Clustering code:

**proc** **import** datafile="Z:\Assignments\Graded Assignment\Topic 11.2 - Clustering & Decision Trees\Credit - Data to be used for Clustering.csv"

out=credit dbms=csv replace;

**run**;

/\*Data exploration\*/

**proc** **means** data=credit;

**run**;

**proc** **univariate** data=credit;

**run**;/\*outlier in age variable age=0 so we replace it with avg age \*/

**proc** **freq** data=credit;

table NPA\_status;/\* 2 missing data\*/

**run**;

**proc** **freq** data=credit;

table Gender Region Rented\_Ownhouse Occupation Education MonthlyIncome1 NumberOfDependents;

/\* Gender missing=2 Region=2 Rented\_Ownhouse freq missing=2,Occupation missing=2

Education missing=2 Monthly Income NA=29731 missing freq=2 NumberOfDependents=3924 we replace it with avg no. of dependent and income variable will not being more effecttive as it has large no of missing data there for it is not participating in clustering \*/

**proc** **export** data=credit

outfile="C:\Users\Jig12620\Desktop\jagriti\clusterretail.csv"

dbms=csv replace;

**run**;

**proc** **import** datafile="C:\Users\Jig12620\Desktop\jagriti\clusterretail.xlsx"

out=Credit1 dbms=excel replace;

sheet="clusterretail";

**run**;

**proc** **means** data=credit1;/\*imputed age outlier with avg age\*/

**run**;

**Data** credit1;

set credit1;

Dependent=Input(NumberOfDependents,8.);

**run**;

**data** credit1;

set credit1;

if Dependent=. then Dependent=1;/\*imputed missing value with avg dependent mean because only 2.6% data are missed \*/

**run**;

/\*Data preparation\*/

/\*we convert categorical variable into binary dummy variable and \*/

**data** credit2;

set credit1;

if Occupation="Non-offi" then non\_offi=1;

else non\_offi=0;

if Occupation="Officer1" then officer1=1;

else officer1=0;

if Occupation="Officer2" then officer2=1;

else officer2=0;

if Occupation="Officer3" then officer3=1;

else officer3=0;

if Occupation="Self\_Emp" then self\_emp=1;

else self\_emp=0;

if Region="Centr" then centr=1;

else centr=0;

if Region ="South" then south=1;

else south=0;

if Region="North" then north=1;

north=0;

if Region="East" then east=1;

else east=0;

if Region="West" then west=1;

else west=0;

if Education="PhD" then edu\_phd=1;

else edu\_phd=0;

if Education="Post-Grad" then edu\_pg=1;

else edu\_pg=0;

if Education="Graduate" then edu\_gradu=1;

else edu\_gradu=0;

if Education="Professional" then edu\_prof=1;

else edu\_prof=0;

if Education="Matric" then edu\_matric=1;

else edu\_matric=0;

if Gender="Male" then Gender1=1;

else Gender1=0;

if Rented\_OwnHouse="Ownhouse" then Rented\_OwnHouse1=1;

else Rented\_OwnHouse1=0;

**run**;

/\*scaling\*/

**proc** **standard** data=credit2(keep= NPA\_Status RevolvingUtilizationOfUnsecuredL Gender1 age centr south east west Rented\_OwnHouse1

non\_offi officer1 officer2 officer3 self\_emp edu\_phd edu\_pg edu\_gradu edu\_prof edu\_matric NumberOfTime30\_59DaysPastDueNotW DebtRatio

NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate NumberRealEstateLoansOrLines NumberOfTime60\_89DaysPastDueNotW Dependent)

mean=0 std=1 out=clustering\_data;

var NPA\_Status RevolvingUtilizationOfUnsecuredL Gender1 age centr south east west Rented\_OwnHouse1

non\_offi officer1 officer2 officer3 self\_emp edu\_phd edu\_pg edu\_gradu edu\_prof edu\_matric NumberOfTime30\_59DaysPastDueNotW DebtRatio

NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate NumberRealEstateLoansOrLines NumberOfTime60\_89DaysPastDueNotW Dependent;

**run** ;

/\*weighing\*/

**data** clustering\_data;

set credit2;

NPA\_status1=3\*NPA\_Status;

**run**;

/\*clustering\*/

**proc** **fastclus** data= clustering\_data maxclusters=7 converge=0 maxiter=40

out =cluster\_output

outstat=cluster\_stat;

var NPA\_Status RevolvingUtilizationOfUnsecuredL Gender1 age centr south east west Rented\_OwnHouse1

non\_offi officer1 officer2 officer3 self\_emp edu\_phd edu\_pg edu\_gradu edu\_prof edu\_matric NumberOfTime30\_59DaysPastDueNotW DebtRatio

NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate NumberRealEstateLoansOrLines NumberOfTime60\_89DaysPastDueNotW Dependent;

**run** ;

/\*sorting and adding the cluster to orginal file \*/

**data** cluster\_output1(keep=NPA\_Status cluster);

set cluster\_output;

**run**;

**proc** **sort** data=cluster\_output1;

by NPA\_Status;

**run**;

**Proc** **sort** data=credit2;

by NPA\_Status;

**run**;

**Data** cluster\_credit;

merge cluster\_output1(in=a) credit2(in=b) ;

by NPA\_Status;

if a and b;

**run**;

/\*profiling\*/

**proc** **sort** data=cluster\_credit;

by cluster;

**proc** **means** mean data=cluster\_credit;

var NPA\_Status RevolvingUtilizationOfUnsecuredL Gender1 age centr south east west Rented\_OwnHouse1

non\_offi officer1 officer2 officer3 self\_emp edu\_phd edu\_pg edu\_gradu edu\_prof edu\_matric NumberOfTime30\_59DaysPastDueNotW DebtRatio

NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate NumberRealEstateLoansOrLines NumberOfTime60\_89DaysPastDueNotW Dependent;

by cluster;

**run**;

**proc** **means** mean std data=cluster\_credit;

var NPA\_Status RevolvingUtilizationOfUnsecuredL Gender1 age centr south east west Rented\_OwnHouse1

non\_offi officer1 officer2 officer3 self\_emp edu\_phd edu\_pg edu\_gradu edu\_prof edu\_matric NumberOfTime30\_59DaysPastDueNotW DebtRatio

NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate NumberRealEstateLoansOrLines NumberOfTime60\_89DaysPastDueNotW Dependent;

**run**;

Deciding the number of Cluster:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster Summary** | | | | | | |
| **Cluster** | **Frequency** | **RMS Std Deviation** | **Maximum Distance from Seed to Observation** | **Distance Between Cluster Centroids** | **Ratio(strength metric)** | **proportion** |
| 1 | 11307 | 487.4 | 58415.7 | 3406.3 | 119.851662 | 7.538 |
| 2 | 3 | 2404.5 | 14034.7 | 100520 | 5.83684758 | 0.002 |
| 3 | 34 | 1704 | 38017.7 | 12698.1 | 22.3108568 | 0.02266667 |
| 4 | 4 | 6172.4 | 46837 | 98518 | 7.58813428 | 0.00266667 |
| 5 | 1 | . | 0 | 98518 | #VALUE! | 0.00066667 |
| 6 | 138651 | 61.6969 | 6321.6 | 3406 | 102.462198 | 92.434 |
|  | 150000 |  |  |  | 43.00828 |  |

NUMBER OF CLUSTER:7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster Summary** | | | | | | |
| **Cluster** | **Frequency** | **RMS Std Deviation** | **Maximum Distance from Seed to Observation** | **Distance Between Cluster Centroids** | **Ratio(strength metric)** | **proportion** |
| 1 | 75 | 2306.2 | 38369.2 | 20293.3 | 8.79945365 | 0.05 |
| 2 | 34 | 1704 | 38017.7 | 12698.4 | 7.45211268 | 0.02266667 |
| 3 | 12064 | 316.1 | 10133 | 3169.9 | 10.0281556 | 8.04266667 |
| 4 | 1 | . | 0 | 98518 | #VALUE! | 0.00066667 |
| 5 | 3 | 2404.5 | 14034.7 | 100520 | 41.8049491 | 0.002 |
| 6 | 137819 | 56.5671 | 6321.5 | 3169.9 | 56.0378736 | 91.8793333 |
| 7 | 4 | 6172.4 | 46837 | 98460.1 | 15.951672 | 0.00266667 |
|  |  |  |  |  | 20.01060237 |  |

NUMBER OF CLUSTER:8

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster Summary** | | | | | | |
| **Cluster** | **Frequency** | **RMS Std Deviation** | **Maximum Distance from Seed to Observation** | **Distance Between Cluster Centroids** | **Ratio(strength metric)** | **proportion** |
| 1 | 137819 | 56.5671 | 6321.5 | 3169.9 | 56.0378736 | 91.8793333 |
| 2 | 34 | 1704 | 38017.7 | 12698.4 | 7.45211268 | 0.02266667 |
| 3 | 1 | . | 0 | 51681 | #VALUE! | 0.00066667 |
| 4 | 3 | 948.3 | 5065.7 | 62449.3 | 65.8539492 | 0.002 |
| 5 | 12064 | 316.1 | 10133 | 3169.9 | 10.0281556 | 8.04266667 |
| 6 | 1 | . | 0 | 51681 | #VALUE! | 0.00066667 |
| 7 | 75 | 2306.2 | 38369.2 | 20293.3 | 8.79945365 | 0.05 |
| 8 | 3 | 2404.5 | 14034.7 | 100520 | 41.8049491 | 0.002 |
|  |  |  |  |  | 23.74706 |  |

