Question 1:

Before we start choosing the 5 most interesting/valuable attributes in a dataset, it may be helpful to find the correlation between the 'credit rating' attributes to set a good base foundation on which attributes we're able to investigate for a start.

We should not fully rely on correlation to find the 5 best attributes as it must be logically linked to credit rating on the other hand. Even though attributes such as self-employed has a rather high correlation score, it is said that self-employment does not directly affect one's credit score. This is the same for the gender attribute.

Find the strength of relations using correlation between "credit rating" and other continuous dataset to determine the 'better' options to select the initial 5 most interesting & valuable attributes

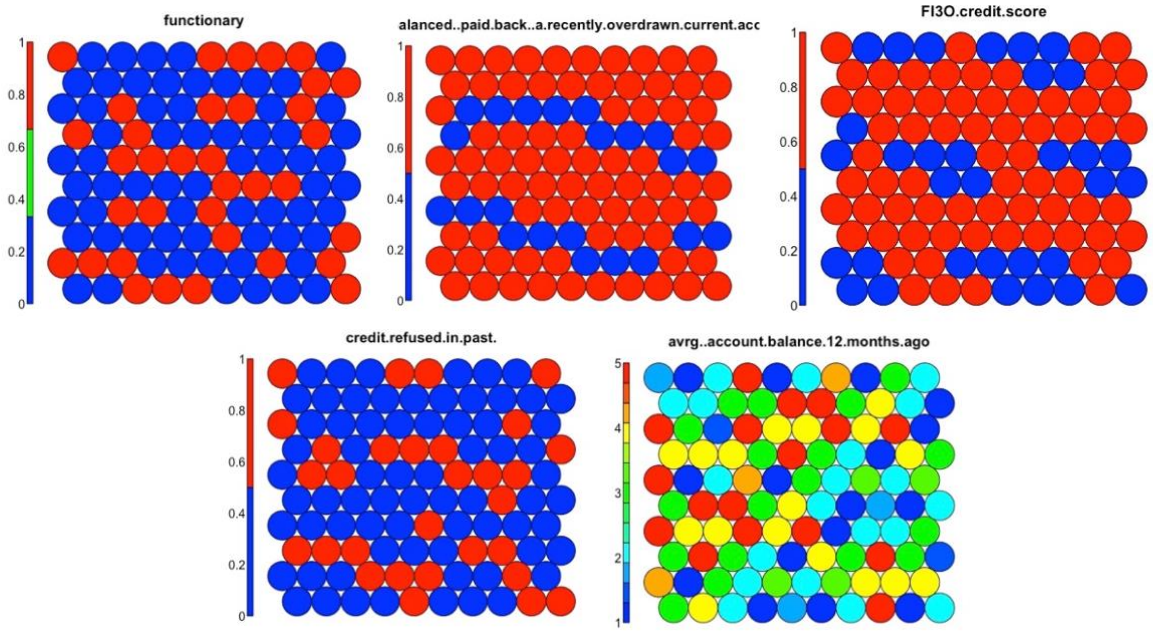
re-balanced (paid back) a recently overdrawn current acount	0.264230
FI3O credit score	0.384746
gender	0.021313
0. accounts at other banks	0.004717
credit refused in past?	0.027453
years employed	-0.004292
savings on other accounts	0.316520
self employed?	0.008360
max. account balance 12 months ago	0.003284
min. account balance 12 months ago	0.043381
avrg. account balance 12 months ago	0.047918
max. account balance 11 months ago	0.026303
min. account balance 11 months ago	-0.030903
avrg. account balance 11 months ago	-0.015363
max. account balance 10 months ago	0.013235
min. account balance 10 months ago	-0.003449
avrg. account balance 10 months ago	-0.004361
max. account balance 9 months ago	-0.014238
min. account balance 9 months ago	-0.014849
avrg. account balance 9 months ago	0.046329
max. account balance 8 months ago	0.060146
min. account balance 8 months ago	0.006631
avrg. account balance 8 months ago	0.024926
max. account balance 7 months ago	-0.013925
min. account balance 7 months ago	-0.023441
avrg. account balance 7 months ago	0.027659
max. account balance 6 months ago	-0.019120
min. account balance 6 months ago	-0.014914
avrg. account balance 6 months ago	-0.010455
max. account balance 5 months ago	0.005235
min. account balance 5 months ago	0.008704
avrg. account balance 5 months ago	-0.017618
max. account balance 4 months ago	-0.009898
min. account balance 4 months ago	-0.039467
avrg. account balance 4 months ago	-0.002622
max. account balance 3 months ago	0.020899
min. account balance 3 months ago	0.011385
avrg. account balance 3 months ago	0.003409
max. account balance 2 months ago	0.000888
min. account balance 2 months ago	0.006640
avrg. account balance 2 months ago	-0.013817
max. account balance 1 months ago	-0.002227
min. account balance 1 months ago	0.000553
avrg. account balance 1 months ago	-0.026852

