Web Personalization, lab 2

Group 10

Per Erik Kristiansson

Iely Chabrouni

Kaisheng Lin

Task 1: Analyze the Dataset (10 pts)

Open the ARFF file in a text editor and familiarize yourself with the nature of the data at hand. What type of task was this data collected for?

It would seem it was used to perhaps model loan application suitability. I'm unsure as to how the "good" and "bad" class has been added to the data. Perhaps in retrospect. So one can indicate what in the data leads to a "bad" rating in future instances.

How was it prepared, ie. can we expect any systematic errors or missing values in the data? Give a brief description of your findings.

It doesn't say exactly how it was prepared. Perhaps that would give us some insight in where some kind of bias has affected the data.

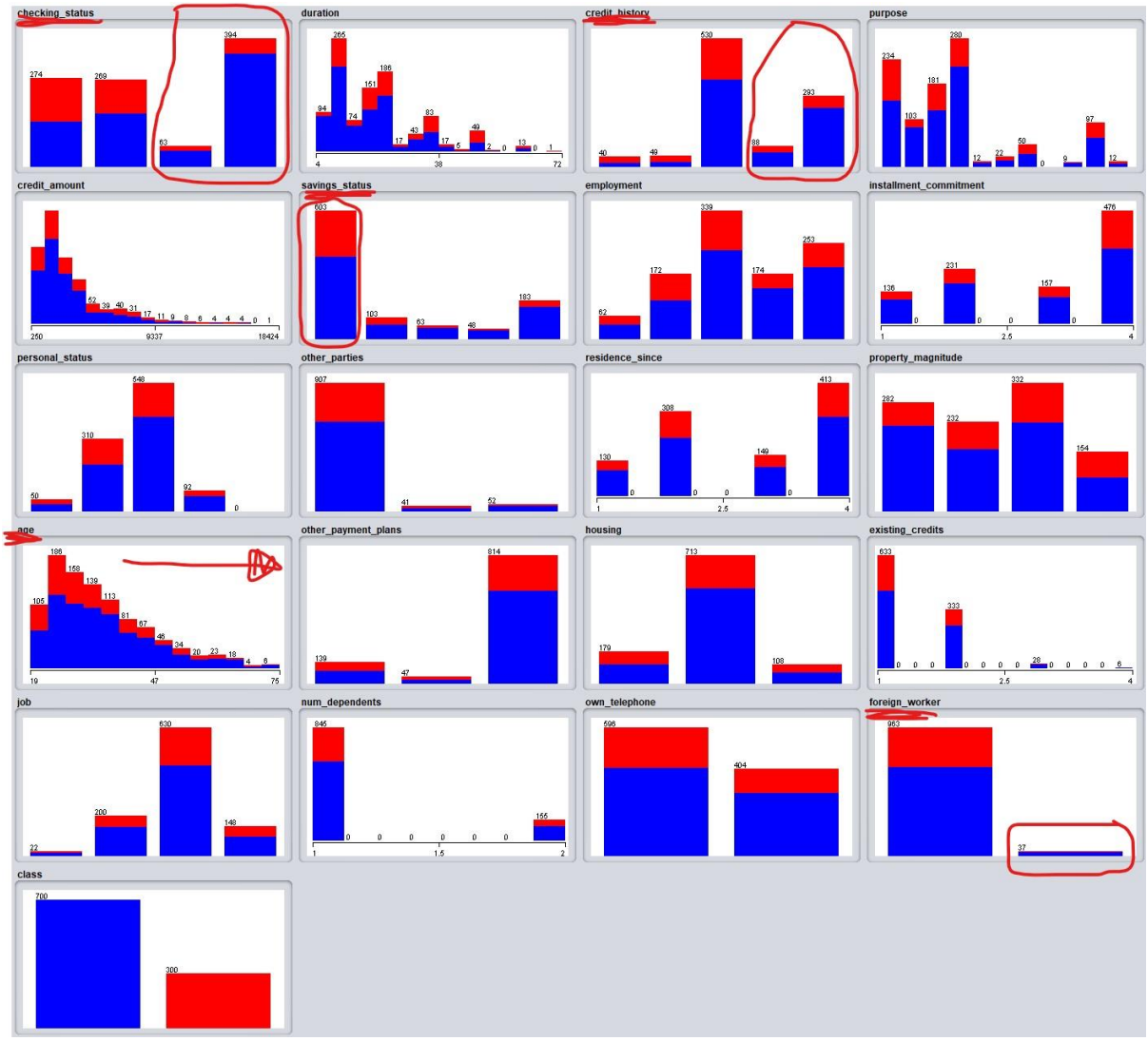
About missing data; the document doesn't really say how it deals with missing values, so unless we validate all the fields, we wont know (WEKA should show us if any fields are missing, however).

When it comes to missing information though, perhaps one could view some of the class values as missing information at least. For example, when it comes to personal_status, some are bunched together while some are not. If at one point, someone created some proof that these statuses (or class values) can be thrown together without affecting the model much negatively, then perhaps it ok. But again, the document says nothing about it.

Open the dataset in WEKA Explorer. Inspect the different attributes and their distributions over all instances. Can you find any interesting aspects, trivial correlations, or even identify useless features (if so, remove them)?

