Vivian Xia

Assignment: Principal Component Analysis

Standardize the Data

	std_TRUNC_IMP_MORTDUE	std_TRUNC_IMP_VALUE	std_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	

	std_TRUNC_IMP_DEROG	std_TRUNC_IMP_DELINQ	std_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	

	std_TRUNC_IMP_NINQ	std_TRUNC_IMP_CLNO	std_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

	std_TRUNC_IMP_MORTDUE	std_TRUNC_IMP_VALUE	std_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	

	std_TRUNC_IMP_DEROG	std_TRUNC_IMP_DELINQ	std_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	

	std_TRUNC_IMP_NINQ	std_TRUNC_IMP_CLNO	std_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 missing values in the numerical data were imputed. The data was also truncated to remove outliers and then standardized. The values are all within -5 and 5.

Principal Component Analysis

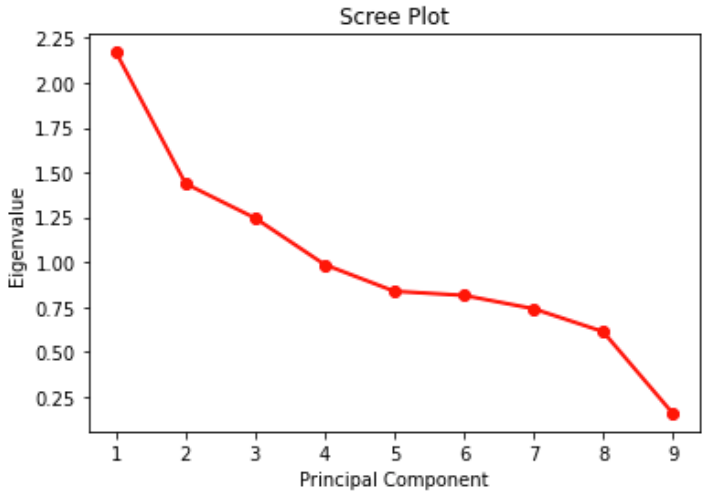
```
Eigen Values
[2.16999301 1.43739568 1.24532789 0.98605536 0.83760722 0.81452937
 0.74040446 0.61242042 0.15777691]
```

```
2.17 variation= 24 %  total= 24 %
1.44 variation= 15 %  total= 40 %
1.25 variation= 13 %  total= 53 %
0.99 variation= 10 %  total= 64 %
0.84 variation= 9 %  total= 74 %
0.81 variation= 9 %  total= 83 %
0.74 variation= 8 %  total= 91 %
0.61 variation= 6 %  total= 98 %
0.16 variation= 1 %  total= 100 %
```

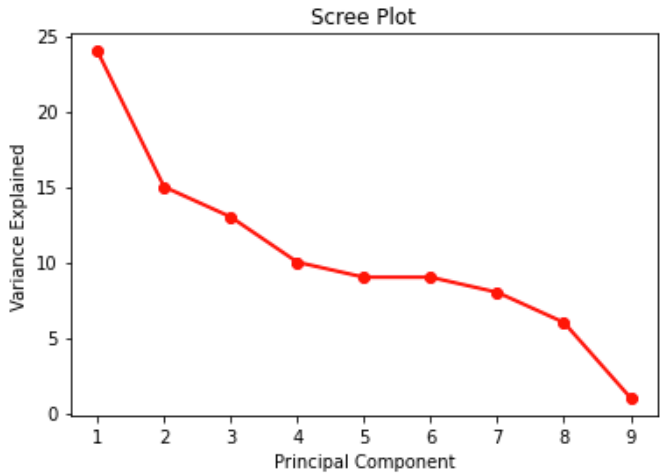
PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9 has as much information as 2.17, 1.44, 1.25, 0.99, 0.84, 0.81, 0.74, 0.61, 0.16 variables respectively. PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9 has 24%, 15%, 13%, 10%, 9%, 9%, 8%, 6%, 1% of the total information.

With PC1 and PC2, we can get 40% of the total information. With PC1, PC2, and PC3, we can get 53% of the total information. With PC1, PC2, PC3, PC4, and PC5, we can get 74% of the total information. With PC1, PC2, PC3, PC4, PC5, PC6, and PC7, we can get 91% of the total information.

