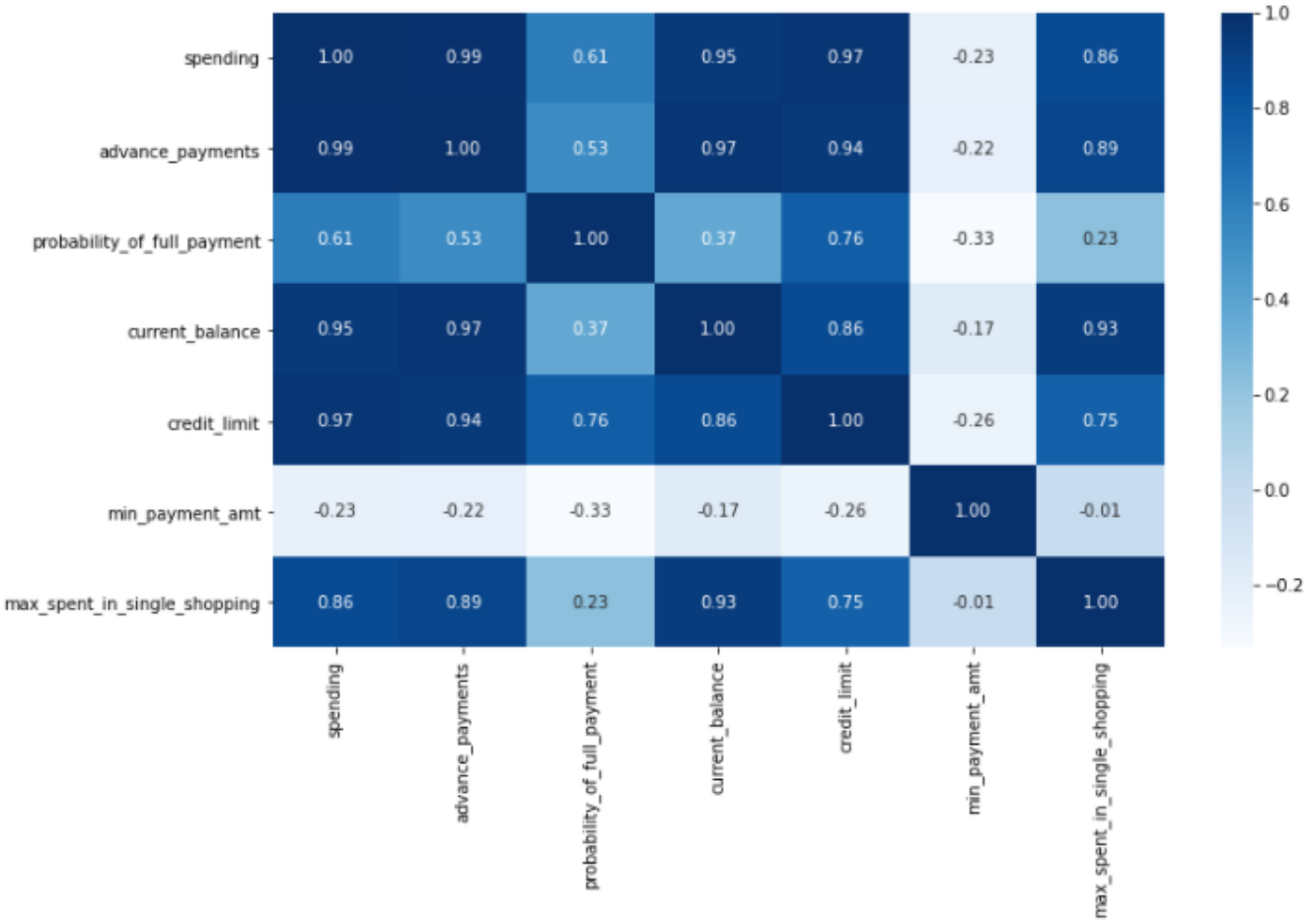
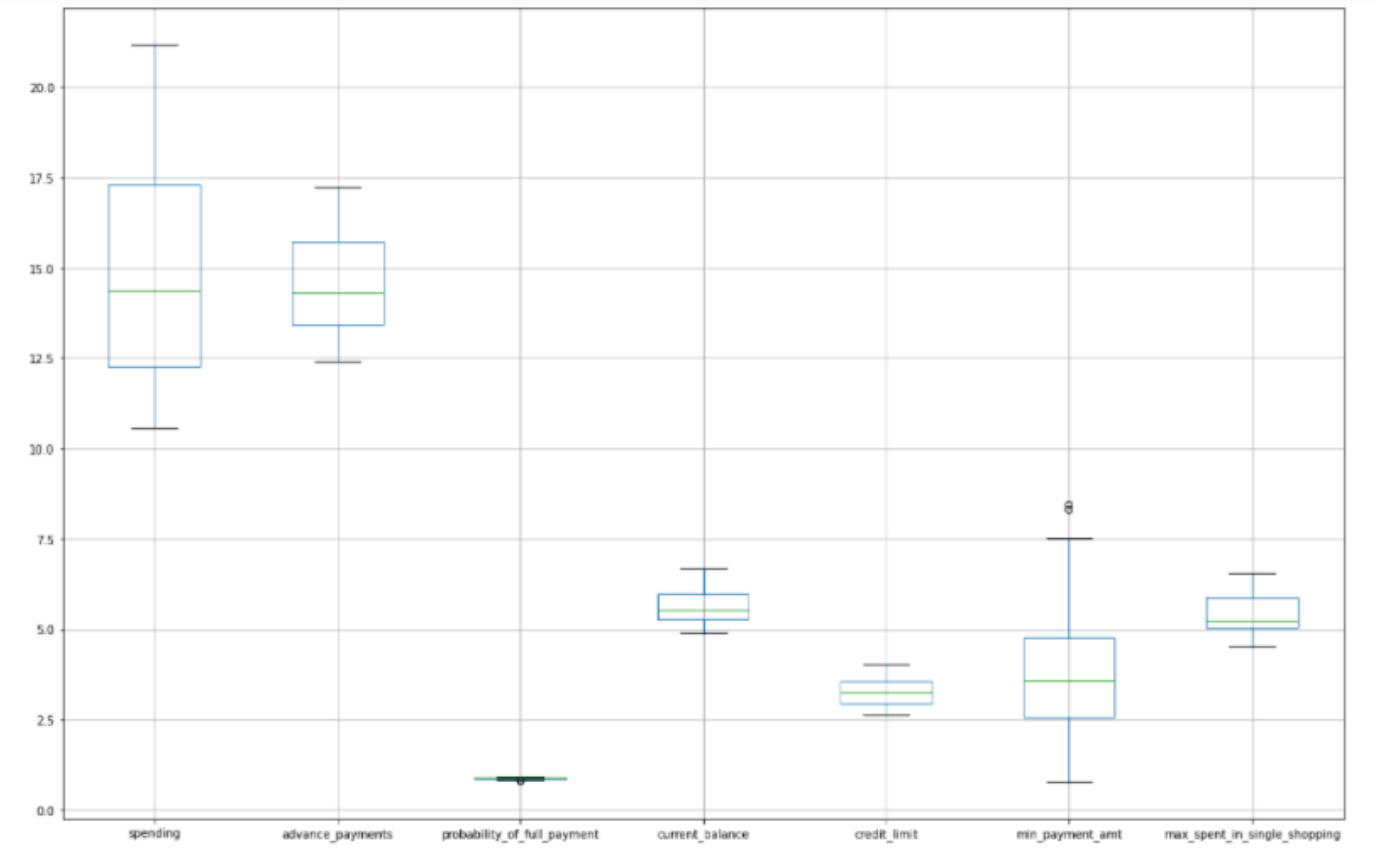