Just looking at the histograms we can find a few interesting qualities:

- checking_status – over 200 seem to have a larger relation of “good” verdicts
- credit_history – “delayed” or “critical” seems to have a larger relation of “good” verdicts (?!)
- savings_status – low savings seem to seems to have a larger relation of “good” verdicts (?!)
- age – the older age, the larger the relation of “good” verdicts
- foreign_worker – not being a foreign worker yields higher ratio of “good” verdicts



Which attributes appear promising wrt. the task? Use the visualizations in the Preprocess and Visualize tabs in WEKA Explorer and include annotated screenshots to support your findings.

We would argue the above mentioned attributes seems promising with regards to the task. Since the value we’re compring to is a nominal value (a class value), it’s hard to see any linear correlations such as with a numerical value. So with that said, there is no point in including an annotated screenshot of this specific task.

Task 2: Clustering (10 pts)

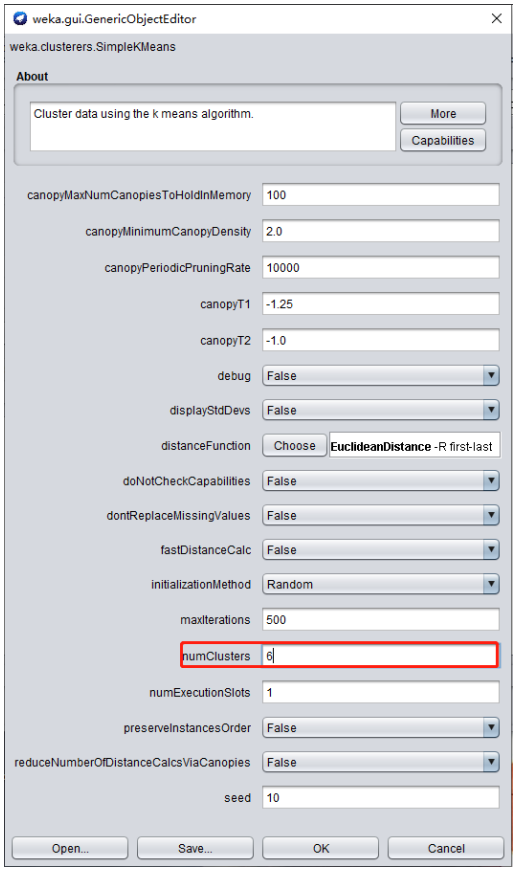
In the Cluster tab, perform several types of clusterings (k-means, Hierarchical Clustering, possibly others). Compare different settings, e.g. by varying the number of cluster centroids. Try to interpret the derived clusters and how well they describe the data.

Choose the ARFF file credit-g and choose and simpleKMeans. Keep the setting in default.

`Within cluster sum of squared errors: 5365.9976202840735` Then got the sum of squared errors is 5365.9976202847035. And there are 2 clusters in it

```
Final cluster centroids:
                                                Cluster#
Attribute                               Full Data                0                1
                                         (1000.0)           (643.0)          (357.0)
================================================================================================
checking_status                        no checking      no checking              <0
duration                                    20.903          19.9285          22.6583
credit_history                        existing paid    existing paid    existing paid
purpose                                    radio/tv         radio/tv         new car
credit_amount                            3271.258        2924.7869        3895.2941
savings_status                               <100             <100             <100
employment                                 1<=X<4           1<=X<4               >=7
installment_commitment                      2.973           2.9611           2.9944
personal_status                       male single      male single      male single
other_parties                                none             none             none
residence_since                             2.845           2.5599           3.3585
property_magnitude                            car              car no known property
age                                        35.546          33.2364          39.7059
other_payment_plans                          none             none             none
housing                                       own              own              own
existing_credits                            1.407           1.3701           1.4734
job                                       skilled          skilled          skilled
num_dependents                              1.155           1.1011           1.2521
own_telephone                                none             none              yes
foreign_worker                                yes              yes              yes
```

Second, changed the numClusters in setting from 2 to 6. Got the sum of squared errors is 4613.817664697715. Because this one is smalle than last one. That means this simpleKMeans with setting numCluster 6 is better than last simpleKMeans . The smaller the value, the smaller the distance between instances of the same cluster.




```
Within cluster sum of squared errors: 4613.817664697715
```

Attribute	Full Data (1000.0)	0 (225.0)	1 (115.0)	2 (81.0)	3 (189.0)	4 (121.0)
checking_status	no checking	no checking	<0	0<=X<200	no checking	<0
duration	20.903	23	29.0522	16.8148	19.8095	18.7934
credit_history	existing paid	critical/other existing credit	existing paid	existing paid	existing paid	critical/other existing credit
purpose	radio/tv	new car	used car	radio/tv	radio/tv	new car
credit_amount	3271.258	3839.7822	5825.0174	2274.4074	2914.5926	2701.7686
savings_status	<100	<100	<100	<100	no known savings	<100
employment	1<=X<4	1<=X<4	>=7	>=7	>=7	>=7
installment_commitment	2.973	3.1067	3.0957	2.8889	3.1058	2.7438
personal_status	male single	male single	male single	male single	male single	female div/dep/mar
other_parties	none	none	none	none	none	none
residence_since	2.845	2.6844	3.4957	2.6914	3.1217	3.4793
property_magnitude	car	car	no known property	real estate	life insurance	car
age	35.546	35.5022	43.4522	36.037	38.1693	32.1322
other_payment_plans	none	none	none	bank	none	none
housing	own	own	for free	own	own	rent
existing_credits	1.407	1.6489	1.2957	1.4074	1.381	1.595
job	skilled	skilled	high qualif/self emp/mgmt	unskilled resident	skilled	skilled
num_dependents	1.155	1.1333	1.2609	1.4691	1.164	1.0909
own_telephone	none	yes	yes	none	none	none
foreign_worker	yes	yes	yes	yes	yes	yes

There are 5 clusters in it.