NUMBER OF CLUSTERS:9

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster Summary** | | | | | | |
| **Cluster** | **Frequency** | **RMS Std Deviation** | **Maximum Distance from Seed to Observation** | **Distance Between Cluster Centroids** | **Ratio(strength metric)** | **proportion** |
| 1 | 1 | . | 0 | 51681 | #VALUE! | 0.00066667 |
| 2 | 34 | 1704 | 38017.7 | 12698.4 | 7.45211268 | 0.02266667 |
| 3 | 137819 | 56.5671 | 6321.5 | 3169.9 | 56.0378736 | 91.8793333 |
| 4 | 75 | 2306.2 | 38369.2 | 20293.3 | 8.79945365 | 0.05 |
| 5 | 1 | . | 0 | 51681 | #VALUE! | 0.00066667 |
| 6 | 1 | . | 0 | 21052 | #VALUE! | 0.00066667 |
| 7 | 3 | 948.3 | 5065.7 | 62449.3 | 65.8539492 | 0.002 |
| 8 | 12064 | 316.1 | 10133 | 3169.9 | 10.0281556 | 8.04266667 |
| 9 | 2 | 446.8 | 1611 | 21052 | 47.1172784 | 0.00133333 |

21.69876

We can see that there is only a little change in proportion and strength of clusters with increase in number of clusters and it's pretty obvious as number of default loan is less than non-default and weight NPA\_status variable 3 times stronger so it is most expected number of clusters.

Here we will take number of cluster 7 as its strength is pretty high and proportion is appropriate as per of analysis and dataset and there for it is best combination of good strength and proportion.

PROFILING OF CLUSTERS:

CLUSTER1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **CMean** | **PMean** | **Std Dev** | **DIFF** | **Z-score** |  |  |  |
| **NPA\_Status** | 75 | 0.13333 | 0.06684 | 0.2497455 | 0.0664933 | **0.26624424** |  |  |  |
| **RevolvingUtilizationOfUnsecuredL** | 75 | 111.36 | 6.04844 | 249.75537 | 105.3111415 | **0.42165716** |  |  |  |
| **Gender1** | 75 | 0.56 | 0.61536 | 0.4865117 | -0.05536 | **-0.11378966** |  |  |  |
| **age** | 75 | 55.1733 | 52.2952 | 14.771866 |  | **0** |  |  |  |
| **centr** | 75 | 0.32 | 0.29304 | 0.4551581 | 0.02696 | **0.05923217** |  |  |  |
| **south** | 75 | 0.09333 | 0.15663 | 0.3634559 |  | **0** |  |  |  |
| **east** | 75 | 0.17333 | 0.13701 | 0.3438555 | 0.0363266 | **0.10564496** |  |  |  |
| **west** | 75 | 0.16 | 0.18599 | 0.3891026 |  | **0** |  |  |  |
| **Rented\_OwnHouse1** | 75 | 0.61333 | 0.57302 | 0.494641 | 0.0403133 | **0.08150012** |  |  |  |
| **non\_offi** | 75 | 0.30667 | 0.27409 | 0.4460544 |  | **0** |  |  |  |
| **officer1** | 75 | 0.12 | 0.10109 | 0.3014533 | 0.0189067 | **0.0627185** |  |  |  |
| **officer2** | 75 | 0.12 | 0.08889 | 0.284581 |  | **0** |  |  |  |
| **officer3** | 75 | 0.09333 | 0.10849 | 0.3110035 | -0.01516 | **-0.04874543** |  |  |  |
| **self\_emp** | 75 | 0.36 | 0.42744 | 0.4947087 |  | **0** |  |  |  |
| **edu\_phd** | 75 | 0.05333 | 0.04201 | 0.200605 | 0.0113266 | **0.0564622** |  |  |  |
| **edu\_pg** | 75 | 0.30667 | 0.24809 | 0.4319077 |  | **0** |  |  |  |
| **edu\_gradu** | 75 | 0.16 | 0.26503 | 0.4413524 | -0.1050333 | **-0.23798058** |  |  |  |
| **edu\_prof** | 75 | 0.4 | 0.33947 | 0.4735299 |  | **0** |  |  |  |
| **edu\_matric** | 75 | 0.08 | 0.1054 | 0.3070692 | -0.0254 | **-0.08271751** |  |  |  |
| **NumberOfTime30\_59DaysPastDueNotW** | 75 | 0.56 | 0.42103 | 4.1927813 |  | **0** |  |  |  |
| **DebtRatio** | 75 | 23538 | 353.005 | 2037.82 | 23184.94492 | **11.3773272** |  |  |  |
| **NumberOfOpenCreditLinesAndLoans** | 75 | 15.5067 | 8.45276 | 5.145951 |  | **0** |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **NumberOfTimes90DaysLate** | 75 | 0.06667 | 0.26597 | 4.1693038 | -0.1993066 | **-0.04780333** |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **NumberRealEstateLoansOrLines** | 75 | 4.98667 | 1.01824 | 1.129771 |  | **0** |  |  |  |
| **NumberOfTime60\_89DaysPastDueNotW** | 75 | 0.10667 | 0.24039 | 4.1551794 | -0.13372 | **-0.03218152** |  |  |  |
| **Dependent** | 75 | 1 | 1 | 0 |  | **#DIV/0!** |  |  |  |

