Vivian Xia

Assignment: KMeans

Standardize the Data

	trn_TRUNC_IMP_MORTDUE	trn_TRUNC_IMP_VALUE	trn_TRUNC_IMP_YOJ	\
0	-1.187821	-1.218785	0.243923	
1	-0.055078	-0.636046	-0.241631	
2	-1.504630	-1.661666	-0.657820	
3	-0.184108	-0.222713	-0.241631	
4	0.656126	0.228887	-0.796550	

	trn_TRUNC_IMP_DEROG	trn_TRUNC_IMP_DELINQ	trn_TRUNC_IMP_CLAGE	\
0	-0.329584	-0.418963	-1.067294	
1	-0.329584	1.837718	-0.718939	
2	-0.329584	-0.418963	-0.368469	
3	-0.329584	-0.418963	-0.064081	
4	-0.329584	-0.418963	-1.080400	

	trn_TRUNC_IMP_NINQ	trn_TRUNC_IMP_CLNO	trn_TRUNC_IMP_DEBTINC
0	-0.081701	-1.247113	0.141543
1	-0.791521	-0.736029	0.141543
2	-0.081701	-1.144896	0.141543
3	-0.081701	-0.122729	0.141543
4	-0.791521	-0.736029	0.141543

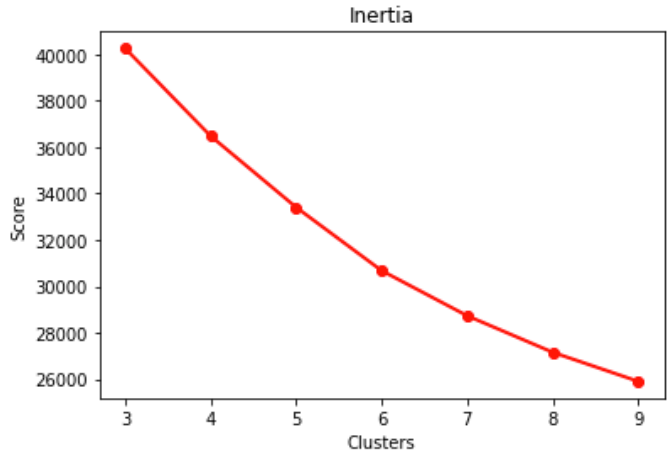
	trn_TRUNC_IMP_MORTDUE	trn_TRUNC_IMP_VALUE	trn_TRUNC_IMP_YOJ	\
count	5.960000e+03	5960.000000	5.960000e+03	
mean	-3.814995e-17	0.000000	1.192186e-16	
std	1.000084e+00	1.000084	1.000084e+00	
min	-1.797780e+00	-1.834256	-1.212739e+00	
25%	-6.167719e-01	-0.673946	-7.965498e-01	
50%	-1.841083e-01	-0.222713	-2.416307e-01	
75%	4.100674e-01	0.367847	4.520181e-01	
max	3.292580e+00	3.405910	3.087884e+00	

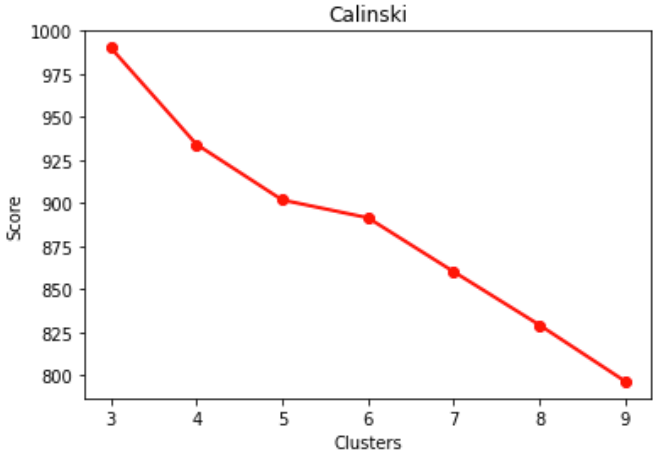
	trn_TRUNC_IMP_DEROG	trn_TRUNC_IMP_DELINQ	trn_TRUNC_IMP_CLAGE	\
count	5.960000e+03	5.960000e+03	5.960000e+03	
mean	7.749208e-18	-2.861246e-17	-2.956621e-16	
std	1.000084e+00	1.000084e+00	1.000084e+00	
min	-3.295844e-01	-4.189627e-01	-2.264132e+00	
25%	-3.295844e-01	-4.189627e-01	-7.755283e-01	
50%	-3.295844e-01	-4.189627e-01	-6.408111e-02	
75%	-3.295844e-01	-4.189627e-01	6.166882e-01	
max	4.817114e+00	4.094399e+00	3.189491e+00	