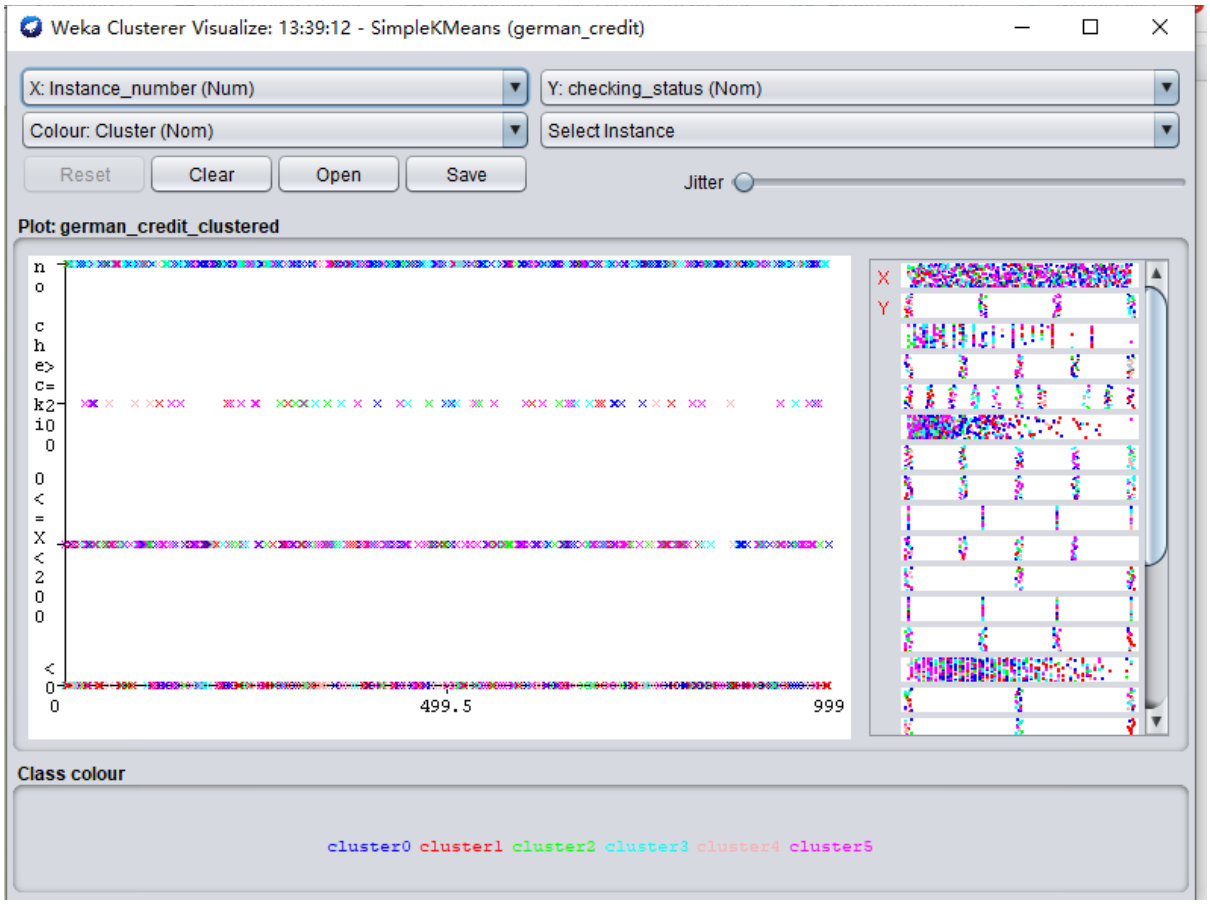
What do cluster centers look like for non-numerical attributes?

Non-numerical attributes cannot directly train the model, it need to encode, convert the attributes to numeric, and then train.

How does clustering work in this case? (Try to investigate or make an educated guess from the output.)

The K-means clustering algorithm first randomly selects K objects as the initial clustering centers. Then calculate the distance between each object and each seed cluster center, and assign each object to the cluster center closest to it. The cluster centers and the objects assigned to them represent a cluster. Once all objects are assigned, the cluster center of each cluster will be recalculated based on the existing objects in the cluster. This process will continue to repeat until a certain termination condition is met. The termination condition can be that no (or minimum number) of objects are reassigned to different clusters, no (or minimum number) of cluster centers change again, and the sum of squared errors is locally minimum

Compare your interpretation using the actual class label for evaluation. Inspect the cluster assignments by visualizing the output (right-click in the Result list to open menu). Document and annotate insightful visualizations over the different attributes.



"Color" is the basis for coloring the scatter plot, and the default is to label instances with different colors based on different "Cluster".

