EXERCISE -3

3.1 Problem Statement:

To List all the categorical attributes and the real-valued attributes separately in 'German credit' data set.

3.2 Description:

About Dataset

The objective of the German Credit Data is to minimize the chances of issuing risky loans to applicants while maximizing the chances of profiting from good loans. An applicant's demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

The German Credit data set is a publically available data set downloaded from the UCI Machine Learning Repository. The German Credit Data contains data on 20 variables and the classification of whether an applicant is considered a Good or Bad credit risk for 1000 loan applicants. The task requires exploring the data and building a predictive model to provide a bank manager guidance for making a decision on whether to approve a loan to a prospective applicant based on his/her profile.

the original dataset that only has 19 variables:

1. Checking_Status
2. Credit_history
3. Purpose
4. Savings_status
5. Employment
6. Personal_status
7. Other_parties
8. Property_Magnitude

9. Other_payment_plans
10. Housing
11. Job
12. Own_telephone
13. Foreign_worker
14. Duration
15. Credit_amout
16. Installment_Commitment
17. Residence_since
18. Age
19. Existing_credits

About Arff:

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types.

Lines that begin with a % are comments. The **@RELATION**, **@ATTRIBUTE** and **@DATA** declarations are case insensitive.

About CSV:

Files with .csv (Comma Separated Values) extension represent plain text files that contain records of data with comma separated values. Each line in a CSV file is a new record from the set of records contained in the file. Such files are generated when data transfer is intended from one storage system to another. Since all applications can recognize records separated by comma, import of such data files to database is done very conveniently. Almost all spreadsheet applications such as Microsoft Excel or OpenOffice Calc can import CSV without much effort. Data imported from such files is arranged in cells of a spreadsheet for representation to user.

Datatypes that are supported by Weka:

- numeric
- integer is treated as numeric
- real is treated as numeric
- [nominal-specification]
- string
- date [date-format]
- relational for multi-instance data (for future use)

where [nominal-specification] and [date-format] are defined below. The keywords **numeric**, **real**, **integer**, **string** and **date** are case insensitive.

Numeric attributes
Numeric attributes can be real or integer numbers.

Nominal attributes
Nominal values are defined by providing an [nominal-specification] listing the possible values: {[nominal-name1], [nominal-name2], [nominal-name3], ...}

