Assignment: Principal Component and Clustering

Purpose

Identify 6 significant clusters in the provided dataset for the marketing department that will transform how customers are viewed by the business

Data Exploration

In [1]:
```
proc print data=assign1.dev (obs=10);
run;
```

```
SAS Connection established. Subprocess id is 16212
```

Out[1]:

The SAS System

Obs	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	grade	annual_inc	dti	delinq_2yrs	inq_last_
1	10000	10000	10000	0.1349	230.05	C	32000	13.05	0	1
2	32000	32000	32000	0.1399	1093.53	C	110000	17.30	0	0
3	3000	3000	3000	0.1559	104.87	C	42000	8.34	0	0
4	9000	9000	9000	0.1399	307.56	C	125000	30.21	2	2
5	10000	10000	10000	0.1099	327.34	B	41000	16.48	2	2
6	8000	8000	8000	0.1449	275.33	C	75000	22.69	1	1
7	20000	20000	20000	0.1699	712.96	D	71000	18.56	0	1
8	4800	4800	4800	0.1349	162.87	C	111000	26.63	0	0
9	35000	35000	35000	0.2399	1372.97	E	180000	15.61	0	3
10	12000	12000	12000	0.1559	419.46	C	70000	22.86	0	1

In [2]:
```
ods exclude enginehost;
proc contents data=assign1.dev;
run;
```

Out[2]:

The SAS System

The CONTENTS Procedure

Data Set Name	ASSIGN1.DEV	**Observations**	27572
Member Type	DATA	**Variables**	77
Engine	V9	**Indexes**	0
Created	11/13/2016 00:58:05	**Observation Length**	611
Last Modified	11/13/2016 00:58:05	**Deleted Observations**	0
Protection		**Compressed**	CHAR
Data Set Type		**Reuse Space**	NO
Label		**Point to Observations**	YES
Data Representation	WINDOWS_64	**Sorted**	NO
Encoding	wlatin1 Western (Windows)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Informat
26	acc_now_delinq	Num	8	3.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Informat
44	acc_open_past_24mths	Num	8	4.
39	all_util	Num	8	6.
7	annual_inc	Num	8	8.
45	avg_cur_bal	Num	8	7.
46	bc_open_to_buy	Num	8	7.
47	bc_util	Num	8	6.
24	collection_recovery_fee	Num	8	5.
77	default	Num	8	
9	delinq_2yrs	Num	8	3.
8	dti	Num	8	7.
2	funded_amnt	Num	8	7.
3	funded_amnt_inv	Num	8	7.
6	grade	Char	3	$3.
35	il_util	Num	8	7.
41	inq_fi	Num	8	3.
43	inq_last_12m	Num	8	3.
10	inq_last_6mths	Num	8	3.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Informat
5	installment	Num	8	9.
4	int_rate	Num	8	PERCENT9.
25	last_pymnt_amnt	Num	8	9.
1	loan_amnt	Num	8	7.
38	max_bal_bc	Num	8	7.
48	mo_sin_old_il_acct	Num	8	5.
49	mo_sin_old_rev_tl_op	Num	8	5.
50	mo_sin_rcnt_rev_tl_op	Num	8	4.
51	mo_sin_rcnt_tl	Num	8	4.
52	mort_acc	Num	8	3.
33	mths_since_rcnt_il	Num	8	4.
53	mths_since_recent_bc	Num	8	4.
54	mths_since_recent_inq	Num	8	4.
55	num_accts_ever_120_pd	Num	8	3.
56	num_actv_bc_tl	Num	8	4.
57	num_actv_rev_tl	Num	8	4.
58	num_bc_sats	Num	8	4.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Informat
59	num_bc_tl	Num	8	4.
60	num_il_tl	Num	8	4.
61	num_op_rev_tl	Num	8	4.
62	num_rev_accts	Num	8	4.
63	num_rev_tl_bal_gt_0	Num	8	4.
64	num_sats	Num	8	4.
65	num_tl_120dpd_2m	Num	8	3.
66	num_tl_30dpd	Num	8	3.
67	num_tl_90g_dpd_24m	Num	8	3.
68	num_tl_op_past_12m	Num	8	4.
11	open_acc	Num	8	4.
29	open_acc_6m	Num	8	3.
31	open_il_12m	Num	8	3.
32	open_il_24m	Num	8	3.
30	open_il_6m	Num	8	3.
36	open_rv_12m	Num	8	3.
37	open_rv_24m	Num	8	4.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Informat
16	out_prncp	Num	8	10.
17	out_prncp_inv	Num	8	10.
69	pct_tl_nvr_dlq	Num	8	6.
70	percent_bc_gt_75	Num	8	6.
12	pub_rec	Num	8	3.
71	pub_rec_bankruptcies	Num	8	3.
23	recoveries	Num	8	5.
13	revol_bal	Num	8	8.
14	revol_util	Num	8	PERCENT7.
72	tax_liens	Num	8	3.
27	tot_coll_amt	Num	8	6.
28	tot_cur_bal	Num	8	8.
73	tot_hi_cred_lim	Num	8	8.
15	total_acc	Num	8	4.
74	total_bal_ex_mort	Num	8	8.
34	total_bal_il	Num	8	8.
75	total_bc_limit	Num	8	8.

Alphabetic List of Variables and Attributes				
#	**Variable**	**Type**	**Len**	**Informat**
42	total_cu_tl	Num	8	4.
76	total_il_high_credit_limit	Num	8	8.
18	total_pymnt	Num	8	9.
19	total_pymnt_inv	Num	8	9.
21	total_rec_int	Num	8	9.
22	total_rec_late_fee	Num	8	5.
20	total_rec_prncp	Num	8	9.
40	total_rev_hi_lim	Num	8	8.

In the give dataset, there there are 27572 observations and 77 variables. All the variables are numeric other than Grade which is char variable

In [3]:
```
proc means data=assign1.dev n nmiss min max mean std  kurt skew;
run;
```

Out[3]:

The SAS System

The MEANS Procedure

Variable	N	N Miss	Minimum	Maximum	Mean	Std Dev	Kurtosis	Skewness
loan_amnt	27572	0	1000.00	40000.00	14824.54	8932.60	-0.1505980	0.7610518
funded_amnt	27572	0	1000.00	40000.00	14824.18	8932.59	-0.1504663	0.7611626
funded_amnt_inv	27572	0	1000.00	40000.00	14817.77	8930.86	-0.1502582	0.7614111
int_rate	27572	0	0.0532000	0.3099000	0.1239266	0.0489008	0.5336314	0.8818232
installment	27572	0	30.1200000	1517.43	445.2166053	269.6470073	0.7788743	1.0512064
annual_inc	27572	0	0	9225000.00	77795.70	83435.72	5857.82	58.8436957
dti	27572	0	0	9999.00	21.1208291	147.5376534	4550.39	67.3198310
delinq_2yrs	27572	0	0	20.0000000	0.3639562	0.9622504	42.5226571	5.0962270
inq_last_6mths	27572	0	0	5.0000000	0.5172276	0.8205323	4.2441639	1.9021461
open_acc	27572	0	1.0000000	71.0000000	11.8524953	5.8043591	4.1026405	1.4348483
pub_rec	27572	0	0	31.0000000	0.2507979	0.6750724	208.5898663	8.0698593
revol_bal	27572	0	0	882984.00	16764.09	24063.26	161.2413165	8.9263876
revol_util	27560	12	0	1.5430000	0.5022455	0.2423468	-0.7720507	-0.0018362
total_acc	27572	0	2.0000000	144.0000000	24.3361381	11.9954876	2.0230860	1.0439330
out_prncp	27572	0	0	38956.11	13479.21	8283.75	-0.1312438	0.7708924
out_prncp_inv	27572	0	0	38956.11	13473.45	8282.25	-0.1308382	0.7711920
total_pymnt	27572	0	0	39927.86	1992.67	1451.31	46.4213916	3.6642903
total_pymnt_inv	27572	0	0	39927.86	1991.78	1450.80	46.3279816	3.6610915
total_rec_prncp	27572	0	0	39020.54	1344.97	1144.28	113.7054449	6.3414812
total_rec_int	27572	0	0	4917.62	647.5591343	556.1270199	4.9640383	1.9375203
total_rec_late_fee	27572	0	0	116.2400000	0.1394201	2.2473621	634.7824906	22.0275810
recoveries	27572	0	0	0	0	0	.	.
collection_recovery_fee	27572	0	0	0	0	0	.	.
last_pymnt_amnt	27572	0	0	36242.23	457.8396743	471.9813272	2007.32	32.8366128
acc_now_delinq	27572	0	0	4.0000000	0.0066734	0.0926700	386.5270173	17.1946384
tot_coll_amt	27572	0	0	63951.00	256.6784782	1599.30	365.7082376	15.7567392