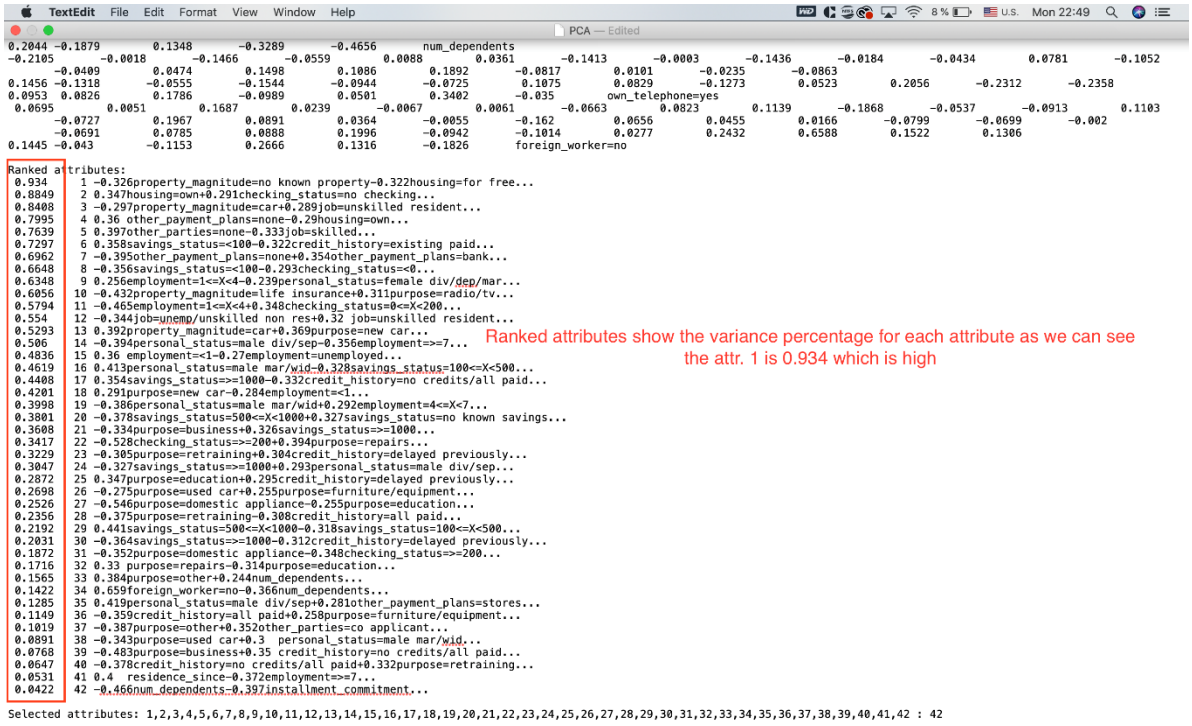
Task 3: Dimensionality Reduction (10 pts)

To reduce the number of features, in the 'Select attributes' tab, perform a Principal Component Analysis. Analyze and interpret the output.

After applying the principle components attribute selection, the output result shown that the Correlation Matrix that used from our data set the 52 features for the matrix dimensions and retrieves the Standard deviation with 1 in diagonal, in the correlation matrix where the standard deviation is 1 we can identify the correlation between the feature from column with the row corresponding with deviation value which is 1 in our matrix.

So we can see the correlation between features from our data set.