	trn_TRUNC_IMP_NINQ	trn_TRUNC_IMP_CLNO	trn_TRUNC_IMP_DEBTINC
count	5.960000e+03	5.960000e+03	5.960000e+03
mean	-5.484055e-17	-1.907497e-17	-3.051996e-16
std	1.000084e+00	1.000084e+00	1.000084e+00
min	-7.915210e-01	-2.167063e+00	-5.016046e+00
25%	-7.915210e-01	-6.338123e-01	-4.683220e-01
50%	-8.170078e-02	-1.227287e-01	1.415426e-01
75%	6.281194e-01	4.905717e-01	6.125221e-01
max	3.467400e+00	3.045990e+00	3.477551e+00

The data has the numerical variables' missing values imputed with the median values. The data was also truncated to remove outliers and then standardized. The values are between -5 and 5.

Score Plots





The inertia plot does not really have any point where there is clear elbow or flattening in its graph. The silhouette plot decreases at 4 and then increases at 5, so it may be a good idea to stick with 4 clusters. The calinksi plot flattens at 5 clusters. I am going to go with 4 clusters.

KMeans Clusters

```
K =  4
=======
   TARGET_BAD_FLAG  TARGET_LOSS_AMT  LOAN  MORTDUE     VALUE   REASON     JOB  \
0                1            641.0  1100  25860.0   39025.0  HomeImp   Other
1                1           1109.0  1300  70053.0   68400.0  HomeImp   Other
2                1            767.0  1500  13500.0   16700.0  HomeImp   Other
3                1           1425.0  1500      NaN       NaN      NaN     NaN
4                0              NaN  1700  97800.0  112000.0  HomeImp  Office

    YOJ  DEROG  DELINQ       CLAGE  NINQ  CLNO  DEBTINC  CLUSTER
0  10.5    0.0     0.0   94.366667   1.0   9.0      NaN        1
1   7.0    0.0     2.0  121.833333   0.0  14.0      NaN        1
2   4.0    0.0     0.0  149.466667   1.0  10.0      NaN        1
3   NaN    NaN     NaN         NaN   NaN   NaN      NaN        1
4   3.0    0.0     0.0   93.333333   0.0  14.0      NaN        1
         TARGET_BAD_FLAG  TARGET_LOSS_AMT          LOAN        MORTDUE  \
CLUSTER
0               0.621749     15627.007605  18216.784870    63935.799544
1               0.193644     10563.637902  16339.351245    57468.788187
2               0.128520     11495.073034  18850.830325    55851.598734
3               0.142727     22616.535032  24747.000000   137851.205894

                 VALUE        YOJ     DEROG    DELINQ       CLAGE      NINQ  \
CLUSTER
0         90237.166240   6.958808  2.436019  1.526961  152.607654  2.476427
1         79922.424457   5.329652  0.060971  0.257088  139.433471  1.203745
2         86345.684562  18.341423  0.059006  0.592730  243.141051  0.775281
3        184883.848084   7.714465  0.078594  0.320513  213.839363  1.167946

              CLNO    DEBTINC
CLUSTER
0        23.669031  34.835882
1        17.899790  33.278767
2        22.280000  32.510588
3        28.051565  36.454990
```

```
CLUSTER  TARGET_BAD_FLAG
0        1                263
         0                160
1        0               2461
         1                591
2        0               1207
         1                178
3        0                943
         1                157
Name: TARGET_BAD_FLAG, dtype: int64
```

If in cluster 0, the person will more likely have their loan default. If in cluster 1, 2, and 3, the loan will probably not default. There is some overlap for all the clusters.