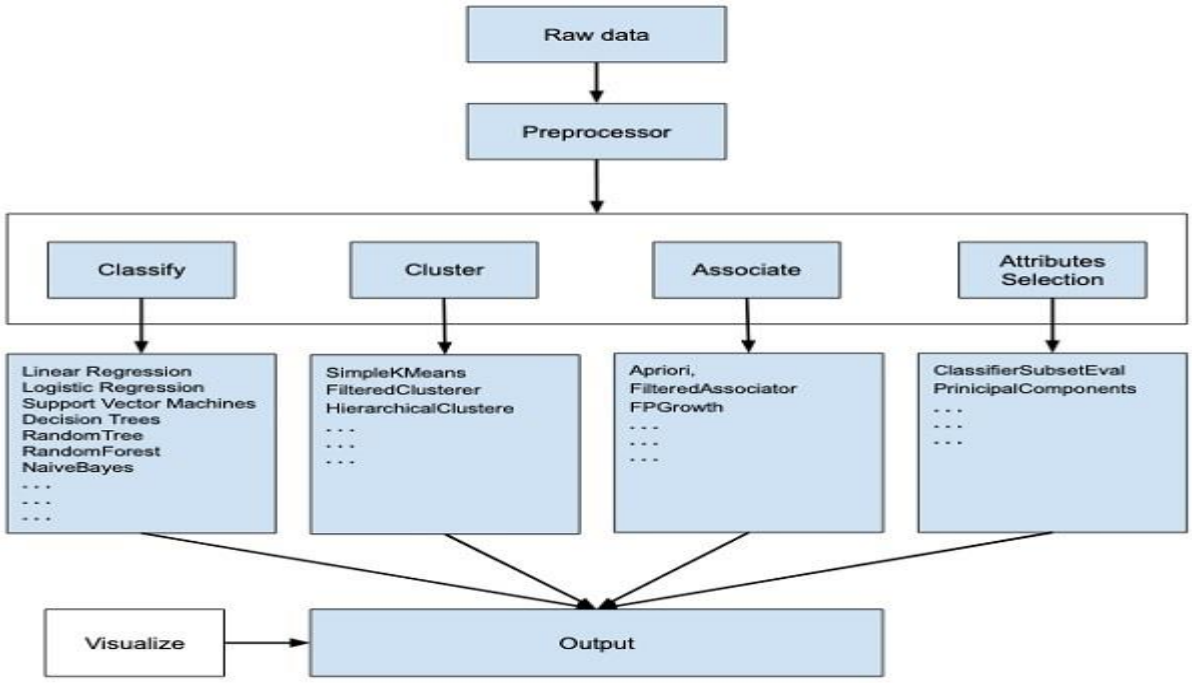
For example, the class value of the Iris dataset can be defined as follows:

```
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Values that contain spaces must be quoted.

About WEKA Software:

WEKA - an opensource software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram −



If you observe the beginning of the flow of the image, you will understand that there are many stages in dealing with Big Data to make it suitable for machine learning −

First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

Then, you would save the preprocessed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify, Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

Now that we have seen what WEKA is and what it does, in the next chapter let us learn how to install WEKA on your local computer

About Attributes:

Different types of attributes

1. **Nominal Attribute:**
   Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Student Roll No., Sex of the Person.