The most important function for Principal Component is the Ranked Attributes which show us the variance covered by the attributes, from this result illustration we can keep just the top ranked variation as the first variation is 93% which is enough or combining other attributes to have a high variation percentage value.




```
0.2044 -0.1879       0.1348       -0.3289       -0.4656       num_dependents
-0.2105       -0.0018       -0.1466       -0.0559       0.0088       0.0361       -0.1413       -0.0003       -0.1436       -0.0184       -0.0434       0.0781       -0.1052
       -0.0409       0.0474       0.1498       0.1086       0.1892       -0.0817       0.0101       -0.0235       -0.0863
0.1456 -0.1318       -0.0555       -0.1544       -0.0944       -0.0725       0.1075       0.0829       -0.1273       0.0523       0.2056       -0.2312       -0.2358
0.0953  0.0826       0.1786       -0.0989       0.0501       0.3402       -0.035       own_telephone=yes
 0.0695       0.0051       0.1687       0.0239       -0.0067       0.0061       -0.0663       0.0823       0.1139       -0.1868       -0.0537       -0.0913       0.1103
       -0.0727       0.1967       0.0891       0.0364       -0.0055       -0.162       0.0656       0.0455       0.0166       -0.0799       -0.0699       -0.002
       -0.0691       0.0785       0.0888       0.1996       -0.0942       -0.1014       0.0277       0.2432       0.6588       0.1522       0.1306
0.1445 -0.043       -0.1153       0.2666       0.1316       -0.1826       foreign_worker=no

Ranked attributes:
 0.934     1 -0.326property_magnitude=no known property-0.322housing=for free...
 0.8849    2 0.347housing=own+0.291checking_status=no checking...
 0.8408    3 -0.297property_magnitude=car+0.289job=unskilled resident...
 0.7995    4 0.36 other_payment_plans=none-0.29housing=own...
 0.7639    5 0.397other_parties=none-0.333job=skilled...
 0.7297    6 0.358savings_status=<100-0.322credit_history=existing paid...
 0.6962    7 -0.395other_payment_plans=none+0.354other_payment_plans=bank...
 0.6648    8 -0.356savings_status=<100-0.293checking_status=<0...
 0.6348    9 0.256employment=1<=X<4-0.239personal_status=female div/dep/mar...
 0.6056   10 -0.432property_magnitude=life insurance+0.311purpose=radio/tv...
 0.5794   11 -0.465employment=1<=X<4+0.348checking_status=0<=X<200...
 0.554    12 -0.344job=unemp/unskilled non res+0.32 job=unskilled resident...
 0.5293   13 0.392property_magnitude=car+0.369purpose=new car...
 0.506    14 -0.394personal_status=male div/sep-0.356employment=>=7...
 0.4836   15 0.36 employment=<1-0.27employment=unemployed...
 0.4619   16 0.413personal_status=male mar/wid-0.328savings_status=100<=X<500...
 0.4408   17 0.354savings_status=>=1000-0.332credit_history=no credits/all paid...
 0.4201   18 0.291purpose=new car-0.284employment=<1...
 0.3998   19 -0.386personal_status=male mar/wid+0.292employment=4<=X<7...
 0.3801   20 -0.378savings_status=500<=X<1000+0.327savings_status=no known savings...
 0.3608   21 -0.334purpose=business+0.326savings_status=>=1000...
 0.3417   22 -0.528checking_status=>=200+0.394purpose=repairs...
 0.3229   23 -0.305purpose=retraining+0.304credit_history=delayed previously...
 0.3047   24 -0.327savings_status=>=1000+0.293personal_status=male div/sep...
 0.2872   25 0.347purpose=education+0.295credit_history=delayed previously...
 0.2698   26 -0.275purpose=used car+0.255purpose=furniture/equipment...
 0.2526   27 -0.546purpose=domestic appliance-0.255purpose=education...
 0.2356   28 -0.375purpose=retraining-0.308credit_history=all paid...
 0.2192   29 0.441savings_status=500<=X<1000-0.318savings_status=100<=X<500...
 0.2031   30 -0.364savings_status=>=1000-0.312credit_history=delayed previously...
 0.1872   31 -0.352purpose=domestic appliance-0.348checking_status=>=200...
 0.1716   32 0.33 purpose=repairs-0.314purpose=education...
 0.1565   33 0.384purpose=other+0.244num_dependents...
 0.1422   34 0.659foreign_worker=no-0.366num_dependents...
 0.1285   35 0.419personal_status=male div/sep+0.281other_payment_plans=stores...
 0.1149   36 -0.359credit_history=all paid+0.258purpose=furniture/equipment...
 0.1019   37 -0.387purpose=other+0.352other_parties=co applicant...
 0.0891   38 -0.343purpose=used car+0.3  personal_status=male mar/wid...
 0.0768   39 -0.483purpose=business+0.35 credit_history=no credits/all paid...
 0.0647   40 -0.378credit_history=no credits/all paid+0.332purpose=retraining...
 0.0531   41 0.4  residence_since-0.372employment=>=7...
 0.0422   42 -0.466num_dependents-0.397installment_commitment...

Selected attributes: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42 : 42
```

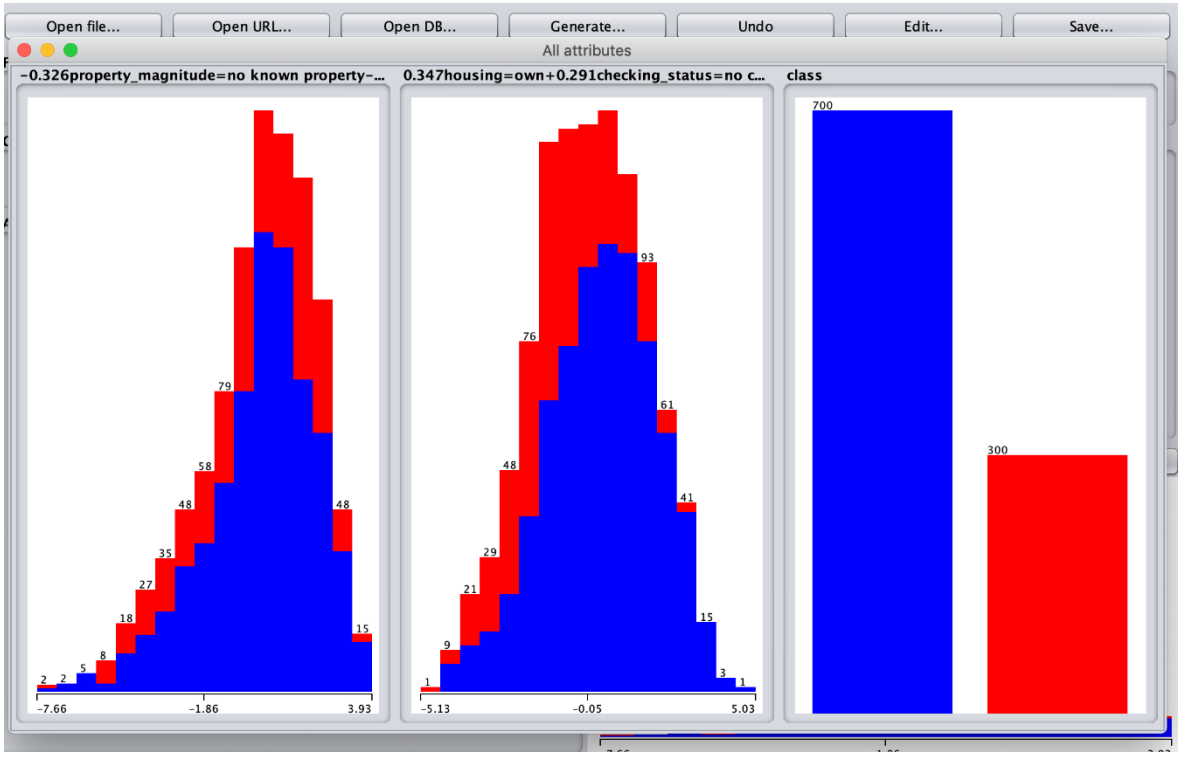
Visualize the transformed data. How many transformed dimensions are responsible for 80% of the variance in the data?