From the above we can first nominate 'rebalanced a recently overdrawn current account', 'credit score', and 'credits refused in past' due to the high correlation score. From the above, 1 to 12 months average account balance, the 'avrg. account balance 12 months ago

has the highest correlation score among other averages account balance from other months. All of which makes a good candidate to compared with 'credit rating'

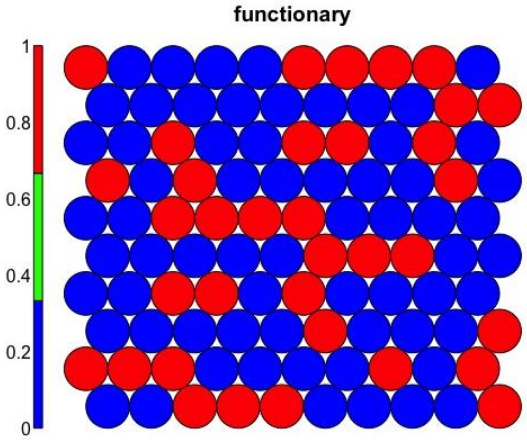
5 Most interesting and valuable attributes:

Currently xdim = 10 and ydim = 10 which creates 100 neurons in the SOM grid. Using a hexagonal topology would mean that each neurons has 6 neighboring neurons hence having the 'honeycomb' look.