Variable	N	N Miss	Minimum	Maximum	Mean	Std Dev	Kurtosis	Skewness
tot_cur_bal	27572	0	0	2326419.00	140937.58	157090.59	14.3482858	2.6174126
open_acc_6m	27572	0	0	17.0000000	0.9519440	1.1601218	5.8367713	1.7908735
open_il_6m	27572	0	0	39.0000000	2.8339257	3.1405978	13.4718207	2.9961226
open_il_12m	27572	0	0	25.0000000	0.7252648	0.9815660	17.7927339	2.2483186
open_il_24m	27572	0	0	51.0000000	1.5772523	1.6393971	35.2252828	2.6897433
mths_since_rcnt_il	26768	804	0	334.0000000	21.6819710	26.3684671	12.9178221	3.0681345
total_bal_il	27572	0	0	774639.00	34805.93	41504.60	23.0354446	3.4617417
il_util	23772	3800	0	229.8000000	70.9311249	23.1807866	0.8917972	-0.5527381
open_rv_12m	27572	0	0	20.0000000	1.3393660	1.5160997	8.0957453	2.0077395
open_rv_24m	27572	0	0	43.0000000	2.8521326	2.5932127	7.8035116	1.9050432
max_bal_bc	27572	0	0	243997.00	5675.86	5538.52	139.9324873	5.4652507
all_util	27572	0	0	169.5000000	60.0201763	20.1307022	0.0192310	-0.2128567
total_rev_hi_lim	27572	0	0	1084400.00	33867.85	35004.41	68.0991335	5.3465159
inq_fi	27572	0	0	19.0000000	0.9321413	1.4468199	10.9206125	2.6097232
total_cu_tl	27572	0	0	39.0000000	1.4795445	2.6773428	17.7024005	3.3534926
inq_last_12m	27572	0	0	31.0000000	2.0979980	2.3941917	9.5481506	2.3094775
acc_open_past_24mths	27572	0	0	56.0000000	4.6776803	3.2809658	6.5835866	1.5344431
avg_cur_bal	27572	0	0	275482.00	13303.87	15974.53	22.5596575	3.3774581
bc_open_to_buy	27255	317	0	264512.00	10507.19	15384.92	27.7821558	3.9585931
bc_util	27241	331	0	166.7000000	57.9976065	28.3109883	-0.9741334	-0.2956761
mo_sin_old_il_acct	26768	804	1.0000000	615.0000000	126.1800284	52.8488914	2.2905370	0.3921634
mo_sin_old_rev_tl_op	27572	0	4.0000000	745.0000000	182.4536124	98.0945964	1.3495909	1.0066305
mo_sin_rcnt_rev_tl_op	27572	0	0	306.0000000	13.7164152	16.9399681	17.7680442	3.3994773
mo_sin_rcnt_tl	27572	0	0	171.0000000	8.2038300	8.9735717	38.2502491	4.4710049
mort_acc	27572	0	0	37.0000000	1.5334035	1.8390719	7.5345117	1.6842322
mths_since_recent_bc	27265	307	0	472.0000000	24.4837337	31.8821122	16.8444432	3.2943346
mths_since_recent_inq	24643	2929	0	25.0000000	6.9945624	6.0000787	-0.0191166	0.8986013
num_accts_ever_120_pd	27572	0	0	34.0000000	0.5239736	1.3779345	60.2276404	5.6717114
num_actv_bc_tl	27572	0	0	35.0000000	3.6384013	2.3362234	6.4569613	1.6438948
num_actv_rev_tl	27572	0	0	47.0000000	5.6761932	3.4692024	6.6053081	1.7428374
num_bc_sats	27572	0	0	54.0000000	4.7413681	3.1253751	9.1164523	1.9969207
num_bc_tl	27572	0	0	70.0000000	7.5624184	4.6391361	5.0334993	1.5399766
num_il_tl	27572	0	0	128.0000000	8.5637603	7.5384063	9.6100515	2.2116162
num_op_rev_tl	27572	0	0	61.0000000	8.2953721	4.7672553	5.2646362	1.6160060

Variable	N	N Miss	Minimum	Maximum	Mean	Std Dev	Kurtosis	Skewness
num_rev_accts	27572	0	2.0000000	92.0000000	13.9892645	8.0645964	3.3109025	1.3839440
num_rev_tl_bal_gt_0	27572	0	0	47.0000000	5.5968736	3.3415590	6.1506038	1.6447164
num_sats	27572	0	1.0000000	71.0000000	11.8001959	5.7829087	4.1445307	1.4406144
num_tl_120dpd_2m	26229	1343	0	3.0000000	0.0012200	0.0409403	2055.46	41.0299584
num_tl_30dpd	27572	0	0	3.0000000	0.0040258	0.0698587	526.8827802	20.5737371
num_tl_90g_dpd_24m	27572	0	0	20.0000000	0.0918686	0.5268978	244.8349330	12.5587212
num_tl_op_past_12m	27572	0	0	25.0000000	2.1891412	1.9223494	5.3343944	1.5644498
pct_tl_nvr_dlq	27572	0	14.3000000	100.0000000	93.6786885	9.2747879	6.6800153	-2.2136206
percent_bc_gt_75	27255	317	0	100.0000000	42.1390754	36.0386589	-1.2316091	0.3236462
pub_rec_bankruptcies	27572	0	0	6.0000000	0.1386189	0.3962117	19.9400160	3.6067170
tax_liens	27572	0	0	30.0000000	0.0724648	0.4654551	782.2300255	18.5064086
tot_hi_cred_lim	27572	0	2700.00	3229815.00	175894.36	176799.73	16.2183633	2.6698643
total_bal_ex_mort	27572	0	0	882984.00	51563.70	48789.28	22.0891674	3.3260501
total_bc_limit	27572	0	0	368500.00	22055.27	22171.39	14.7544160	2.8028951
total_il_high_credit_limit	27572	0	0	870000.00	44293.22	44598.10	19.2915283	2.8986559
default	27572	0	0	1.0000000	0.0154142	0.1231955	59.9020295	7.8675081