When we had the variance is set to 95% we used 42 ranked attributes and 21 input attributes for achieving the result, but now for 80% variance the transformed dimensions are only 32 ranked attributes with 21 input attributes, and the eigenvalue is still having the same value.

```
Ranked attributes:
 0.934     1 -0.326property_magnitude=no known property-0.322housing=for free-0.267credit_amount-0.237personal_status=male single-0.232job=high qualif/self emp/mgmt-0.222age...
 0.885     2 0.347housing=own+0.291checking_status=no checking+0.289credit_history=critical/other existing credit-0.266housing=rent+0.248existing_credits-0.247credit_history=existing paid...
 0.841     3 -0.297property_magnitude=car+0.289job=unskilled resident-0.289other_parties=none+0.249other_parties=guarantor+0.241property_magnitude=real estate-0.234job=skilled...
 0.799     4 0.36 other_payment_plans=none-0.29housing=own-0.268other_payment_plans=bank+0.242residence_since-0.223other_payment_plans=stores-0.223purpose=business...
 0.764     5 0.397other_parties=none-0.333job=skilled-0.314other_parties=guarantor-0.236purpose=radio/tv-0.232personal_status=male single-0.23other_parties=co applicant...
 0.73      6 0.358savings_status=<100-0.322credit_history=existing paid+0.318credit_history=critical/other existing credit+0.318existing_credits-0.219savings_status=no known savings-0.208job=unskilled
resident...
 0.696     7 -0.395other_payment_plans=none+0.354other_payment_plans=bank+0.295job=skilled-0.294employment=unemployed-0.294job=high qualif/self emp/mgmt+0.219credit_history=all paid...
 0.665     8 -0.356savings_status=<100-0.293checking_status=<0+0.274housing=rent-0.267other_parties=none+0.251savings_status=100<=X<500+0.222checking_status=0<=X<200...
 0.635     9 0.256employment=1<=X<4-0.239personal_status=female div/dep/mar+0.225purpose=business+0.219num_dependents-0.21employment=unemployed+0.209credit_history=delayed previously...
 0.606    10 -0.432property_magnitude=life insurance+0.311purpose=radio/tv-0.255savings_status=no known savings-0.236purpose=furniture/equipment+0.221savings_status=<100+0.218checking_status=0<=X<200...
 0.579    11 -0.465employment=1<=X<4+0.348checking_status=0<=X<200-0.33checking_status=no checking+0.3  property_magnitude=life insurance+0.236employment=<1-0.209purpose=used car...
 0.554    12 -0.344job=unemp/unskilled non res+0.32 job=unskilled resident-0.304job=skilled+0.25 housing=rent-0.248employment=1<=X<4-0.233employment=unemployed...
 0.529    13 0.392property_magnitude=car+0.369purpose=new car+0.33 employment=4<=X<7-0.269purpose=furniture/equipment-0.268property_magnitude=life insurance-0.238personal_status=male div/sep...
 0.506    14 -0.394personal_status=male div/sep-0.356employment=>=7-0.272property_magnitude=car+0.217purpose=education-0.213residence_since+0.198employment=4<=X<7...
 0.484    15 0.36 employment=<1-0.27employment=unemployed+0.251checking_status=>=200-0.238residence_since-0.229installment_commitment-0.225savings_status=500<=X<1000...
 0.462    16 0.413personal_status=male mar/wid-0.328savings_status=100<=X<500-0.303purpose=education-0.244personal_status=female div/dep/mar-0.235job=unskilled resident+0.23 purpose=used car...
 0.441    17 0.354savings_status=>=1000-0.332credit_history=no credits/all paid+0.308employment=4<=X<7-0.304savings_status=no known savings-0.256employment=1<=X<4-0.204purpose=retraining...
 0.42     18 0.291purpose=new car-0.284employment=<1+0.258employment=1<=X<4-0.257job=unemp/unskilled non res-0.242personal_status=male single-0.215checking_status=>=200...
 0.4      19 -0.386personal_status=male mar/wid+0.292employment=4<=X<7+0.272personal_status=female div/dep/mar-0.225property_magnitude=life insurance+0.211savings_status=>=1000+0.208property_magnitude=real
estate...
 0.38     20 -0.378savings_status=500<=X<1000+0.327savings_status=no known savings+0.291other_payment_plans=stores+0.271installment_commitment-0.257employment=4<=X<7+0.252checking_status=<0...
 0.361    21 -0.334purpose=business+0.326savings_status=>=1000+0.307checking_status=0<=X<200+0.296purpose=used car-0.262employment=4<=X<7-0.26purpose=domestic appliance...
 0.342    22 -0.528checking_status=>=200+0.394purpose=repairs+0.367checking_status=no checking-0.195purpose=radio/tv+0.173credit_history=no credits/all paid-0.172housing=rent...
 0.323    23 -0.305purpose=retraining+0.304credit_history=delayed previously-0.243personal_status=male mar/wid-0.221credit_history=all paid+0.22 other_parties=guarantor-0.203credit_history=critical/other
existing credit...
 0.305    24 -0.327savings_status=>=1000+0.293personal_status=male div/sep-0.257credit_history=no credits/all paid-0.253purpose=retraining+0.237credit_history=delayed
previously+0.227savings_status=100<=X<500...
 0.287    25 0.347purpose=education+0.295credit_history=delayed previously-0.285savings_status=500<=X<1000-0.276purpose=domestic appliance+0.25 purpose=repairs-0.248credit_history=no credits/all paid...
 0.27     26 -0.275purpose=used car+0.255purpose=furniture/equipment-0.249purpose=education-0.239savings_status=100<=X<500-0.224checking_status=>=200-0.217purpose=domestic appliance...
 0.253    27 -0.546purpose=domestic appliance-0.255purpose=education+0.227credit_history=no credits/all paid-0.222checking_status=0<=X<200-0.207employment=<1+0.203checking_status=>=200...
 0.236    28 -0.375purpose=retraining-0.308credit_history=all paid+0.299savings_status=100<=X<500-0.289savings_status=500<=X<1000-0.278purpose=new car+0.223other_payment_plans=bank...
 0.219    29 0.441savings_status=500<=X<1000-0.318savings_status=100<=X<500+0.317savings_status=>=1000+0.291purpose=education+0.224purpose=repairs-0.221purpose=retraining...
 0.203    30 -0.364savings_status=>=1000-0.312credit_history=delayed previously-0.304purpose=other+0.271other_payment_plans=stores+0.247savings_status=no known savings-0.21num_dependents...
 0.187    31 -0.352purpose=domestic appliance-0.348checking_status=>=200-0.34purpose=repairs-0.289other_payment_plans=stores+0.241purpose=business+0.22 purpose=education...

Selected attributes: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31 : 31
```