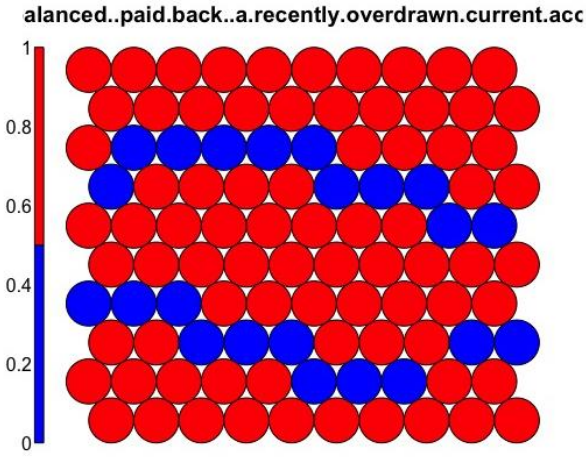
---

1. <u>Functionary</u>



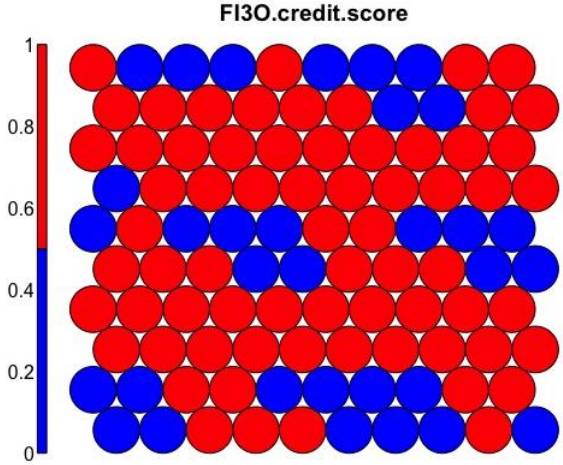
Functionary shows if a client is of someone who works official duties and jobs such as office-bearer, public servant, or civil servant. Functionary can also mean different types of people such as retired, contribution to economy, or businessmen for example. From the above plot, most of the dataset is closely related and are of distinct colors such as low-valued nodes blues which represent false and high-valued nodes for true. This can be said that the likelihood of having a higher credit rating as the number of high-valued nodes (red) is lesser and sparser. Hence it has close correlation with the 'credit rating' attribute which makes the functionary attribute a good candidate to select.

2. Re-balanced (paid back) a recently overdrawn current account



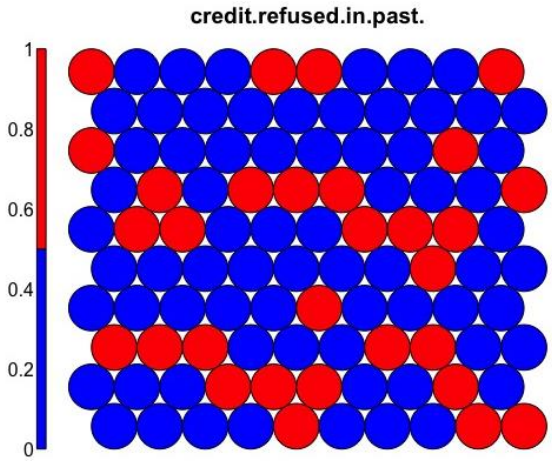
Similarly, to the functionary attribute, this attribute has a distinct plot which makes a good candidate to select. As seem from the plot above due to the high number of high-valued nodes (reds) and a lot lesser low-valued nodes, it can be said that this feature has very high correlation to determine a client's chances of having a high credit rating. it can be said that many from the has successfully paid back a recently overdrawn account. In such making these group of training dataset to cause a good/increased/positive credit rating score.

3. FI3O credit score



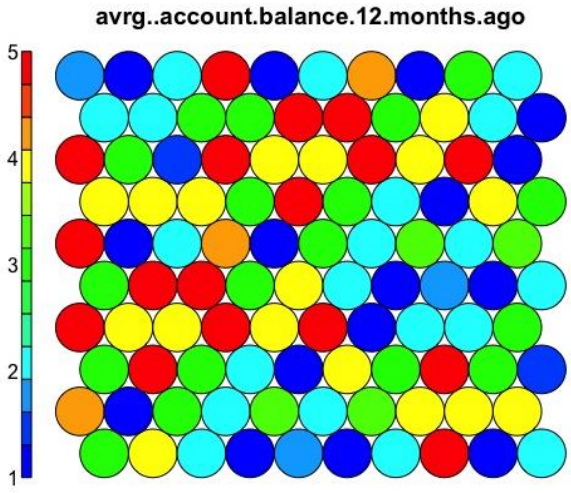
From the above plot, most of the dataset is showing red (good credit score) as compared to a handful of blue (bad credit score). It is an important indicator for banks to determine one person's ability to repay loans by using credit score. Credit scores is also used to determine if banks will most likely approve for personal loans and the interest rates to pay.

4. Credit refused in past



Credits refused in the past may be highly linked to that client having late or missed payments in their credit history. It can be used to determine if someone will have no problem or lots of problems repaying debts in the past. As seem from above, even though there's more blues dataset (no credits refused) as compared to the red dataset (credits refused), there's still a lot of people who have had troubles repaying debts. This will hence cause a (significant) drop or refusal in future loans.