There are 7601 missing values, which will be ignored in PCA and Clustering proceedure. A lot of variables are heavily skewed. Also, some variables have really strong kurtosis - which can mean that people with respect to those variables lie in a similar bracket or have similar behaviour. Another noticeable aspect is both recoveries and collection_recovery_fee are 0 values. Hennce, we can ignore them in our principal component proceedure

Principal Component prodeedure

PCA proceedure is used to reduce the dimension of the data and it gives us a direction that approximately how many variables can be used to explain the data. It does it by ignoring the effect of multi-collinearity. It also standardizes the data

In [15]:
```
ods select Eigenvalues ScreePlot PatternPlot  ;
proc princomp data=assign1.dev (drop=default recoveries collection_recovery_fee )
out=dev_pca_score outstat= dev_pca_stat
plots(only)=(scree
pattern(ncomp=3)
score(ellipse ncomp=3));
run;
```

Out[15]:

The SAS System

The PRINCOMP Procedure

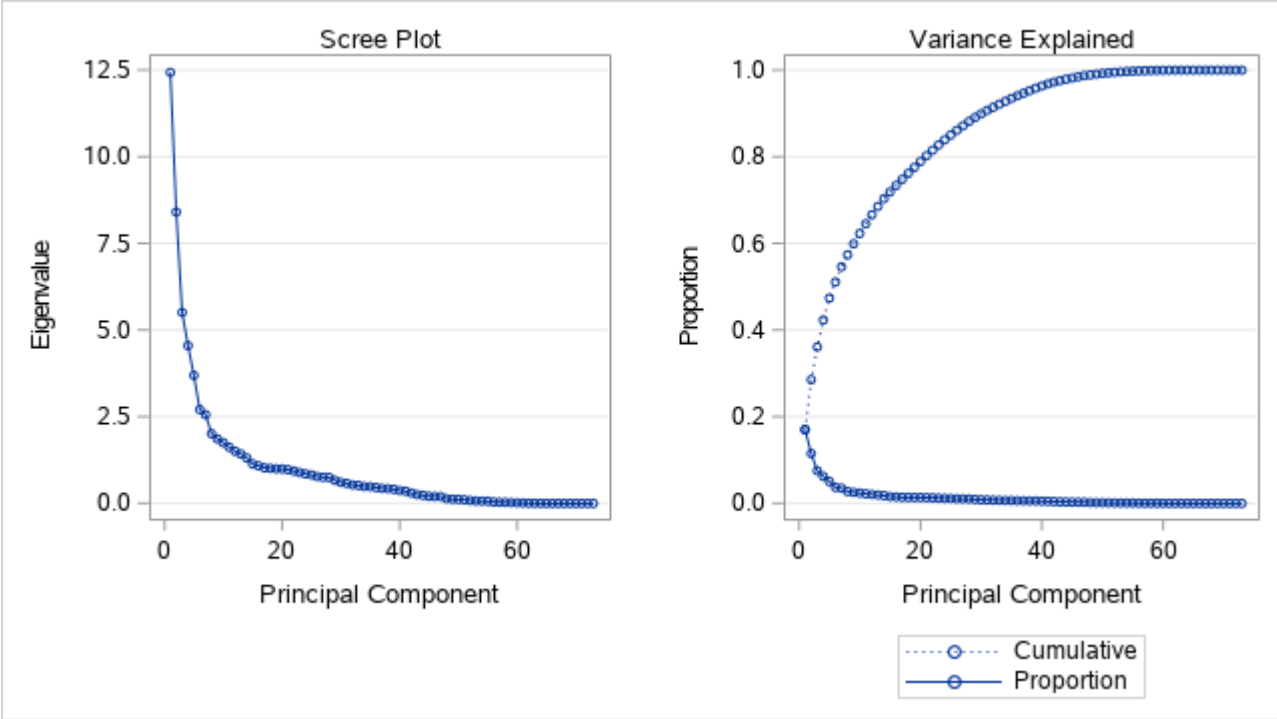
Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	12.4344182	4.0278404	0.1703	0.1703
2	8.4065779	2.8961841	0.1152	0.2855
3	5.5103938	0.9579837	0.0755	0.3610
4	4.5524101	0.8596258	0.0624	0.4233
5	3.6927843	0.9837031	0.0506	0.4739
6	2.7090812	0.1492128	0.0371	0.5110
7	2.5598684	0.5476458	0.0351	0.5461
8	2.0122227	0.1516531	0.0276	0.5737
9	1.8605696	0.1081305	0.0255	0.5992
10	1.7524390	0.1303310	0.0240	0.6232

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
11	1.6221081	0.1165728	0.0222	0.6454
12	1.5055353	0.0801479	0.0206	0.6660
13	1.4253874	0.1144061	0.0195	0.6855
14	1.3109813	0.1753390	0.0180	0.7035
15	1.1356424	0.0522649	0.0156	0.7190
16	1.0833774	0.0576276	0.0148	0.7339
17	1.0257498	0.0126710	0.0141	0.7479
18	1.0130789	0.0167629	0.0139	0.7618
19	0.9963159	0.0016321	0.0136	0.7755
20	0.9946838	0.0215743	0.0136	0.7891
21	0.9731095	0.0384508	0.0133	0.8024
22	0.9346588	0.0430496	0.0128	0.8152
23	0.8916091	0.0416947	0.0122	0.8274
24	0.8499144	0.0248477	0.0116	0.8391
25	0.8250667	0.0514294	0.0113	0.8504
26	0.7736373	0.0190560	0.0106	0.8610
27	0.7545813	0.0065225	0.0103	0.8713

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
28	0.7480588	0.0822264	0.0102	0.8816
29	0.6658324	0.0579462	0.0091	0.8907
30	0.6078861	0.0290676	0.0083	0.8990
31	0.5788186	0.0441510	0.0079	0.9069
32	0.5346676	0.0153881	0.0073	0.9143
33	0.5192795	0.0325830	0.0071	0.9214
34	0.4866965	0.0045491	0.0067	0.9280
35	0.4821474	0.0237520	0.0066	0.9347
36	0.4583954	0.0252264	0.0063	0.9409
37	0.4331690	0.0082228	0.0059	0.9469
38	0.4249462	0.0178975	0.0058	0.9527
39	0.4070486	0.0372757	0.0056	0.9583
40	0.3697730	0.0161752	0.0051	0.9633
41	0.3535977	0.0521856	0.0048	0.9682
42	0.3014122	0.0464641	0.0041	0.9723
43	0.2549481	0.0201432	0.0035	0.9758
44	0.2348048	0.0293416	0.0032	0.9790