Fig.1 – Correlation Heatmap

Check Outliers:



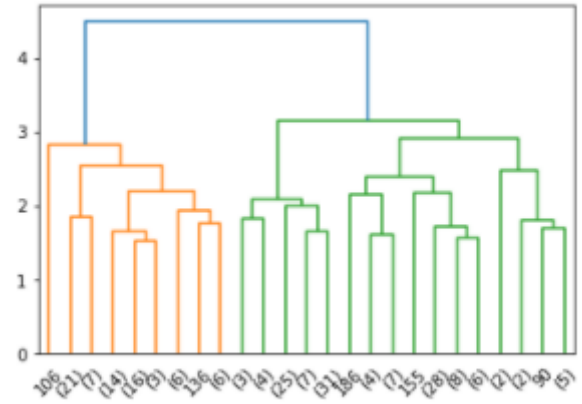
1. No missing values found
2. There are outliers present in only 2 variables: min_payment_amt and probability_of_full_payment
3. There is a small outlier hence no treatment is needed

1.2 Do you think scaling is necessary for clustering in this case? Justify

Yes, it's necessary as we need to rescale the data for further clustering use as the variables are different from each other and range needs to be added

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.178230	2.367533	1.338579	-0.298806	2.328998
1	0.393582	0.253840	1.501773	-0.600744	0.858236	-0.242805	-0.538582
2	1.413300	1.428192	0.504874	1.401485	1.317348	-0.221471	1.509107
3	-1.384034	-1.227533	-2.591878	-0.793049	-1.639017	0.987884	-0.454961
4	1.082581	0.998364	1.196340	0.591544	1.155464	-1.088154	0.874813

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	H_clusters
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1

Cluster Frequency:

```
1    75
2    70
3    65
Name: H_clusters, dtype: int64
```