5. Average account balance 12 months ago

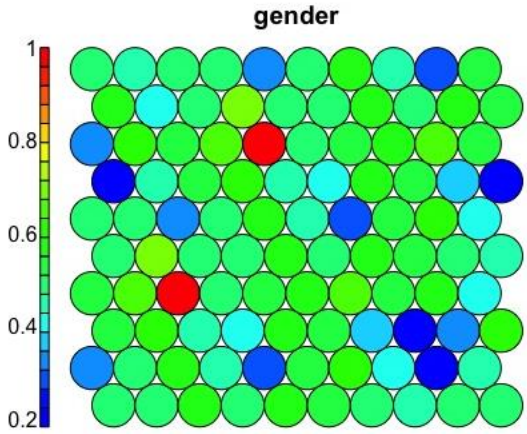


The term Trailing Twelve Months (TTM) is an important factor in the finance field as it is an effective way to analyze the most recent financial data of a person in an annualized format. It shows the spending/saving power of a person which can highly determine a person's credit rating as seen from above, there're more nodes in the 1 – 2 low-valued zone. Banks will be more likely to see as a good gauge to determine how much money one have in the bank 12 months ago.

Examples of attributes with high correlation with 'credits rating' but was not chosen:

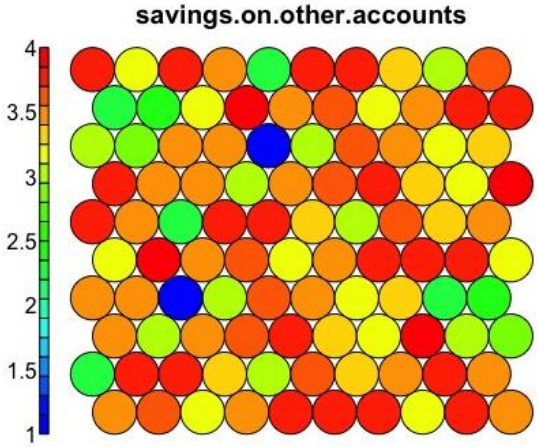
Even though some of these attributes has a high correlation score, it is not suitable as we can only use correlation as a good stepping-stone but should not rely 100% on it. We should also consider factors such as to logically complement the 'credits rating' attribute all in all.

1. Gender



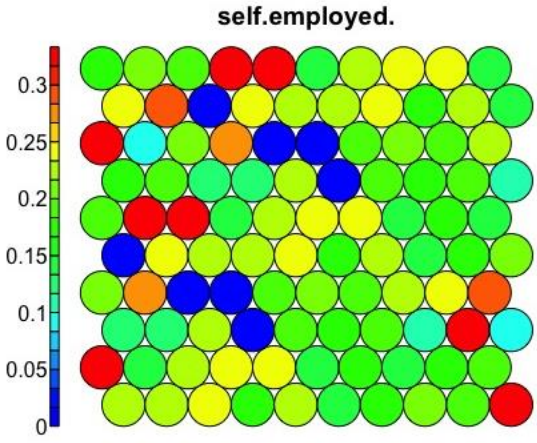
Even though the gender attribute has a high correlation of 0.021313, because most of the colors in this plot is of the neutral zone with a handful of 2 low-valued nodes and +/- 5 high-valued nodes. This makes gender a less favorable candidate to nominate as a potential 5 most interesting attributes. Gender also does have close to no impact on the 'credits rating' making it not useful to be nominated.

2. Savings on other accounts



Savings on other accounts has a correlation of 0.0316520 against 'credits rating.' Even though this attribute returns a relatively high correlation score, as seen by the plot above, most of the nodes are in the 3 to 4 high-valued zone. The data here may be too large to have a distinctive impact against the 'credit rating' and one client can have many accounts with little money. It is also hard to achieve any useful data from other banks to determine a good credit rating due to many missing data. Hence the data in this attribute may not be useful to nominate for a proper data analysis