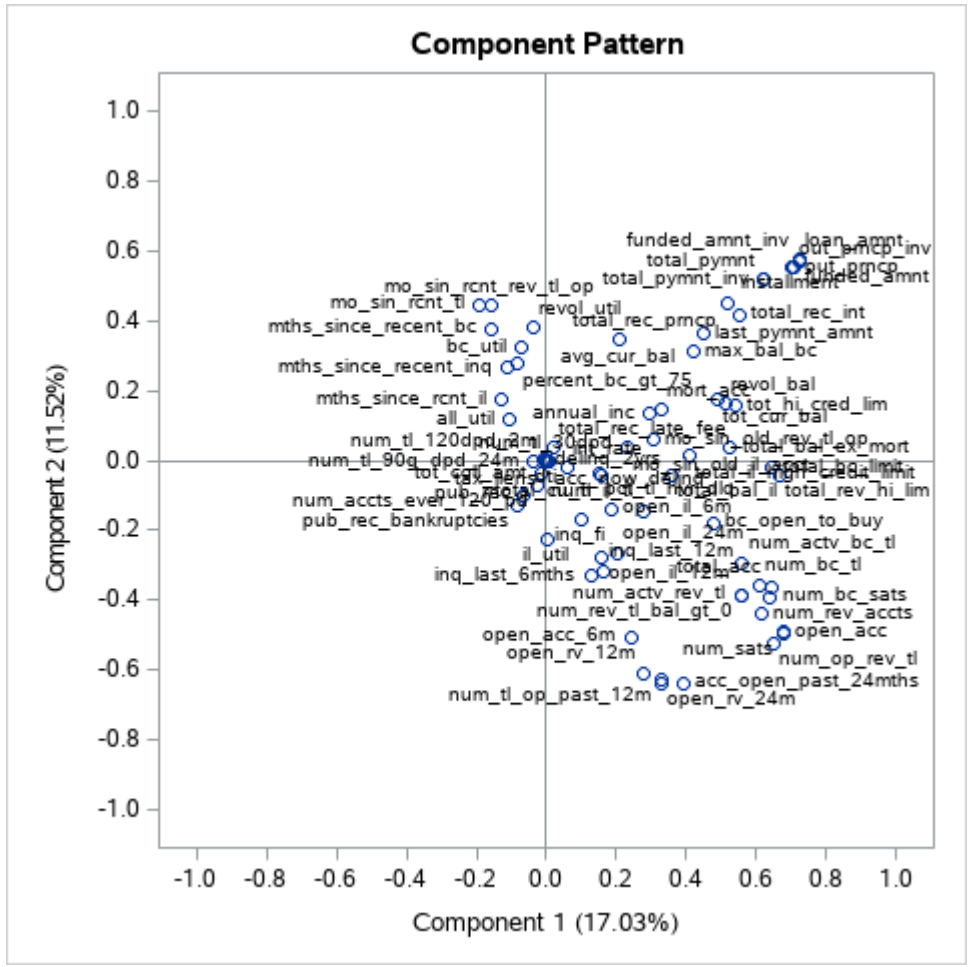
Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
45	0.2054632	0.0065014	0.0028	0.9818
46	0.1989618	0.0021606	0.0027	0.9845
47	0.1968012	0.0613276	0.0027	0.9872
48	0.1354736	0.0133602	0.0019	0.9891
49	0.1221134	0.0055569	0.0017	0.9908
50	0.1165565	0.0135460	0.0016	0.9924
51	0.1030105	0.0197051	0.0014	0.9938
52	0.0833054	0.0116726	0.0011	0.9949
53	0.0716328	0.0061392	0.0010	0.9959
54	0.0654935	0.0023798	0.0009	0.9968
55	0.0631137	0.0232341	0.0009	0.9977
56	0.0398796	0.0077088	0.0005	0.9982
57	0.0321709	0.0025226	0.0004	0.9987
58	0.0296483	0.0060279	0.0004	0.9991
59	0.0236204	0.0055726	0.0003	0.9994
60	0.0180478	0.0071015	0.0002	0.9996
61	0.0109463	0.0039346	0.0001	0.9998

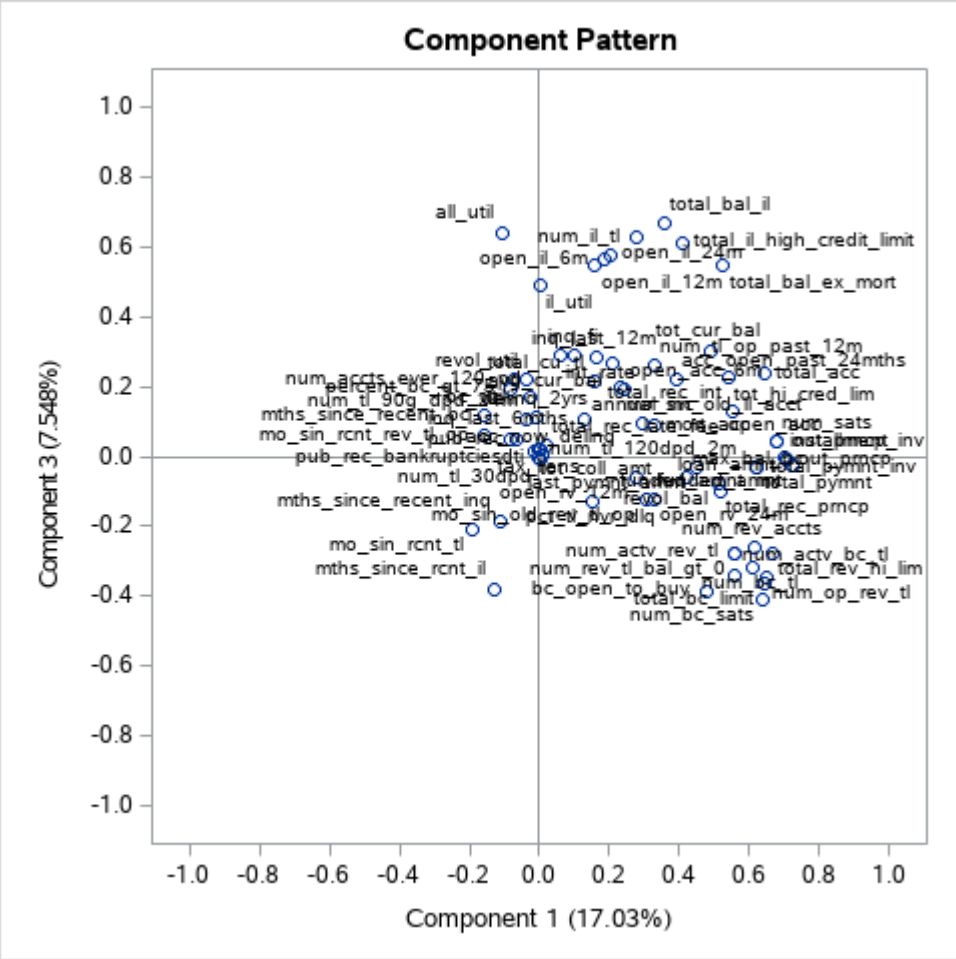
Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
62	0.0070118	0.0023024	0.0001	0.9999
63	0.0047094	0.0027150	0.0001	0.9999
64	0.0019944	0.0007158	0.0000	1.0000
65	0.0012786	0.0007389	0.0000	1.0000
66	0.0005397	0.0000802	0.0000	1.0000
67	0.0004595	0.0004012	0.0000	1.0000
68	0.0000583	0.0000346	0.0000	1.0000
69	0.0000237	0.0000226	0.0000	1.0000
70	0.0000011	0.0000011	0.0000	1.0000
71	0.0000000	0.0000000	0.0000	1.0000
72	0.0000000	0.0000000	0.0000	1.0000
73	0.0000000		0.0000	1.0000