1.In cluster1 NPA\_status is 6.64% is higher than population so must figure out why this high degree of default for this cluter?

2. Debt ratio of cluster1 is much much higher than population mean. Why this huge difference?

3.In this cluster open creditlines and loans mean is 15.5 which much higher than population mean that is 8.45 so must investigate why this difference in open credit line.

4.Number of realEstate Loan or lines is approx. 5 time higher than population mean.

5.There for people of this cluster having high number of credit lines rather than rest of the population.

6.NPA\_status and Debt ratio have highest Z-score.

CLUSTER2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **CMean** | **PMean** | **Std Dev** | **Diff** | **zscore** |
| **NPA\_Status** | 0 | 0.06684 | 0.24975 | -0.0668 | -0.26763245 |
| **RevolvingUtilizationOfUnsecuredL** | 12693.8 | 6.04844 | 249.755 | 12687.7 | 50.8005555 |
| **Gender1** | 0.55882 | 0.61536 | 0.48651 | -0.0565 | -0.11620789 |
| **age** | 48.9118 | 52.2952 | 14.7719 | -3.3834 | -0.22904635 |
| **centr** | 0.29412 | 0.29304 | 0.45516 | 0.00108 | 0.00236753 |
| **south** | 0.17647 | 0.15663 | 0.36346 | 0.01984 | 0.05457966 |
| **east** | 0.17647 | 0.13701 | 0.34386 | 0.03946 | 0.11476885 |
| **west** | 0.14706 | 0.18599 | 0.3891 | -0.0389 | -0.1000623 |
| **Rented\_OwnHouse1** | 0.55882 | 0.57302 | 0.49464 | -0.0142 | -0.02870061 |
| **non\_offi** | 0.23529 | 0.27409 | 0.44605 | -0.0388 | -0.08696832 |
| **officer1** | 0.05882 | 0.10109 | 0.30145 | -0.0423 | -0.14022006 |
| **officer2** | 0.08824 | 0.08889 | 0.28458 | -0.0007 | -0.00228898 |
| **officer3** | 0.08824 | 0.10849 | 0.311 | -0.0203 | -0.06513753 |
| **self\_emp** | 0.52941 | 0.42744 | 0.49471 | 0.10197 | 0.20612494 |
| **edu\_phd** | 0.02941 | 0.04201 | 0.20061 | -0.0126 | -0.06278458 |
| **edu\_pg** | 0.26471 | 0.24809 | 0.43191 | 0.01661 | 0.03846331 |
| **edu\_gradu** | 0.35294 | 0.26503 | 0.44135 | 0.08791 | 0.19917848 |
| **edu\_prof** | 0.29412 | 0.33947 | 0.47353 | -0.0453 | -0.09576819 |
| **edu\_matric** | 0.05882 | 0.1054 | 0.30707 | -0.0466 | -0.15168079 |
| **NumberOfTime30\_59DaysPastDueNotW** | 0.17647 | 0.42103 | 4.19278 | -0.2446 | -0.05832947 |
| **DebtRatio** | 515.092 | 353.005 | 2037.82 | 162.087 | 0.07953931 |
| **NumberOfOpenCreditLinesAndLoans** | 6.5 | 8.45276 | 5.14595 | -1.9528 | -0.37947505 |
| **NumberOfTimes90DaysLate** | 0.02941 | 0.26597 | 4.1693 | -0.2366 | -0.05673885 |
| **NumberRealEstateLoansOrLines** | 1.41176 | 1.01824 | 1.12977 | 0.39352 | 0.34832254 |
| **NumberOfTime60\_89DaysPastDueNotW** | 0 | 0.24039 | 4.15518 | -0.2404 | -0.0578523 |
| **Dependent** | 1 | 1 | 0 | 0 | #DIV/0! |

1.In cluster2 NPA\_status is 0 means all in this cluster all customers are non-defaulter that is different from population mean. Why?