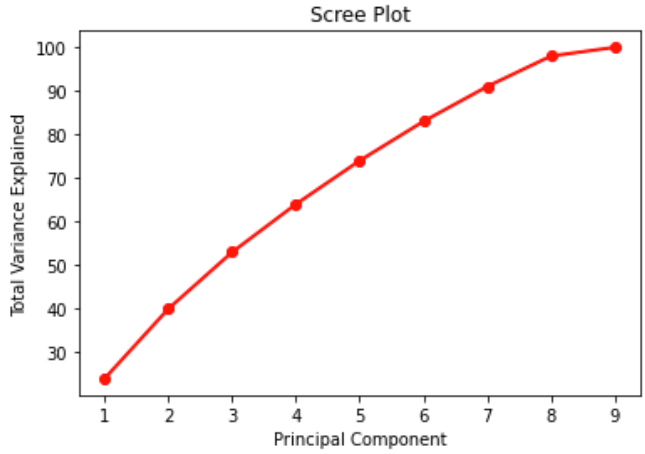
Scree Plot



This plot shows the eigenvalues of each principal component. In the plot, we can see that PC1 has as much information as approximately 2.20 variables. From the analysis, we know that is true and the more exact value is 2.17 variables worth of information for PC1. The plot as well as the analysis shows that around PC4 is when the principal components start to have less than 1 variable's worth of information. The plot starts to flatten out at PC5.



This scree plot looks at the percentage of the total information that each principal component has. This plot also starts to flatten out at PC5.



This scree plot looks at the running total percentage of information.

After looking at the scree plots, I decided to use PC1, PC2, PC3, PC4, and PC5 to get 74% of the total information.

Principal Components with its Weight Values

	PC1	PC2	PC3	PC4	PC5
TRUNC_IMP_MORTDUE	0.599264	-0.048214	-0.233698	0.196359	0.104302
TRUNC_IMP_VALUE	0.595345	-0.107601	-0.178048	0.196608	0.156791
TRUNC_IMP_YOJ	0.012565	-0.266796	0.552498	-0.271267	0.460117
TRUNC_IMP_DEROG	-0.028194	0.520258	0.200053	0.425923	0.236893
TRUNC_IMP_DELINQ	0.062710	0.378211	0.489581	0.309360	-0.372277
TRUNC_IMP_CLAGE	0.245347	-0.316222	0.480113	-0.114906	-0.055883
TRUNC_IMP_NINQ	0.061608	0.519724	-0.062358	-0.357743	0.592034
TRUNC_IMP_CLNO	0.420290	0.159582	0.288644	-0.139703	-0.164546
TRUNC_IMP_DEBTINC	0.202304	0.325006	-0.104911	-0.641396	-0.421212

The weights of each principal component in each variable are shown in the figure above. PC1 is made up of larger TRUNC_IMP_MORTDUE, TRUNC_IMP_VALUE, so it looks at larger values in TRUNC_IMP_MORTDUE and TRUNC_IMP_VALUE. PC2 has a larger TRUNC_IMP_NINQ and TRUNC_IMP_DEROG component, so PC2 looks at larger values in those two variables. PC3 has a larger TRUC_IMP_YOJ component. PC4 has a larger TRUNC_IMP_DEBTINC component, and it looks at smaller debt to income ratios due to the negative sign. PC5 has a larger TRUNC_IMP_NINQ component.

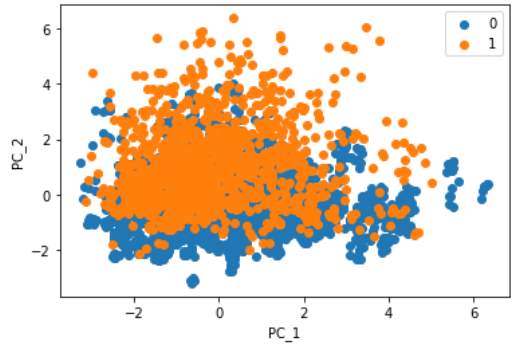
Principal Components with Target and Categorical Variables Dataframe

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

	TARGET_LOSS_AMT	REASON	JOB
0	641.0	HomeImp	Other
1	1109.0	HomeImp	Other
2	767.0	HomeImp	Other
3	1425.0	NaN	NaN
4	NaN	HomeImp	Office

Scatter plot with PC_1 and PC_2 and values of TARGET_FLAG

```
: for Name, Group in X_PCA.groupby(TARGET_FLAG):
      plt.scatter(Group.PC_1, Group.PC_2, label=Name)
  plt.xlabel("PC_1")
  plt.ylabel("PC_2")
  plt.legend()
  plt.show()
```



The plot shows that if the PC_1 value is greater than 5, then the loan will most likely not default. If the PC_2 value is larger than 4, then the loan is more likely to default. If the PC_2 value is less than -2, then the loan will more likely not default.