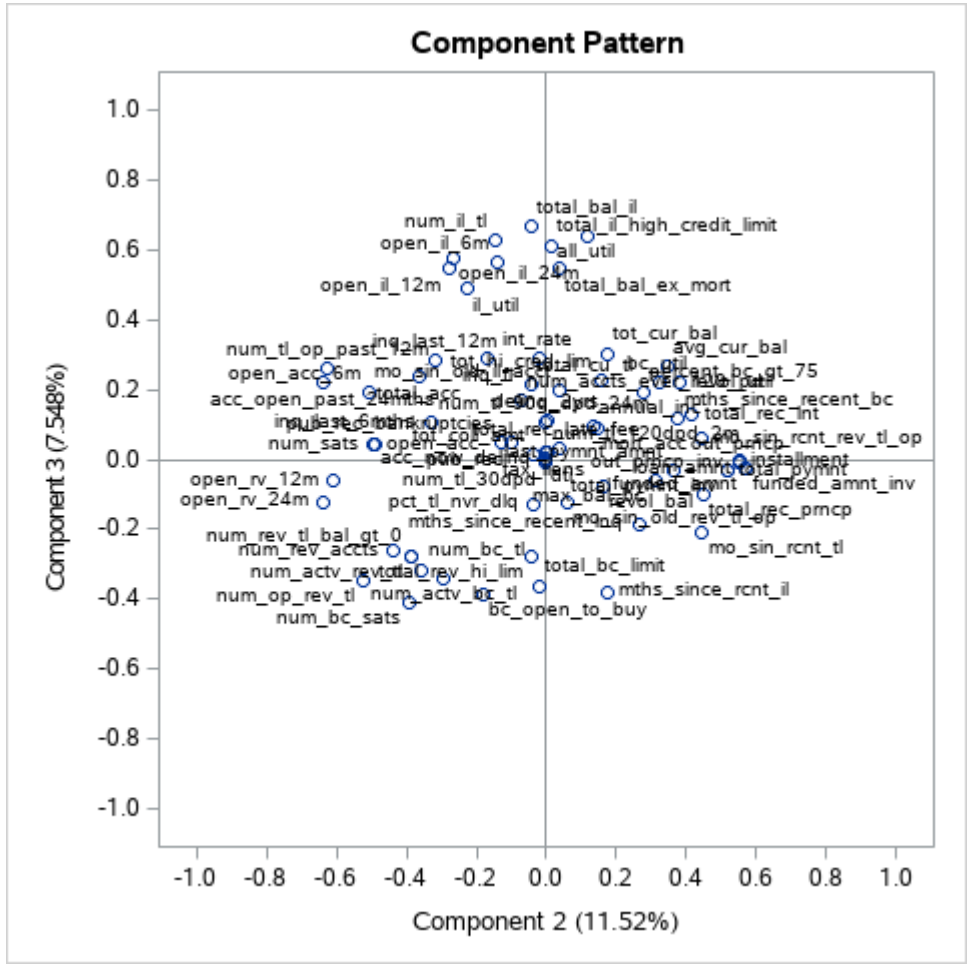


The SAS System

The PRINCOMP Procedure


Component Pattern
Component 2 (11.52%)
Component 1 (17.03%)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
funded_amnt_inv
loan_amnt
total_pymnt
out_prncp_inv
total_pymnt_inv
out_prncp
funded_amnt
installment
mo_sin_rcnt_rev_tl_op
mo_sin_rcnt_tl
revol_util
total_rec_int
mths_since_recent_bc
total_rec_prncp
last_pymnt_amnt
bc_util
avg_cur_bal
max_bal_bc
mths_since_recent_inq
percent_bc_gt_75
revol_bal
mths_since_rcnt_il
mort_acc
tot_hi_cred_lim
all_util
annual_inc
tot_cur_bal
total_rec_late_fee
num_tl_120dpd_2m
mo_sin_old_rev_tl_op
total_bal_ex_mort
num_tl_90g_dpd_24m
total_il_high_credit_limit
num_accts_ever_120_pd
open_il_6m
total_rev_hi_lim
pub_rec_bankruptcies
bc_open_to_buy
inq_fi
open_il_24m
num_actv_bc_tl
il_util
inq_last_12m
total_acc
num_bc_tl
inq_last_6mths
num_actv_rev_tl
num_bc_sats
num_rev_tl_bal_gt_0
num_rev_accts
open_acc_6m
open_acc
num_sats
open_rv_12m
num_op_rev_tl
acc_open_past_24mths
num_tl_op_past_12m
open_rv_24m


Component Pattern
Component 3 (7.548%)
Component 1 (17.03%)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
all_util
total_bal_il
num_il_tl
total_il_high_credit_limit
open_il_6m
open_il_24m
open_il_12m
total_bal_ex_mort
il_util
tot_cur_bal
num_tl_op_past_12m
acc_open_past_24mths
total_acc
total_rec_int
tot_hi_cred_lim
num_tl_120dpd_2m
total_pymnt
total_rec_prncp
num_rev_accts
num_actv_rev_tl
num_actv_bc_tl
num_rev_tl_bal_gt_0
total_rev_hi_lim
num_bc_tl
num_op_rev_tl
bc_open_to_buy
total_bc_limit
num_bc_sats
mo_sin_rcnt_tl
mths_since_rcnt_il
mths_since_recent_inq



From the above Eigenvalues of the Correlation Matrix it can be observed that First 18 principal components explain 76% of total variation, considering eigenvalue of 1 as the threshhold. The same aspect can be also observed in our scree plot - that the curve is almost flat after principal component 18 and eigen value 1, which means other principal components contributes very less to explain variation.

Hence, for now we can ignore principal components greater than 18

Clustering

Next step would be clustering, by infusing principal components in our clustering proceedure. Several iterations were tried with different cluster sizes. Reducing and incresing the number of principal components in the clustering proceedure was also tried. We would use Fastclus proceedure for clustering, which uses kmeans to find significant clusters in the set

In [6]:
```
ods select initialseeds mindist IterHistory ConvergenceStatus Criterion ClusterSum
 PseudoFStat ApproxExpOverAllRSq ;
proc fastclus data=dev_pca_score out = dev_clus outstat=dev_clus_stat
maxclusters=14 l=2 maxiter=86;
var prin1 prin2 prin3 prin4 prin5 prin6 prin7 prin8 prin9 prin10
prin11 prin12 prin13 prin14 prin15 prin16 prin17 prin18;
run;
```

Out[6]:

The SAS System

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=14 Maxiter=86 Converge=0.0001 Least=2

Initial Seeds

Cluster	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Pri
1	-3.80085666	-0.04568519	-3.14041625	-1.52604460	-0.37582391	-0.90370187	1.06433597	-1.68716337	-0.
2	1.49502516	-0.04052799	-0.96729997	2.72865886	-1.44884692	-4.17955538	11.72075250	-1.47559568	31
3	0.70648100	-1.36102664	0.38214222	-3.12370761	-1.16243370	-3.12931983	5.73903782	-0.43387160	-3.
4	26.13326092	7.75693709	-8.94053246	-13.17325911	-8.59502778	-5.33473254	-1.72293579	-1.39030878	-1.
5	10.86048449	-7.68296731	-9.45070158	6.88663384	-8.07936157	0.76657821	-3.30083317	-1.16913327	2.0
6	3.30701006	5.08070567	3.22226588	-6.13378791	-1.57741417	-1.57824580	-0.69466905	-0.14987723	1.3
7	-1.18127271	-0.34446961	5.80226952	1.87589617	-0.41793761	-19.86098508	29.88373777	-8.82153312	-5.
8	14.48234853	7.20265965	-0.15366021	15.09214115	1.28342897	11.19081481	-4.78068666	-4.02604953	1.9
9	12.92428240	9.71432450	13.14083652	16.62983674	-9.21302156	2.93932102	-1.29745801	0.41462012	-3.
10	18.71471474	-23.67030094	1.46693214	-6.21816992	3.86751273	0.43006978	1.57623731	6.40460517	-0.