2.In this segment Self-employed customers are 10% higher than population so must figure out why this high?

3.Debt ratio is higher by 162.087.why?

4.In cluster2 number of times90dayslate and number of time60\_89dayspastduenotw are quite lower than population.

5.so from above finding in this cluster customers are pretty good .

Cluster3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **CMean** | **PMean** | **Std Dev** | **DIFF** | **Z-score** |
| **NPA\_Status** | 0.04982 | 0.06684 | 0.24975 | -0.017022 | **-0.06815899** |
| **RevolvingUtilizationOfUnsecuredL** | 4.93339 | 6.04844 | 249.755 | -1.115048 | **-0.00446456** |
| **Gender1** | 0.6134 | 0.61536 | 0.48651 | -0.001965 | **-0.00403855** |
| **age** | 53.8078 | 52.2952 | 14.7719 | 1.5125685 | **0.102395223** |
| **centr** | 0.29833 | 0.29304 | 0.45516 | 0.0052856 | **0.011612668** |
| **south** | 0.15418 | 0.15663 | 0.36346 | -0.002456 | **-0.00675625** |
| **east** | 0.13992 | 0.13701 | 0.34386 | 0.0029137 | **0.008473618** |
| **west** | 0.17987 | 0.18599 | 0.3891 | -0.006119 | **-0.0157267** |
| **Rented\_OwnHouse1** | 0.57352 | 0.57302 | 0.49464 | 0.0005045 | **0.001019932** |
| **non\_offi** | 0.27487 | 0.27409 | 0.44605 | 0.0007807 | **0.001750235** |
| **officer1** | 0.10428 | 0.10109 | 0.30145 | 0.0031839 | **0.010561835** |
| **officer2** | 0.08803 | 0.08889 | 0.28458 | -0.000856 | **-0.00300863** |
| **officer3** | 0.10875 | 0.10849 | 0.311 | 0.00026 | **0.000836003** |
| **self\_emp** | 0.42407 | 0.42744 | 0.49471 | -0.003368 | **-0.00680886** |
| **edu\_phd** | 0.04012 | 0.04201 | 0.20061 | -0.001887 | **-0.00940804** |
| **edu\_pg** | 0.24876 | 0.24809 | 0.43191 | 0.0006633 | **0.001535745** |
| **edu\_gradu** | 0.26981 | 0.26503 | 0.44135 | 0.0047777 | **0.010825137** |
| **edu\_prof** | 0.3382 | 0.33947 | 0.47353 | -0.00127 | **-0.00268283** |
| **edu\_matric** | 0.10312 | 0.1054 | 0.30707 | -0.002283 | **-0.00743578** |
| **NumberOfTime30\_59DaysPastDueNotW** | 0.24619 | 0.42103 | 4.19278 | -0.174846 | **-0.04170175** |
| **DebtRatio** | 3244.98 | 353.005 | 2037.82 | 2891.9749 | **1.419151311** |
| **NumberOfOpenCreditLinesAndLoans** | 9.81051 | 8.45276 | 5.14595 | 1.3577506 | **0.263848334** |
| **NumberOfTimes90DaysLate** | 0.06117 | 0.26597 | 4.1693 | -0.2048 | **-0.04912081** |
| **NumberRealEstateLoansOrLines** | 1.65899 | 1.01824 | 1.12977 | 0.6407454 | **0.567146262** |
| **NumberOfTime60\_89DaysPastDueNotW** | 0.05521 | 0.24039 | 4.15518 | -0.185181 | **-0.04456633** |
| **Dependent** | 1 | 1 | 0 | 0 | **#DIV/0!** |

1.In this cluster customers having 2891 more debt ration than population means, reason should be figure out?

2.Numberoftimes90dayslate is 20%less and numberoftime60\_89dayspastduenotw is 18% less than population mean. Why such difference?