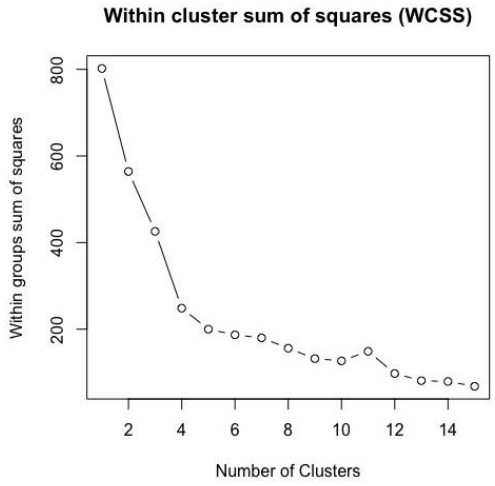
3. <u>Self employed</u>



4. Self-employed may be an important factor when trying to determine credit worthiness but does not have a good correlation score of 0.008360 when compared with 'credits rating.' As seen in the plot above, a large percentage of the data is of green color making most of the data in the neutral zone making it difficult to interpret the relationship with 'credits ratings' and will not help in the predictions in the later stage. Being self-employed has also close to no relationship between the 'credits rating' one will get based on being self-employed or not making this attribute not useful to nominate.

Conclusion:

As the number of clusters increases the value of Within Groups Sum of Squares (WCSS) starts to decrease significantly. The value of K is selected as the rate of WCSS decreases. As seen in the plot below, a big drop is seen as the number of clusters starts to get more. At the 'elbow' of the plot at the 5th number of cluster is when the drop starts slowing down hence it is good to take 6th as the optimal value of K. The number of clusters shown here is an estimate by plotting the 'elbow' point against the WCSS for each value of K.



Despite having 6 as the optimal K-value, the values in the credit rating attribute are only up till 3 making this model of prediction unable to achieve 100% accuracy even with the help of using correlation to find the most 'optimal' 5 most valuable attributes. Some of the data may also contains missing or NULL data making the analysis an inaccurate analysis.

Question 2:

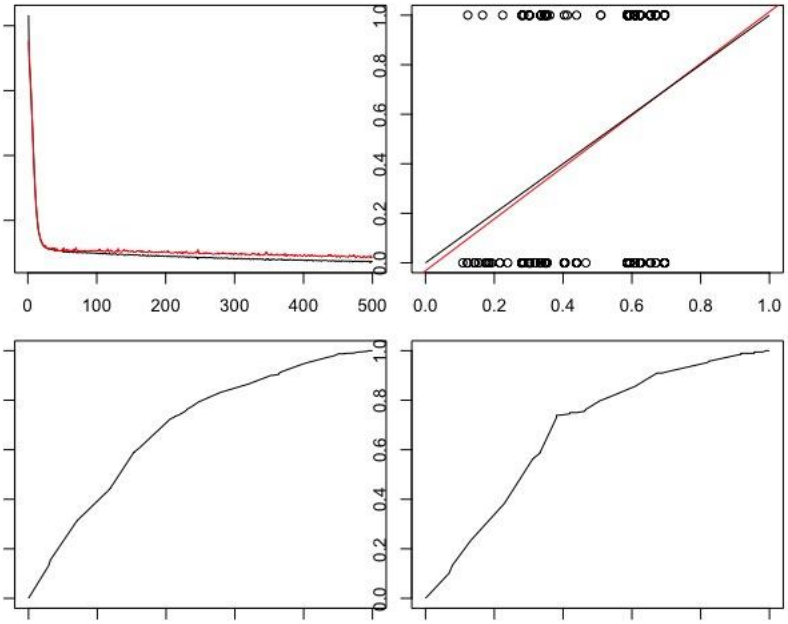
a.) Describe a valid strategy that maximises the accuracy of predicting the credit rating. Explain why your strategy can be expected to maximise the prediction capabilities.