Initial Seeds

Cluster	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Pri
11	11.41379228	-15.50311914	22.96452383	-8.77181166	-6.06173486	3.60689798	-5.14669777	1.98641628	7.4
12	10.05097271	-6.99515845	1.41824431	0.08584539	-4.62343237	-6.98059013	10.50318418	-0.53966350	31
13	-4.44035639	2.58319284	-0.02153754	-4.91690009	-0.77623536	-2.18181807	2.98331183	38.25278987	-0.
14	24.44045699	26.11123075	-5.26230107	-22.05817415	-6.83951133	-11.68076987	-1.14284910	-0.26316856	-4.

Minimum Distance Between Initial Seeds =	32.55035

Iteration History

		Relative Change in Cluster Seeds										
Iteration	**Criterion**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	5.0568	0.1734	0.8609	0.9513	0.9670	0.9187	0.9891	0.9827	0.6519	0.8739	0.8036	0.680
2	1.5544	0	0	0.2657	1.2546	0.1283	0.1174	0.8164	1.1856	0.7602	0.9657	0
3	1.4954	0	0	0.2298	0.5617	0.0745	0.0889	0.3821	0.7463	0.4925	0.4574	0
4	1.4545	0	0	0.2215	0.4430	0.0403	0.0720	0.2006	0.5396	0.3572	0.2344	0
5	1.4209	0	0	0.2050	0.1608	0.0391	0.0713	0.2838	0.4126	0.2664	0.1287	0
6	1.3964	0	0	0.1143	0.0508	0.0489	0.0620	0.2381	0.2890	0.1534	0.0846	0

Iteration History

Iteration	Criterion	Relative Change in Cluster Seeds										
		1	2	3	4	5	6	7	8	9	10	11
7	1.3859	0	0	0.0574	0.0541	0.0518	0.0399	0.1808	0.2573	0.0912	0.0608	0
8	1.3809	0	0	0.0396	0.0569	0.0464	0.0273	0.1217	0.1863	0.0582	0.0501	0
9	1.3779	0	0	0.0326	0.0275	0.0375	0.0221	0.0963	0.1604	0.0393	0.0419	0
10	1.3758	0	0	0.0276	0	0.0253	0.0194	0.0761	0.1135	0.0347	0.0276	0
11	1.3743	0	0	0.0276	0	0.0202	0.0171	0.0275	0.0935	0.0295	0.0201	0
12	1.3732	0	0	0.0258	0.0277	0.0176	0.0153	0.0459	0.0789	0.0275	0.0250	0
13	1.3724	0	0	0.0240	0	0.0171	0.0136	0.0343	0.0561	0.0237	0.0236	0
14	1.3718	0	0	0.0207	0	0.0135	0.0115	0.0408	0.0449	0.0204	0.0270	0
15	1.3713	0	0	0.0242	0	0.0152	0.00980	0.0460	0.0345	0.0171	0.0296	0
16	1.3709	0	0	0.0232	0	0.0157	0.00849	0.0343	0.0302	0.0144	0.0198	0
17	1.3706	0	0	0.0227	0	0.0197	0.00642	0.0253	0.0269	0.0115	0.0135	0
18	1.3703	0	0	0.0241	0	0.0192	0.00654	0.0245	0.0243	0.0108	0.0108	0
19	1.3700	0	0	0.0260	0	0.0221	0.00691	0.0344	0.0207	0.0101	0.00922	0
20	1.3696	0	0	0.0293	0	0.0234	0.00622	0.0197	0.0206	0.00892	0.0113	0
21	1.3693	0	0	0.0218	0	0.0224	0.00541	0.00882	0.0180	0.00855	0.00954	0
22	1.3690	0	0	0.0219	0.0362	0.0203	0.00549	0.00781	0.0155	0.00701	0.0123	0

Iteration History

Iteration	Criterion	Relative Change in Cluster Seeds										
		1	2	3	4	5	6	7	8	9	10	11
23	1.3688	0	0	0.0216	0	0.0219	0.00529	0.0123	0.0166	0.00631	0.0132	0
24	1.3686	0	0	0.0184	0	0.0209	0.00433	0.0125	0.0166	0.00453	0.0138	0
25	1.3684	0	0	0.0198	0	0.0190	0.00548	0.00552	0.0173	0.00563	0.0130	0
26	1.3682	0	0	0.0184	0	0.0192	0.00487	0.00700	0.0153	0.00725	0.0124	0
27	1.3680	0	0	0.0199	0	0.0157	0.00331	0.0116	0.0162	0.00539	0.00981	0
28	1.3678	0	0	0.0234	0	0.0191	0.00379	0.0106	0.0128	0.00540	0.00983	0
29	1.3676	0	0	0.0203	0	0.0191	0.00378	0.0149	0.0134	0.00523	0.0124	0
30	1.3674	0	0	0.0202	0	0.0156	0.00409	0.0117	0.0203	0.00404	0.0115	0
31	1.3672	0	0	0.0172	0	0.0127	0.00421	0.00930	0.0130	0.00415	0.00797	0
32	1.3671	0	0	0.0130	0	0.0120	0.00390	0.00763	0.00890	0.00294	0.00900	0
33	1.3670	0	0	0.0129	0	0.0117	0.00382	0.00433	0.0102	0.00291	0.00884	0
34	1.3670	0	0	0.0140	0	0.0113	0.00313	0.00382	0.00854	0.00196	0.00863	0
35	1.3669	0	0	0.0152	0	0.0120	0.00335	0.00432	0.00759	0.00314	0.00523	0
36	1.3668	0	0	0.0101	0	0.0122	0.00393	0.00301	0.00556	0.00189	0.00535	0
37	1.3667	0	0	0.0101	0	0.0126	0.00475	0.00314	0.00570	0.00235	0.00508	0
38	1.3667	0	0	0.00946	0	0.0117	0.00440	0	0.00506	0.00248	0.00798	0

Iteration History

Iteration	Criterion	Relative Change in Cluster Seeds										
		1	2	3	4	5	6	7	8	9	10	11
39	1.3666	0	0	0.00960	0	0.0118	0.00467	0.00314	0.00403	0.00327	0.00800	0
40	1.3665	0	0	0.00729	0	0.0129	0.00429	0.00324	0.00494	0.00444	0.00729	0
41	1.3665	0	0	0.00615	0	0.0124	0.00579	0.00275	0.00434	0.00358	0.00548	0
42	1.3664	0	0	0.00803	0	0.0148	0.00682	0.00369	0.00630	0.00359	0.0112	0
43	1.3663	0	0	0.0112	0	0.0138	0.00613	0	0.00590	0.00398	0.0118	0
44	1.3662	0	0	0.00966	0	0.0130	0.00779	0.00600	0.00836	0.00465	0.0123	0
45	1.3661	0	0	0.0104	0	0.0121	0.00589	0.00488	0.00871	0.00364	0.0135	0
46	1.3660	0	0	0.0112	0	0.0117	0.00494	0.00506	0.00762	0.00266	0.0144	0
47	1.3659	0	0	0.0105	0	0.0107	0.00503	0.00735	0.00696	0.00290	0.0116	0
48	1.3659	0	0	0.00888	0	0.0105	0.00425	0.00340	0.00475	0.00272	0.0100	0
49	1.3658	0	0	0.00882	0	0.00974	0.00448	0	0.00657	0.00271	0.0115	0
50	1.3657	0	0	0.00872	0	0.0109	0.00477	0.00845	0.00538	0.00237	0.0113	0
51	1.3657	0	0	0.00815	0	0.00982	0.00438	0.00364	0.00464	0.00200	0.0135	0
52	1.3656	0	0	0.00527	0	0.00841	0.00489	0.00285	0.00415	0.00202	0.00877	0
53	1.3656	0	0	0.00306	0	0.00897	0.00507	0	0.00301	0.00183	0.00954	0
54	1.3655	0	0	0.00397	0	0.00721	0.00382	0.00299	0.00428	0.00177	0.00874	0