Investigate a representation using only the first two dimensions for easier visualization. What does the distribution look like?

```
Ranked attributes:
 0.934    1 -0.326property_magnitude=no known property-0.322housing=for free-0.267credit_amount-0.237personal_status=male single-0.232job=high qualif/self emp/mgmt-0.222age...
 0.885    2 0.347housing=own+0.291checking_status=no checking+0.289credit_history=critical/other existing credit-0.266housing=rent+0.248existing_credits-0.247credit_history=existing paid...

Selected attributes: 1,2 : 2
```



How many cluster centers would you consider a good choice to model the data? Document and annotate the visualization.

The best results we got when we use the K-Means with splitting the data set with 77% and 4 clusters, but with 7 iteration we got the best results with 4 clusters, 10 iterations retrieve the kMeans

Number of iterations: 10

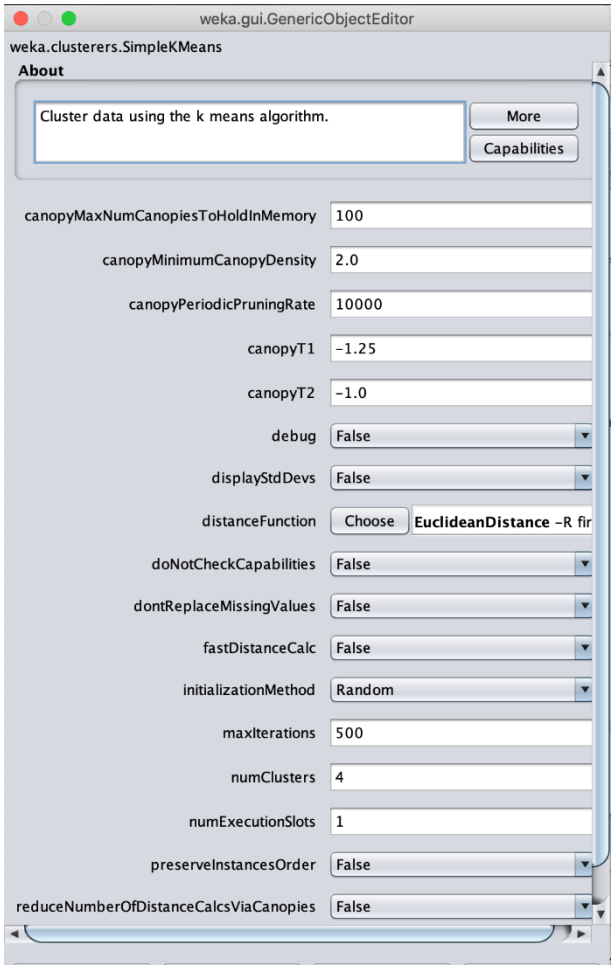
Within cluster sum of squared errors: 5226.216042637847

gives the best results for the model.