CLUSTER4:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **CMean** | **PMean** | **Std Dev** | **DIFF** | **Z-score** |
| **NPA\_Status** | 0 | 0.06684 | 0.24975 | -0.0668 | **-0.268** |
| **RevolvingUtilizationOfUnsecuredL** | 0.18264 | 6.04844 | 249.755 | -5.8658 | **-0.023** |
| **Gender1** | 1 | 0.61536 | 0.48651 | 0.38464 | **0.7906** |
| **age** | 37 | 52.2952 | 14.7719 | -15.295 | **-1.035** |
| **centr** | 1 | 0.29304 | 0.45516 | 0.70696 | **1.5532** |
| **south** | 0 | 0.15663 | 0.36346 | -0.1566 | **-0.431** |
| **east** | 0 | 0.13701 | 0.34386 | -0.137 | **-0.398** |
| **west** | 0 | 0.18599 | 0.3891 | -0.186 | **-0.478** |
| **Rented\_OwnHouse1** | 0 | 0.57302 | 0.49464 | -0.573 | **-1.158** |
| **non\_offi** | 0 | 0.27409 | 0.44605 | -0.2741 | **-0.614** |
| **officer1** | 0 | 0.10109 | 0.30145 | -0.1011 | **-0.335** |
| **officer2** | 0 | 0.08889 | 0.28458 | -0.0889 | **-0.312** |
| **officer3** | 0 | 0.10849 | 0.311 | -0.1085 | **-0.349** |
| **self\_emp** | 1 | 0.42744 | 0.49471 | 0.57256 | **1.1574** |
| **edu\_phd** | 0 | 0.04201 | 0.20061 | -0.042 | **-0.209** |
| **edu\_pg** | 0 | 0.24809 | 0.43191 | -0.2481 | **-0.574** |
| **edu\_gradu** | 0 | 0.26503 | 0.44135 | -0.265 | **-0.601** |
| **edu\_prof** | 1 | 0.33947 | 0.47353 | 0.66053 | **1.3949** |
| **edu\_matric** | 0 | 0.1054 | 0.30707 | -0.1054 | **-0.343** |
| **NumberOfTime30\_59DaysPastDueNotW** | 0 | 0.42103 | 4.19278 | -0.421 | **-0.1** |
| **DebtRatio** | 220516 | 353.005 | 2037.82 | 220163 | **108.04** |
| **NumberOfOpenCreditLinesAndLoans** | 8 | 8.45276 | 5.14595 | -0.4528 | **-0.088** |
| **NumberOfTimes90DaysLate** | 0 | 0.26597 | 4.1693 | -0.266 | **-0.064** |
| **NumberRealEstateLoansOrLines** | 3 | 1.01824 | 1.12977 | 1.98176 | **1.7541** |
| **NumberOfTime60\_89DaysPastDueNotW** | 0 | 0.24039 | 4.15518 | -0.2404 | **-0.058** |
| **Dependent** | 1 | 1 | 0 | 0 | **#DIV/0!** |

1.IN this cluster4 there is only one observation means this customer is outlier.

2.this customer is male,self\_employed and his NPA\_status is non\_default .He has much higer Debt ratio than population mean about 220163 and the reason is high number of open credit loans and real state loans.

4.With many loan this customer number of times90 day slate and number of time 60\_89days past duenotw is 0 so it define this customer is a good customer and beneficial for bank.