Iteration History

Iteration	Criterion	Relative Change in Cluster Seeds										
		1	2	3	4	5	6	7	8	9	10	11
55	1.3655	0	0	0.00323	0	0.00621	0.00368	0	0.00413	0.00186	0.00671	0
56	1.3655	0	0	0.00260	0	0.00493	0.00310	0	0.00361	0.00160	0.00816	0
57	1.3655	0	0	0.00255	0	0.00416	0.00304	0.00301	0.00587	0.00163	0.00613	0
58	1.3655	0	0	0.00179	0	0.00359	0.00221	0.00345	0.00342	0.000914	0.00583	0
59	1.3654	0	0	0.000568	0	0.00287	0.00150	0	0.00544	0.00186	0.00475	0
60	1.3654	0	0	0.000799	0	0.00369	0.00189	0	0.00366	0.00162	0.00587	0
61	1.3654	0	0	0.00183	0	0.00396	0.00208	0.00380	0.00255	0.00115	0.00504	0
62	1.3654	0	0	0.00151	0	0.00364	0.00187	0	0.00616	0.00195	0.00464	0
63	1.3654	0	0	0.00114	0	0.00363	0.00152	0	0.00664	0.00187	0.00988	0
64	1.3654	0	0	0.000817	0	0.00462	0.00237	0	0.00556	0.00202	0.00779	0
65	1.3654	0	0	0.00252	0	0.00358	0.00151	0	0.00582	0.00250	0.00688	0
66	1.3654	0	0	0.00174	0	0.00336	0.00188	0	0.00574	0.00210	0.00583	0
67	1.3654	0	0	0.00353	0	0.00437	0.00223	0	0.00685	0.00136	0.00708	0
68	1.3653	0	0	0.00293	0	0.00419	0.00238	0	0.00967	0.00187	0.00680	0
69	1.3653	0	0	0.00158	0	0.00448	0.00188	0	0.00319	0.00152	0.00711	0
70	1.3653	0	0	0.00116	0	0.00361	0.00216	0	0.00258	0.00130	0.00429	0

Iteration History

Iteration	Criterion	Relative Change in Cluster Seeds										
		1	2	3	4	5	6	7	8	9	10	11
71	1.3653	0	0	0.000847	0	0.00291	0.00227	0	0.00351	0.00138	0.00306	0
72	1.3653	0	0	0.000705	0	0.00239	0.00195	0.00340	0.00263	0.00133	0.00169	0
73	1.3653	0	0	0.000666	0	0.00194	0.00129	0	0.000967	0.000834	0.00233	0
74	1.3653	0	0	0.000793	0	0.00153	0.000658	0	0	0	0.00244	0
75	1.3653	0	0	0	0	0.00112	0.000733	0	0	0.000406	0.00154	0
76	1.3653	0	0	0	0	0.00114	0.000861	0	0	0	0.000858	0
77	1.3653	0	0	0.000464	0	0.000719	0.000553	0	0	0	0.000596	0
78	1.3653	0	0	0	0	0.00114	0.000730	0	0	0.000528	0.000922	0
79	1.3653	0	0	0.000756	0	0.000921	0.000820	0	0	0.000367	0.00117	0
80	1.3653	0	0	0.000486	0	0.00136	0.00128	0	0.00120	0.000285	0.000762	0
81	1.3653	0	0	0.000822	0	0.00126	0.000864	0	0.000786	0.000265	0.000549	0
82	1.3653	0	0	0.000423	0	0.000548	0.000540	0	0	0.000286	0.000431	0
83	1.3653	0	0	0	0	0.000289	0.000274	0	0.000992	0.000334	0	0
84	1.3653	0	0	0	0	0.000232	0.000220	0	0.000807	0.000272	0	0
85	1.3653	0	0	0	0	0	0	0	0	0	0	0

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	1.3653

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	3	1.0844	4.1625		6	63.5064
2	72	2.0201	22.9147		5	19.9644
3	2129	1.5031	19.3230		5	5.3547
4	56	2.2617	23.0317		9	12.5713
5	4856	1.1539	25.0175		6	3.7067
6	5133	1.1377	22.4996		5	3.7067
7	369	1.8935	36.5104		5	7.7503
8	1231	1.8426	32.1394		9	6.1719
9	3654	1.3657	20.8775		6	5.8641
10	1937	1.5935	31.8992		5	6.1886
11	2	4.8441	14.5324		3	26.0381

Cluster Summary						
Cluster	**Frequency**	**RMS Std Deviation**	**Maximum Distance from Seed to Observation**	**Radius Exceeded**	**Nearest Cluster**	**Distance Between Cluster Centroids**
12	21	2.9627	25.0581		2	23.0315
13	505	1.5897	40.1296		5	6.0448
14	3	4.4749	19.3992		4	46.2990

Pseudo F Statistic =	1009.26

Approximate Expected Over-All R-Squared =	0.42780

Our convergence criteria was satisfied at 85th iteration and the cluster summary distinguishes 6 major clusters - (3, 5, 6, 8, 9, and 10). Also, it can be observed that cluster 5 is the closest cluster to all other clusters.

Scoring PCA

In [7]:
```
proc score data=assign1.new score=dev_pca_stat out= new_pca_score_from_devstat;
run;
```

Out[7]:

```
94   ods listing close;ods html5 (id=saspy_internal) file=stdout options(bitmap_mode='inline') device=sv
g; ods graphics on /
94 ! outputfmt=png;
NOTE: Writing HTML5(SASPY_INTERNAL) Body file: STDOUT
95
96   proc score data=assign1.new score=dev_pca_stat out= new_pca_score_from_devstat;
NOTE: Data file ASSIGN1.NEW.DATA is in a format that is native to another host, or the file encoding does
 not match the session
      encoding. Cross Environment Data Access will be used, which might require additional CPU resources
 and might reduce
      performance.
97   run;
NOTE: No VAR statement is given.  All numeric variables in the SCORE= data set will be used to compute th
e scores.
NOTE: There were 15828 observations read from the data set ASSIGN1.NEW.
NOTE: There were 150 observations read from the data set WORK.DEV_PCA_STAT.
NOTE: The data set WORK.NEW_PCA_SCORE_FROM_DEVSTAT has 15828 observations and 149 variables.
NOTE: PROCEDURE SCORE used (Total process time):
      real time           0.62 seconds
      cpu time            0.41 seconds

98
99
100  ods html5 (id=saspy_internal) close;ods listing;

101
```

Now scoring the New Dataset for clustering

In [8]:
```
proc fastclus instat= dev_clus_stat data= new_pca_score_from_devstat out=new_clus;
var prin1 prin2 prin3 prin4 prin5 prin6 prin7 prin8 prin9 prin10
prin11 prin12 prin13 prin14 prin15 prin16 prin17 prin18;
run;
```

Out[8]:

```
103  ods listing close;ods html5 (id=saspy_internal) file=stdout options(bitmap_mode='inline') device=sv
g; ods graphics on /
103! outputfmt=png;
NOTE: Writing HTML5(SASPY_INTERNAL) Body file: STDOUT
104
105  proc fastclus instat= dev_clus_stat data= new_pca_score_from_devstat out=new_clus;
106  var prin1 prin2 prin3 prin4 prin5 prin6 prin7 prin8 prin9 prin10
107  prin11 prin12 prin13 prin14 prin15 prin16 prin17 prin18;
108  run;
NOTE: 14 clusters in the INSTAT=WORK.DEV_CLUS_STAT.DATA data set.
NOTE: There were 80 observations read from the data set WORK.DEV_CLUS_STAT.
NOTE: The data set WORK.NEW_CLUS has 15828 observations and 151 variables.
NOTE: PROCEDURE FASTCLUS used (Total process time):
      real time           0.04 seconds
      cpu time            0.03 seconds

109
110  ods html5 (id=saspy_internal) close;ods listing;

111
```

Clustering adds two new columns to the produced clustering output.