Cluster 0 has a larger average DEROG, DELINQ, and NINQ value than that of the other clusters. If the person has more derogatory and delinquencies on their credit report as well as number of inquiries looking for credit, it is more likely this person's loan will default.

Cluster 1 has a smaller average LOAN, VALUE, YOJ, DELINQ, CLAGE, CLNO value than the other clusters. A person that has a smaller loan amount, value of their house, years on job, delinquencies on their report, credit line age, and number of credit lines will more likely have their loan not default. This cluster seems to include younger people who are more likely just starting out with a job and have a newer credit line with less years on the job and fewer credit lines.

Cluster 2 has a large average YOJ and CLAGE value compared to the others. It has a smaller average NINQ value than the others as well. A person that has been at their job longer and has had their credit line for longer will more likely not have their loan default. Also, if the person has a lower number of inquiries looking for credit, there is a more likely chance their loan will not default.

Cluster 3 has significantly larger average values in TARGET_LOSS_AMT, LOAN, MORTDUE, VALUE than any of the other clusters. So those people with higher loan amounts, current outstanding mortgage balance, and value of their house will most likely not have their loan default.

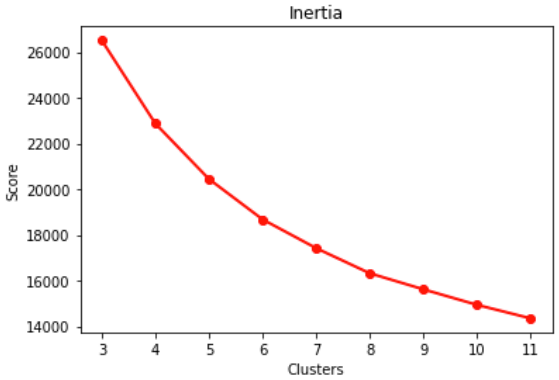
Bingo Bonus

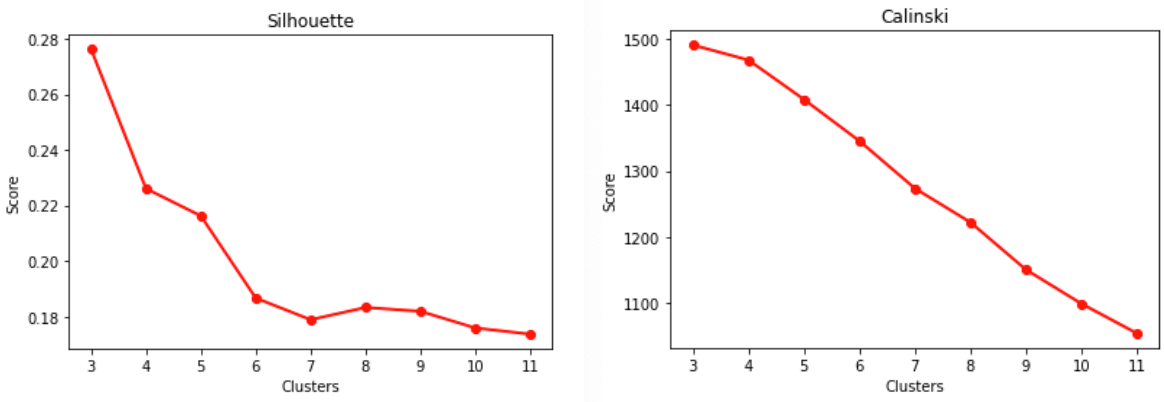
Principal Components

	PC_1	PC_2	PC_3	PC_4	PC_5
0	-2.213738	-0.064566	-0.523838	-0.573712	0.032002
1	-0.796038	0.403654	0.303282	0.735630	-1.345805
2	-2.464173	0.034273	-0.504138	-0.572960	-0.541261
3	-0.306641	-0.228400	-0.397817	-0.321428	-0.171620
4	-0.092079	-0.314853	-1.601887	0.539274	-0.531030

I am going to use the first 5 principal components that contain 74% of the total information as found in Assignment: PCA.

Score Plots





The inertia plot does not show any clear flattening along its graph. The silhouette plot has flattening at 4 before dropping steeply down after 5. The calinksi plot has a bit of flattening at 3 to 4. I will go with 4 clusters.

KMeans Clusters

```
K =  4
=======
   TARGET_BAD_FLAG  TARGET_LOSS_AMT  LOAN  MORTDUE     VALUE   REASON     JOB  \
0                1            641.0  1100  25860.0   39025.0  HomeImp   Other
1                1           1109.0  1300  70053.0   68400.0  HomeImp   Other
2                1            767.0  1500  13500.0   16700.0  HomeImp   Other