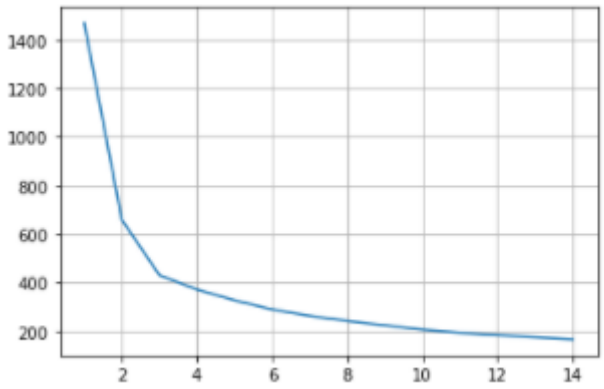
rs	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
1	18.129200	16.058000	0.881595	6.135747	3.648120	3.650200	5.987040	75
2	11.916857	13.291000	0.846766	5.258300	2.846000	4.619000	5.115071	70
3	14.217077	14.195846	0.884869	5.442000	3.253508	2.768418	5.055569	65

The observation for clustering would nominal be 3, based on the hierarchical clustering we have a pattern of high, medium and low spending with variables max_spent_in single_shopping and probability_of_full_payment.

1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

Within sum of squares ranging from 1 to 15:

```
[1469.9999999999998,
 659.171754487041,
 430.6589731513006,
 371.38509060801096,
 327.21278165661346,
 289.31599538959495,
 262.98186570162267,
 241.81894656086033,
 223.91254221002725,
 206.39612184786694,
 193.2835133180646,
 182.97995389115258,
 175.11842017053073,
 166.02965682631788]
```



Its observed there are 3 to 4 points however we will go with 3 points for this Calculation.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

H_clusters	KMeans_clusters	sil_width
1	2	0.573699
3	0	0.366386
1	2	0.637784
2	1	0.512458
1	2	0.362276

The optimal number of clusters here would be 3.

spendi	advanc	probab	current	credit_	min_pa	max_spent_in	H_clust	KMean	sil_wid
19.94	16.92	0.8752	6.675	3.763	3.252	6.55	1	2	0.573699
18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1	2	0.637784
17.99	15.86	0.8992	5.89	3.694	2.068	5.837	1	2	0.362276
18.17	16.26	0.8637	6.271	3.512	2.853	6.273	1	2	0.520285
18.55	16.22	0.8865	6.153	3.674	1.738	5.894	1	2	0.467592
18.98	16.57	0.8687	6.449	3.552	2.144	6.453	1	2	0.524781
17.98	15.85	0.8993	5.979	3.687	2.257	5.919	1	2	0.432773
15.56	14.89	0.8823	5.776	3.408	4.972	5.847	1	2	0.065752
19.51	16.71	0.878	6.366	3.801	2.962	6.185	1	2	0.622415

The KMeans group 2 is the high spending group

spendi	advanc	probab	current	credit_	min_pa	max_spent_in	H_clust	KMean	sil_wid
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	1	0.512458
12.02	13.33	0.8503	5.35	2.81	4.271	5.308	2	1	0.472867
11.23	12.88	0.8511	5.14	2.795	4.325	5.003	2	1	0.532517
12.15	13.45	0.8443	5.417	2.837	3.638	5.338	2	1	0.389668
10.8	12.57	0.859	4.981	2.821	4.773	5.063	2	1	0.499902
13.22	13.84	0.868	5.395	3.07	4.157	5.088	2	1	0.031553
12.7	13.71	0.8491	5.386	2.911	3.26	5.316	2	1	0.235757
12.37	13.47	0.8567	5.204	2.96	3.919	5.001	2	1	0.359037
13.07	13.92	0.848	5.472	2.994	5.304	5.395	2	1	0.366128
12.62	13.67	0.8481	5.41	2.911	3.306	5.231	2	1	0.261362

The KMeans group 1 is the medium spending group

spendi	advanc	probab	current	credit_	min_pa	max_spent_in	H_clust	KMean	sil_wid
13.74	14.05	0.8744	5.482	3.114	2.932	4.825	2	0	0.361812
14.09	14.41	0.8529	5.717	3.186	3.92	5.299	1	0	0.132241
13.78	14.06	0.8759	5.479	3.156	3.136	4.872	2	0	0.377073
15.26	14.85	0.8696	5.714	3.242	4.543	5.314	1	0	0.281318
14.49	14.61	0.8538	5.715	3.113	4.116	5.396	1	0	0.112237
15.38	14.9	0.8706	5.884	3.268	4.462	5.795	1	0	0.005457
15.6	15.11	0.858	5.832	3.286	2.725	5.752	1	0	0.132331
12.74	13.67	0.8564	5.395	2.956	2.504	4.869	2	0	0.007584
16.2	15.27	0.8734	5.826	3.464	2.823	5.527	1	0	0.119541

The KMeans is the lowest spending group

Conclusion and Recommendation:

There are 3 clustering groups with high, medium and low spending

Promotional strategy here can be:

group2: high spending
there can be a raise of the credit limit, reward points as we see the probability of full payment is also high loans can be offered with a good history tracked record of users

group1: medium spending
they are maintaining the account offers like loyalty bonus, increase credit limit provide more customer points to increase spending habits

group0: low spending
give more payment plans so can catch up with balance and offers like daily transaction points should be provided