And for 7 iteration is

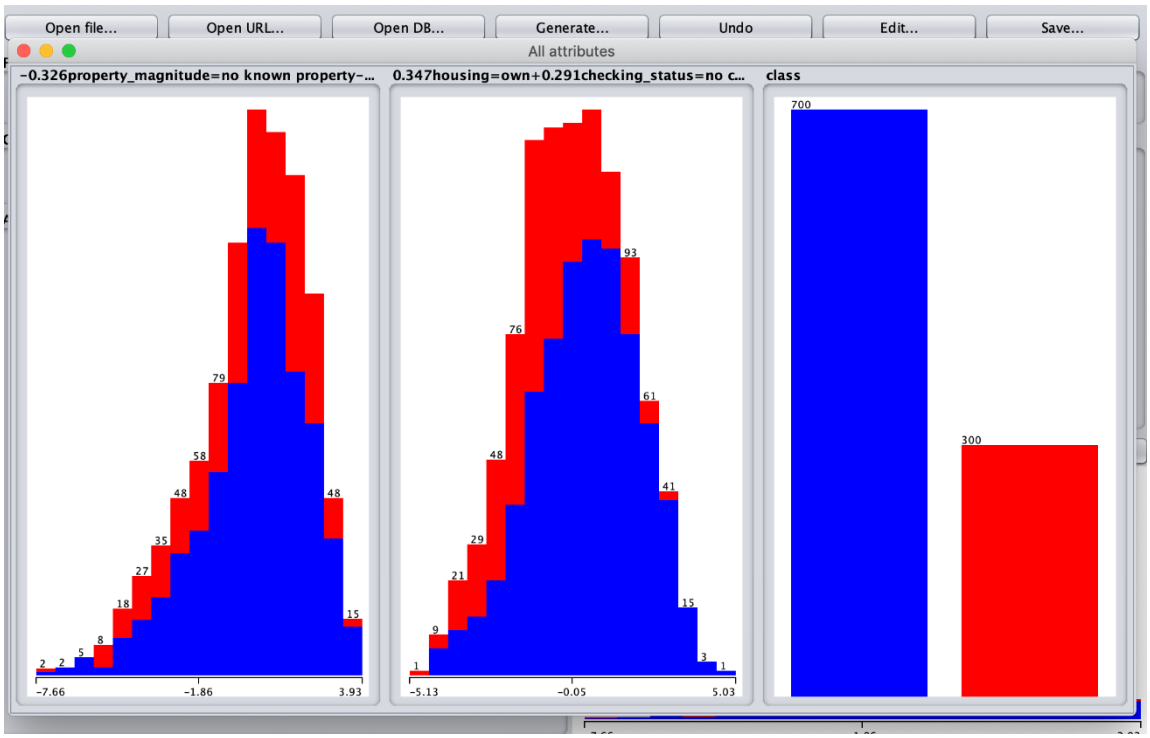
Number of iterations: 7

Within cluster sum of squared errors: 3996.8056845794376




```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on test split ===

kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 3996.8056845794376

Initial starting points (random):

Cluster 0: 'no checking',15,'critical/other existing credit',furniture/equipment,2788,<100,4<=X<7,2,'female div/dep/mar','co applicant',3,car,24,bank,own,2,skilled,1,none,yes,good
Cluster 1: 0<=X<200,12,'critical/other existing credit','new car',958,<100,4<=X<7,2,'male single',none,3,'real estate',47,none,own,2,'unskilled resident',2,none,yes,good
Cluster 2: >=200,15,'existing paid',business,2687,<100,4<=X<7,2,'male single',none,4,'life insurance',26,none,rent,1,skilled,1,yes,yes,good
Cluster 3: 0<=X<200,45,'critical/other existing credit','used car',4576,100<=X<500,unemployed,3,'male single',none,4,car,27,none,own,1,skilled,1,none,yes,good

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                               Full Data           0                1                2                3
                                          (770.0)     (215.0)          (147.0)          (181.0)          (227.0)
=============================================================================================================================
checking_status                       no checking  no checking           0<=X<200               <0       no checking
duration                                  21.1169     19.6884            17.1701          25.0884          21.859
credit_history                      existing paid  existing paid       existing paid    existing paid    existing paid
purpose                                radio/tv furniture/equipment          new car         radio/tv         radio/tv
credit_amount                          3333.939     2965.3023          2960.483         4543.3039        2960.6344
savings_status                            <100         <100               <100             <100             <100
employment                              1<=X<4       1<=X<4             1<=X<4           1<=X<4             >=7
installment_commitment                   2.9727       2.6698             2.6939           3.0055           3.4141
personal_status                     male single  female div/dep/mar  male single     male single     male single
other_parties                              none         none               none             none             none
residence_since                          2.8961       2.5488             2.6327           3.0331           3.2863
property_magnitude                          car          car          real estate  life insurance          car
age                                     35.3039      30.7163            34.5782          36.8729          38.8678
other_payment_plans                        none         none               none             none             none
housing                                     own          own                own              own              own
existing_credits                         1.4247       1.3349             1.3741           1.4199           1.5463
job                                     skilled      skilled  unskilled resident          skilled          skilled
num_dependents                           1.1558       1.0186             1.2857           1.1547           1.2026
own_telephone                              none         none               none              yes             none
foreign_worker                              yes          yes                yes              yes              yes
class                                      good         good               good             good             good
```

Repeat the steps of Task 2, this time with the lower dimensional representation obtained before. In the 'Preprocess' tab, use the PrincipalComponents filter to reduce the data to two dimensions. Perform a clustering with the number of identified cluster centers. Interpret the results.



By using K-Means with reduced dimensions and used the filter Principal components in the preprocess, we got those results.

With K=2 and iteration nr of 6 we got the best result which is K-Means and set split with 77%

======

Number of iterations: 6

Within cluster sum of squared errors: 30.877542416578706

Initial starting points (random):

Cluster 0: 0.871844,-0.066082, good

Cluster 1: 0.211264,2.028087, good

Cluster 2: 0.312764,-1.243008, good

Clustered Instances

0 74 (32%)

1 95 (41%)

2 61 (27%)

By using the Hierarchical Cluster and same splitting with 77% we get results as shown with 2 clusters get the best results

Clustered Instances

0 61 (27%)

1 169 (73%)

The correlation matrix show that we get the same results in both for correlation between instances.

……..