Cluster5:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **PMean** | **Std Dev** | **DIFF** | **Z-score** |
| **NPA\_Status** | 0 | 0.06684 | 0.24975 | -0.0668 | **-0.268** |
| **RevolvingUtilizationOfUnsecuredL** | 0.012024 | 6.04844 | 249.755 | -6.0364 | **-0.024** |
| **Gender1** | 0.6666667 | 0.61536 | 0.48651 | 0.05131 | **0.1055** |
| **age** | 62.6666667 | 52.2952 | 14.7719 | 10.3715 | **0.7021** |
| **centr** | 0.3333333 | 0.29304 | 0.45516 | 0.04029 | **0.0885** |
| **south** | 0 | 0.15663 | 0.36346 | -0.1566 | **-0.431** |
| **east** | 0.6666667 | 0.13701 | 0.34386 | 0.52966 | **1.5404** |
| **west** | 0 | 0.18599 | 0.3891 | -0.186 | **-0.478** |
| **Rented\_OwnHouse1** | 0 | 0.57302 | 0.49464 | -0.573 | **-1.158** |
| **non\_offi** | 0.3333333 | 0.27409 | 0.44605 | 0.05925 | **0.1328** |
| **officer1** | 0 | 0.10109 | 0.30145 | -0.1011 | **-0.335** |
| **officer2** | 0.3333333 | 0.08889 | 0.28458 | 0.24445 | **0.859** |
| **officer3** | 0.3333333 | 0.10849 | 0.311 | 0.22484 | **0.723** |
| **self\_emp** | 0 | 0.42744 | 0.49471 | -0.4274 | **-0.864** |
| **edu\_phd** | 0 | 0.04201 | 0.20061 | -0.042 | **-0.209** |
| **edu\_pg** | 0.3333333 | 0.24809 | 0.43191 | 0.08524 | **0.1974** |
| **edu\_gradu** | 0 | 0.26503 | 0.44135 | -0.265 | **-0.601** |
| **edu\_prof** | 0.6666667 | 0.33947 | 0.47353 | 0.3272 | **0.691** |
| **edu\_matric** | 0 | 0.1054 | 0.30707 | -0.1054 | **-0.343** |
| **NumberOfTime30\_59DaysPastDueNotW** | 0.3333333 | 0.42103 | 4.19278 | -0.0877 | **-0.021** |
| **DebtRatio** | 321035.67 | 353.005 | 2037.82 | 320683 | **157.37** |
| **NumberOfOpenCreditLinesAndLoans** | 10.3333333 | 8.45276 | 5.14595 | 1.88057 | **0.3654** |
| **NumberOfTimes90DaysLate** | 0 | 0.26597 | 4.1693 | -0.266 | **-0.064** |
| **NumberRealEstateLoansOrLines** | 2.6666667 | 1.01824 | 1.12977 | 1.64843 | **1.4591** |
| **NumberOfTime60\_89DaysPastDueNotW** | 0 | 0.24039 | 4.15518 | -0.2404 | **-0.058** |
| **Dependent** | 1 | 1 | 0 | 0 | **#DIV/0!** |

1.In cluster5 average age of customere is higher than popution average age approx 10 years high.why this age variation in this cluster,find out?

2.In cluster 5 NPA\_status mean is 0 this means in this cluster all custer are non-default but population mean is different then this,why???

3.In cluster 5 customers are only belong to Centre and east but population mean have different mean and customer participation from all the location. why this difference?

4.In this cluster people lives in only Rented house but population mean shows different structure of living .

5. In this cluster belong generally belongs to non\_official,officer2 and officer3,there is no self\_employed customer in this customer but in population 42.7% belongs to self\_employed. why this difference?

6.in this cluster people are highly educated as they only belong to PG (33.3%)and profession course (66.6%)but in population education layout is different .why?

CLUSTER6:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **PMean** | **Std Dev** | **DIFF** | **Z-score** |
| **NPA\_Status** | 0.06831 | 0.06684 | 0.24975 | 0.0014742 | **0.005902809** |
| **RevolvingUtilizationOfUnsecuredL** | 2.95901 | 6.04844 | 249.755 | -3.089425 | **-0.0123698** |
| **Gender1** | 0.61556 | 0.61536 | 0.48651 | 0.000201 | **0.000413145** |
| **age** | 52.1616 | 52.2952 | 14.7719 | -0.133603 | **-0.00904445** |
| **centr** | 0.29256 | 0.29304 | 0.45516 | -0.000482 | **-0.00105985** |
| **south** | 0.15687 | 0.15663 | 0.36346 | 0.0002319 | **0.000638042** |
| **east** | 0.13672 | 0.13701 | 0.34386 | -0.000291 | **-0.00084687** |
| **west** | 0.18656 | 0.18599 | 0.3891 | 0.0005702 | **0.001465423** |
| **Rented\_OwnHouse1** | 0.57297 | 0.57302 | 0.49464 | -5.11E-05 | **-0.00010331** |
| **non\_offi** | 0.27401 | 0.27409 | 0.44605 | -7.51E-05 | **-0.00016837** |
| **officer1** | 0.10082 | 0.10109 | 0.30145 | -0.000273 | **-0.00090462** |
| **officer2** | 0.08894 | 0.08889 | 0.28458 | 5.6E-05 | **0.000196781** |
| **officer3** | 0.10848 | 0.10849 | 0.311 | -1.04E-05 | **-3.344E-05** |
| **self\_emp** | 0.42774 | 0.42744 | 0.49471 | 0.0003022 | **0.000610865** |
| **edu\_phd** | 0.04216 | 0.04201 | 0.20061 | 0.0001573 | **0.000784128** |
| **edu\_pg** | 0.24801 | 0.24809 | 0.43191 | -8.69E-05 | **-0.0002012** |
| **edu\_gradu** | 0.26466 | 0.26503 | 0.44135 | -0.000375 | **-0.00084875** |
| **edu\_prof** | 0.33955 | 0.33947 | 0.47353 | 8.74E-05 | **0.000184571** |
| **edu\_matric** | 0.10562 | 0.1054 | 0.30707 | 0.0002168 | **0.00070603** |
| **NumberOfTime30\_59DaysPastDueNotW** | 0.43634 | 0.42103 | 4.19278 | 0.0153071 | **0.003650822** |
| **DebtRatio** | 75.0901 | 353.005 | 2037.82 | -277.915 | **-0.13637858** |
| **NumberOfOpenCreditLinesAndLoans** | 8.33048 | 8.45276 | 5.14595 | -0.122283 | **-0.02376297** |
| **NumberOfTimes90DaysLate** | 0.28408 | 0.26597 | 4.1693 | 0.0181094 | **0.004343507** |
| **NumberRealEstateLoansOrLines** | 0.95982 | 1.01824 | 1.12977 | -0.058416 | **-0.05170597** |
| **NumberOfTime60\_89DaysPastDueNotW** | 0.25674 | 0.24039 | 4.15518 | 0.0163558 | **0.003936244** |
| **Dependent** | 1 | 1 | 0 | 0 | **#DIV/0!** |