3                1           1425.0  1500      NaN       NaN      NaN     NaN
4                0              NaN  1700  97800.0  112000.0  HomeImp  Office

     YOJ  DEROG  DELINQ       CLAGE  NINQ  CLNO  DEBTINC  CLUSTER
0   10.5    0.0     0.0   94.366667   1.0   9.0      NaN        0
1    7.0    0.0     2.0  121.833333   0.0  14.0      NaN        0
2    4.0    0.0     0.0  149.466667   1.0  10.0      NaN        0
3    NaN    NaN     NaN         NaN   NaN   NaN      NaN        0
4    3.0    0.0     0.0   93.333333   0.0  14.0      NaN        0
         TARGET_BAD_FLAG  TARGET_LOSS_AMT          LOAN        MORTDUE  \
CLUSTER
0               0.187369     10049.392924  16336.252617   57990.433496
1               0.134964     21744.120805  24640.036232  137699.466055
2               0.612013     16210.143236  17486.850649   63231.598017
3               0.091703      9542.142857  19002.401747   53982.755000

                 VALUE        YOJ     DEROG    DELINQ       CLAGE      NINQ  \
CLUSTER
0         80035.813487   5.333959  0.070632  0.158717  136.347029  1.156162
1        184415.767241   7.932401  0.086777  0.268891  213.853840  1.174785
2         88703.069686   7.566318  1.772964  2.262458  156.984195  2.228426
3         85309.414132  17.775699  0.045879  0.298417  245.817093  0.785011

              CLNO    DEBTINC
CLUSTER
0        17.328213  33.476926
1        28.106422  36.206824
2        24.907468  35.423826
3        21.987528  32.016252
```

```
CLUSTER  TARGET_BAD_FLAG
0        0                2329
         1                 537
1        0                 955
         1                 149
2        1                 377
         0                 239
3        0                1248
         1                 126
Name: TARGET_BAD_FLAG, dtype: int64
```

The PCA doesn't seem to do have made a big difference compared to using all the variables. Cluster 3 is purer than the corresponding original cluster. Otherwise, the other clusters seem to be more pure or just as pure as their corresponding original cluster. There is still overlap for all the clusters. Similar to the original cluster, there are 3 clusters (Cluster 0, Cluster 1, and Cluster 3) whose members are more likely to not have their loan default and 1 cluster (Cluster 2) whose members are more likely to have their loan default. This PC version of the cluster that has member more likely to have their loan default is less pure than the original version. I would probably use the original version or try to add more principal components and rerun.