In the creditworthiness file, there contains 2500 dataset with 46 variables. After removing any data with no value, int 0 or NULL in the 'credit rating' attribute, 1962 data is left remaining before commencing the SOM training. This is an important step to remove any noise or outliers in the database before the SOM training to get maximize the quality of prediction.

data_train	1962 obs. of 5 variables
fullDataSet	2500 obs. of 46 variables

This show that there's 538 such dataset which was not used in the SOM training. A way to include all the 2500 dataset within creditworthiness.csv is to calculate the mean of these missing values so that all the dataset is included and there is no biasness in the findings.

With the remaining 1962 dataset, the data is further split into 80% training and 20% test data. After commencing the training, the 20% of the test data will be used to determine the accuracy of the data. A second round of splitting will commence with the remaining initial 80% of the data left but this time into 60% training and 20% to testing using the confusion matrix.



Top Left Plot: The red line for test set and black lines for training set is very close to each other. This tells us that the predictions in the training set is very close to being accurate with a steady downward trend from both lines after iterating for 500 times. It shows a strong correlation pattern with each other.

Top Right Plot: This plot shows the linear regression error from the predictive model. The optimal line is very linearly close to the actual dataset. This tells us that the output from the

prediction model is almost nearly matching with the target values. The ideal and optimal plot would be the red line and the black line matches as close as possible for a perfect match.

Bottom Left Plot: This plot shows the Receiver Operating Characteristic (ROC) curve based on the training set. This helps with displaying the tradeoff between true-positive and false-positive rate. Since the curve is close to the top left corner of the boundary, it can be identified as a good classification model. It shows that the model not being underfitted nor overfitted with the data. This will then improve the accuracy of implementing the model on the test set.

Bottom Right Plot: This plot shows the Receiver Operating Characteristic (ROC) curve based on the testing set. The test set prediction is better as compared to the validation set predictions. If we do not overtrain the model, the model becomes more flexible to generalize when approaching the unseen data.

---

b.) Use your strategy to train MLP(s) then report your results. Give an interpretation of your results. What is the best classification accuracy (expressed in % of correctly classified data) that you can obtain for data that were not used during training (i.e. The test set)?

From the initial results as shown below, with the help of Machine Learning Program (MLP) the model is said to have an accuracy of 61.50%.

```
> confusionMatrix(trainset$targetsTrain, fitted.values(model))
        predictions
targets   1   2   3
      1 215 149  13
      2 123 581  78
      3  42 199 169
```

$$Accuracy = \frac{215 + 581 + 169}{215 + 149 + 13 + 123 + 581 + 78 + 42 + 199 + 169} * 100$$

$$= 61.50414277 \approx 61.50\%$$

By including these unknown values (NULL, empty or int 0) in the prediction that was initially not used during the training prediction, it helps the model to have a 100% usage of data instead of discarding useful and informative information (538 out of 2500 data). We therefore do not make any form of assumptions and weak correlations which will result in a more accurate model.
To include unknown values effectively, if the data is continuous like creditworthiness.csv, we can replace unknown values with either mean, median or mode values to include them. We can later treat them in a separate class to build models for those missing values.

The second method is to increase the hidden layers. The `mlp()` function helps to create and trains a multi-layer perceptron. This function learns the mapping of a set of input data to a

set of target values. Currently the MPL is trained to contain 5 neurons in a single hidden layer. By increasing to size 7 for example can create two hidden layer MPL.

The third method is to reselect the 5 most valuable data to try to achieve a higher accuracy rate to improve on the relationship with 'credit rating'. Selecting attributes with higher and more logically suitable attribute can be taken into consideration.

After the inclusion of the unknown data, the best classification accuracy has increased by 1.6% to 63.10%.

```
> confusionMatrix(trainset$targetsTest, predictTestSet)
       predictions
targets   1   2   3
      1  66  38   2
      2  37 139  12
      3  13  43  43
```

$$Accuracy = \frac{66 + 139 + 43}{66 + 38 + 2 + 37 + 139 + 12 + 13 + 43 + 43} * 100$$

$$= 63.1043257 \approx 63.10\%$$