1In this cluster NPA\_status mean is equal to population mean and most of variable mean is equal to population mean.

2.Debt ratio is much lesser(277.915) than population mean. Region should be investigate.

Cluster7:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **PMean** | **Std Dev** | **DIFF** | **Z-score** |
| **NPA\_Status** | 0 | 0.06684 | 0.24975 | -0.0668 | **-0.268** |
| **RevolvingUtilizationOfUnsecuredL** | 0.2259 | 6.04844 | 249.755 | -5.8225 | **-0.023** |
| **Gender1** | 1 | 0.61536 | 0.48651 | 0.38464 | **0.7906** |
| **age** | 64.5 | 52.2952 | 14.7719 | 12.2048 | **0.8262** |
| **centr** | 0.25 | 0.29304 | 0.45516 | -0.043 | **-0.095** |
| **south** | 0.75 | 0.15663 | 0.36346 | 0.59337 | **1.6326** |
| **east** | 0 | 0.13701 | 0.34386 | -0.137 | **-0.398** |
| **west** | 0 | 0.18599 | 0.3891 | -0.186 | **-0.478** |
| **Rented\_OwnHouse1** | 0.75 | 0.57302 | 0.49464 | 0.17698 | **0.3578** |
| **non\_offi** | 0.25 | 0.27409 | 0.44605 | -0.0241 | **-0.054** |
| **officer1** | 0 | 0.10109 | 0.30145 | -0.1011 | **-0.335** |
| **officer2** | 0 | 0.08889 | 0.28458 | -0.0889 | **-0.312** |
| **officer3** | 0 | 0.10849 | 0.311 | -0.1085 | **-0.349** |
| **self\_emp** | 0.75 | 0.42744 | 0.49471 | 0.32256 | **0.652** |
| **edu\_phd** | 0.25 | 0.04201 | 0.20061 | 0.20799 | **1.0368** |
| **edu\_pg** | 0 | 0.24809 | 0.43191 | -0.2481 | **-0.574** |
| **edu\_gradu** | 0.25 | 0.26503 | 0.44135 | -0.015 | **-0.034** |
| **edu\_prof** | 0 | 0.33947 | 0.47353 | -0.3395 | **-0.717** |
| **edu\_matric** | 0.5 | 0.1054 | 0.30707 | 0.3946 | **1.2851** |
| **NumberOfTime30\_59DaysPastDueNotW** | 0 | 0.42103 | 4.19278 | -0.421 | **-0.1** |
| **DebtRatio** | 121998 | 353.005 | 2037.82 | 121645 | **59.694** |
| **NumberOfOpenCreditLinesAndLoans** | 9.75 | 8.45276 | 5.14595 | 1.29724 | **0.2521** |
| **NumberOfTimes90DaysLate** | 0 | 0.26597 | 4.1693 | -0.266 | **-0.064** |
| **NumberRealEstateLoansOrLines** | 1.75 | 1.01824 | 1.12977 | 0.73176 | **0.6477** |
| **NumberOfTime60\_89DaysPastDueNotW** | 0 | 0.24039 | 4.15518 | -0.2404 | **-0.058** |
| **Dependent** | 1 | 1 | 0 | 0 | **#DIV/0!** |

1.This cluster is outlier as it consists only 4 observation and all are non\_defaulter.

2.those average age is 64.5years that is 12.20 years higher than avg population age and among them 75% belong to South and 25% belong to centr but in population only 15% belong to south and 29% belong to centr. why this difference?