1. Cluster - That defines the clusters
2. Distance - That defines the distance

In [9]:
```
proc freq data=new_clus;
tables cluster;
run;
```

Out[9]:

The SAS System

The FREQ Procedure

Cluster				
CLUSTER	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
2	52	0.43	52	0.43
3	1470	12.14	1522	12.57
4	34	0.28	1556	12.86
5	3092	25.55	4648	38.40
6	2629	21.72	7277	60.12
7	239	1.97	7516	62.10
8	554	4.58	8070	66.67
9	2483	20.51	10553	87.19
10	1193	9.86	11746	97.04
11	2	0.02	11748	97.06
12	13	0.11	11761	97.17

Frequency Missing = 3724

Cluster				
CLUSTER	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
13	339	2.80	12100	99.97
14	4	0.03	12104	100.00
Frequency Missing = 3724				

From the above table it can be observed that for the new dataset, the significant clusters are still the same - (3, 5, 6, 8, 9, 10). They obviously can not be the same frequency but their frequency is distributed in similar way

Merging the clustering output to the original dataset

In [14]:
```
data clus_profile;
merge assign1.dev dev_clus (keep=cluster);
run;
```

Out[14]:

```
150  ods listing close;ods html5 (id=saspy_internal) file=stdout options(bitmap_mode='inline') device=sv
g; ods graphics on /
150! outputfmt=png;
NOTE: Writing HTML5(SASPY_INTERNAL) Body file: STDOUT
151
152  data clus_profile;
NOTE: Data file ASSIGN1.DEV.DATA is in a format that is native to another host, or the file encoding does
 not match the session
      encoding. Cross Environment Data Access will be used, which might require additional CPU resources
 and might reduce
      performance.
153  merge assign1.dev dev_clus (keep=cluster);
154  run;
NOTE: There were 27572 observations read from the data set ASSIGN1.DEV.
NOTE: There were 27572 observations read from the data set WORK.DEV_CLUS.
NOTE: The data set WORK.CLUS_PROFILE has 27572 observations and 78 variables.
NOTE: DATA statement used (Total process time):
      real time           1.21 seconds
      cpu time            0.34 seconds

155
156  ods html5 (id=saspy_internal) close;ods listing;

157
```

Insights and Comments

Profiling Cluster

Variables were selected using some business sense and by looking at the correlation table between Principal components and original variables.

In [17]:
```
proc tabulate data=clus_profile;
class cluster grade;
var revol_bal loan_amnt all_util annual_inc  num_sats total_il_high_credit_limit
total_acc int_rate percent_Bc_Gt_75 inq_fi num_il_tl delinq_2Yrs pub_rec_bankruptcies tot_cur_bal;
table cluster,loan_amnt*mean revol_bal*mean all_util*mean all annual_inc*mean
num_sats*mean total_il_high_credit_limit*mean total_acc*mean int_rate*mean
percent_Bc_Gt_75*mean inq_fi*mean num_il_tl*mean delinq_2Yrs*mean pub_rec_bankruptcies*mean tot_cur_bal*mea
n;
run;
```

Out[17]:

The SAS System

	loan_amnt	revol_bal	all_util	All	annual_inc	num_sats	total_il_high_credit_limit	total_acc	int_rate	percer
	Mean	Mean	Mean	N	Mean	Mean	Mean	Mean	Mean	Mean
Cluster 1	12800.00	7106.00	41.37	3	0.00	6.67	23866.00	16.00	0.10	70.00
2	13954.17	16532.25	56.33	72	80773.00	13.10	54414.24	29.89	0.13	32.80
3	13159.61	14022.69	78.08	2129	87579.82	17.05	111273.56	37.17	0.13	45.05
4	18922.32	18305.63	54.56	56	93590.98	12.41	58393.29	26.77	0.12	25.26
5	9485.52	8420.21	56.51	4856	60610.93	11.15	33041.62	22.38	0.13	25.42
6	9928.74	10662.23	64.97	5133	59037.99	8.44	33560.35	17.21	0.12	56.51
7	10783.06	7363.39	69.38	369	75317.28	11.68	52191.47	29.16	0.13	36.71
8	28047.12	54019.92	55.96	1231	160959.79	17.93	88869.76	36.68	0.11	35.69
9	25811.07	19134.58	62.48	3654	95808.60	11.19	50341.41	24.32	0.14	48.02
10	14408.25	20877.88	51.81	1937	81642.78	21.06	47468.51	38.56	0.13	28.79

	loan_amnt	revol_bal	all_util	All	annual_inc	num_sats	total_il_high_credit_limit	total_acc	int_rate	percer
	Mean	Mean	Mean	N	Mean	Mean	Mean	Mean	Mean	Mean
11	11300.00	10986.00	78.85	2	85397.50	20.50	161291.50	126.50	0.15	50.00
12	13525.00	9545.14	58.21	21	89314.71	10.57	70071.57	26.19	0.13	25.23
13	12075.84	8705.98	61.71	505	94069.33	10.18	39433.48	22.01	0.13	35.11
14	36666.67	9217.00	31.37	3	102666.67	20.00	25513.00	43.33	0.10	5.57

In the above table we have 6 main clusters that would help us in segmentation

Cluster 3 are the people who earn around 80K but also tend to use a credit oftenly, although they have are in good status with the bank and they like to maintain good balance in their accounts - hence bank should focus on them

In cluster 5, it can be noticed that these customers borrow less and they earn around 60K, has lesser revolving balance on their cards and accounts, relatively less card utilization than most and less delinquency in 2years states that these people use banking and they don't abuse the banking system. Also due to less income they are not able to maintain big balance in their account

Cluster 6, has the least income and is the most prominent cluster. This cluster includes one of our most common customer, who uses credit frequently, but doesn't borrow big amounts. They try to be in good standing with the financial institutions and they don't make a lot of financial inquiries. Also, due to less income they are not able to maintain big balance in their account

Cluster 8, is our smallest and the most important cluster. They have the highest annual income, have the most bank accounts and they normally loan bigger amounts . Also that is one reason why these customers has the least interest rate.

Cluster 9 have significant annual income, but they tend to use a lot of credit and loans. They are profitable customer for bank as they pay slightly higher interest rate

Cluster 10 normally maintain high revolving balance with respect to their income. Therefore banks should be slightly careful while providing them with loans

Cluster	**Analysis**	
Segment 3	Medium-high Income, high current balance	Customer who can be induced to invest their current a/c funds
Segment 5	Medium income, less loans, less deliquency	Customers who can be persuaded to get more loans
Segment 6	Medium-Low income, less current a/c balance, less financial inquiries	Provide them with more smaller credits and loans as they are less risky
Segment 8	Highest income, wealthy segment	Try to bring loyalty amongst these customers and get them to invest more
segment 9	Medium high income, but a lot of credit usage	Lower interest rates to induce more credit
Segment 10	Medium high income, high revolving balance	some risk of defaultin