2. **Ordinal Attribute:**
   The ordinal attribute value provides sufficient information to order the objects. Such as Rankings, Grades, Height
3. **Binary Attribute:**
   These are 0 and 1. Where 0 is the absence of any features and 1 is the inclusion of any characteristics.
4. **Numeric attribute:**It is quantitative, such that quantity can be measured and represented in integer or real values ,are of two types
   Interval Scaled attribute:
   It is measured on a scale of equal size units,these attributes allows us to compare such as temperature in C or F and thus values of attributes have order.

5. **Ratio Scaled attribute:**
   Both differences and ratios are significant for Ratio. For eg. age, length, Weight.
6 categorical attribute

An attribute where the values correspond to discrete categories. For example, *state* is a categorical attribute with discrete values (CA, NY, MA, etc.). Categorical attributes are either non-ordered (nominal) like state, gender, etc., or ordered (ordinal) such as high, medium, or low temperatures.

3.3 Steps to identify the different types of Attributes:

3.3.1 steps for identifying categorical attributes

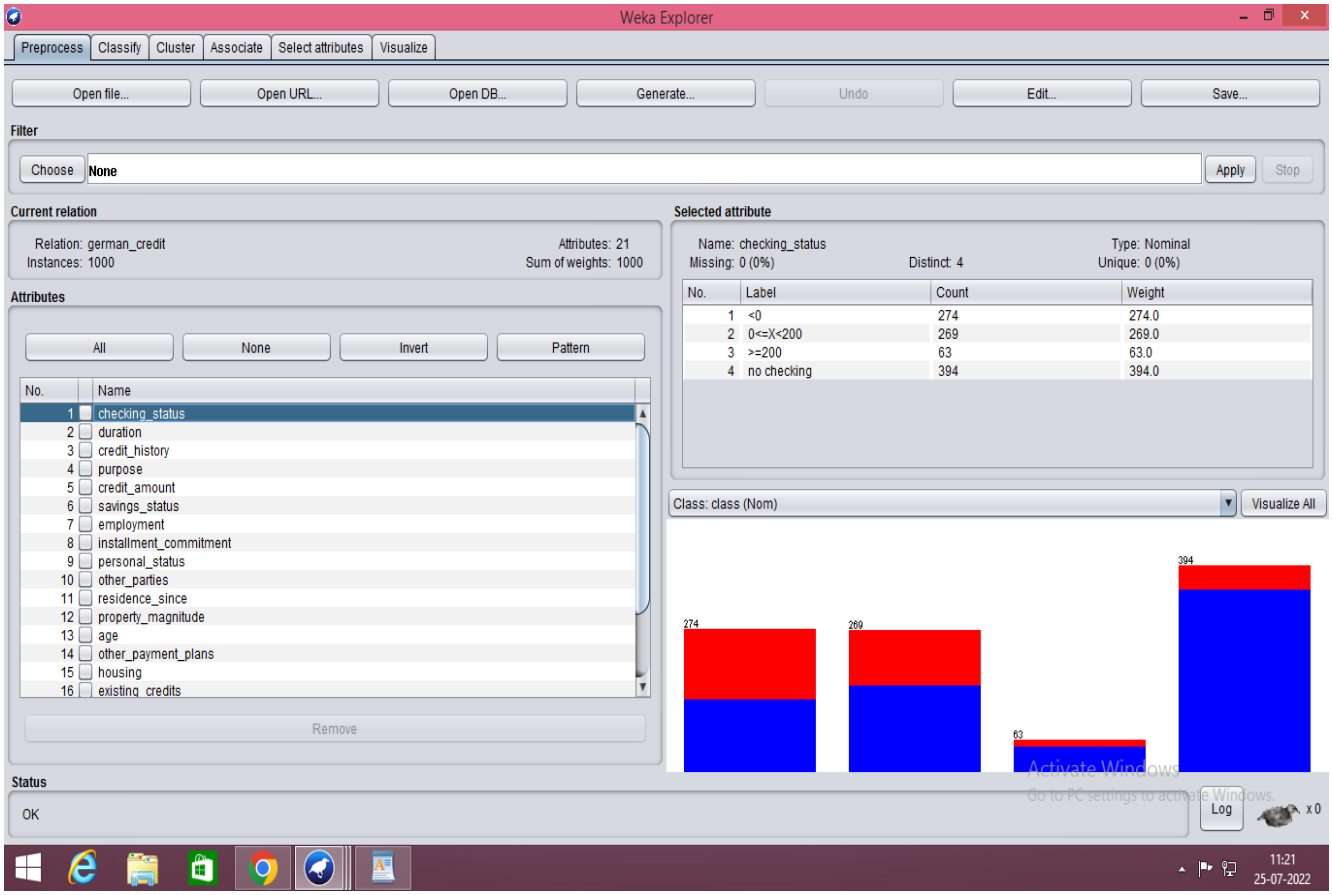
1. Double click on credit-g.arff file.
2. Select all categorical attributes.
3. Click on invert.
4. Then we get all real valued attributes selected
5. Click on remove
6. Click on visualize all.

3.3.2 Steps for identifying real valued attributes

1. Double click on credit-g.arff file.
2. Select all real valued attributes.
2. Click on invert.
3. Then we get all categorial attributes selected
4. Click on remove
5. Click on visualize all.

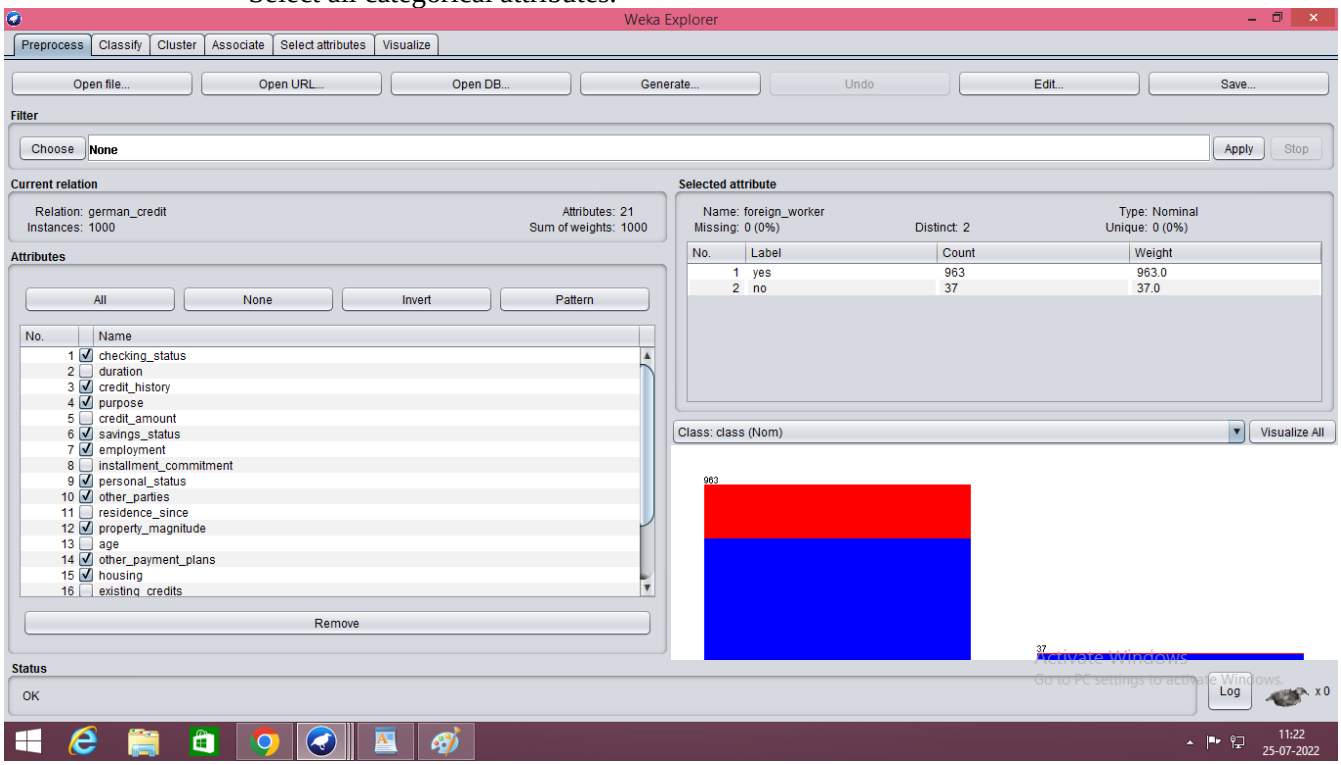
3.4 procedure:

1. To List all the categorical attributes and the real-valued attributes separately in 'German credit' data set first upload the German credit dataset into weka software by choosing open file option and selecting German Credit arff file.

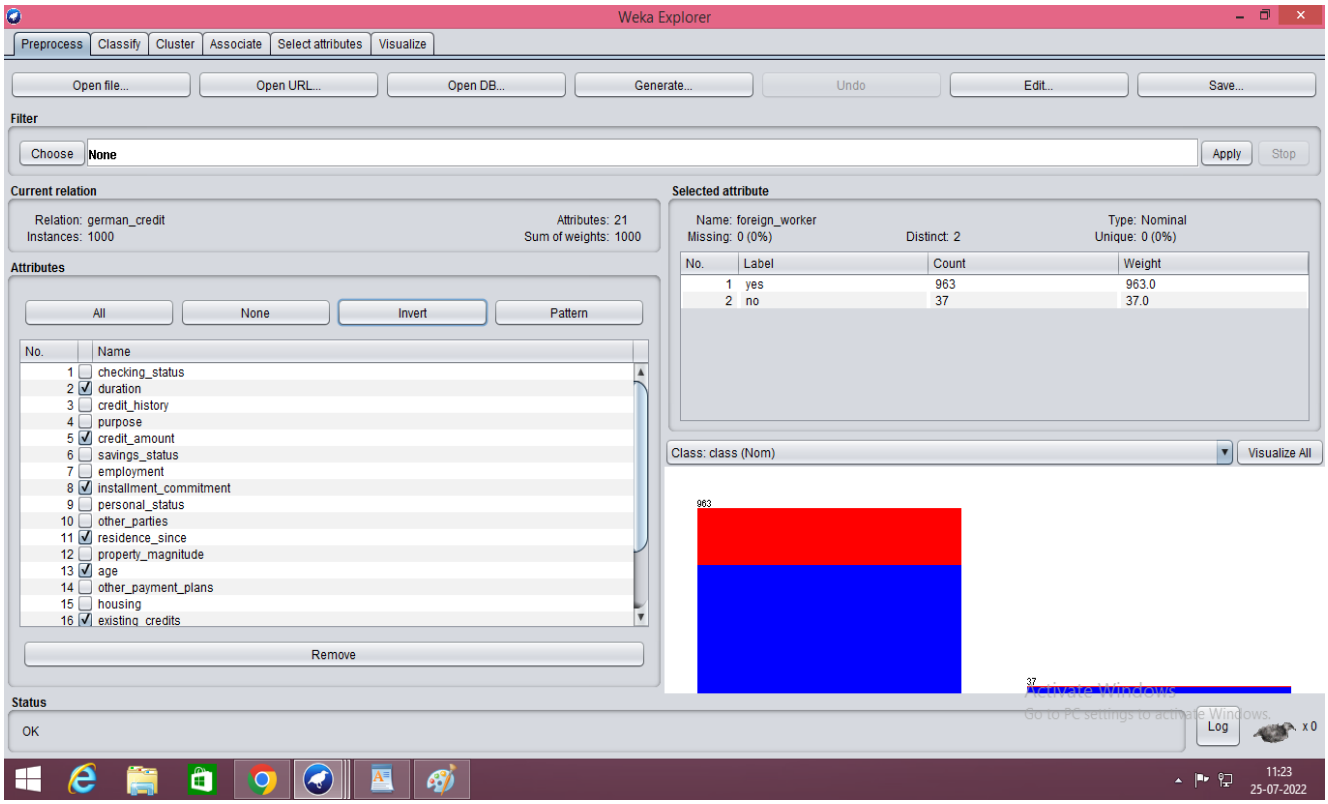


2. Now to identify the categorical attributes

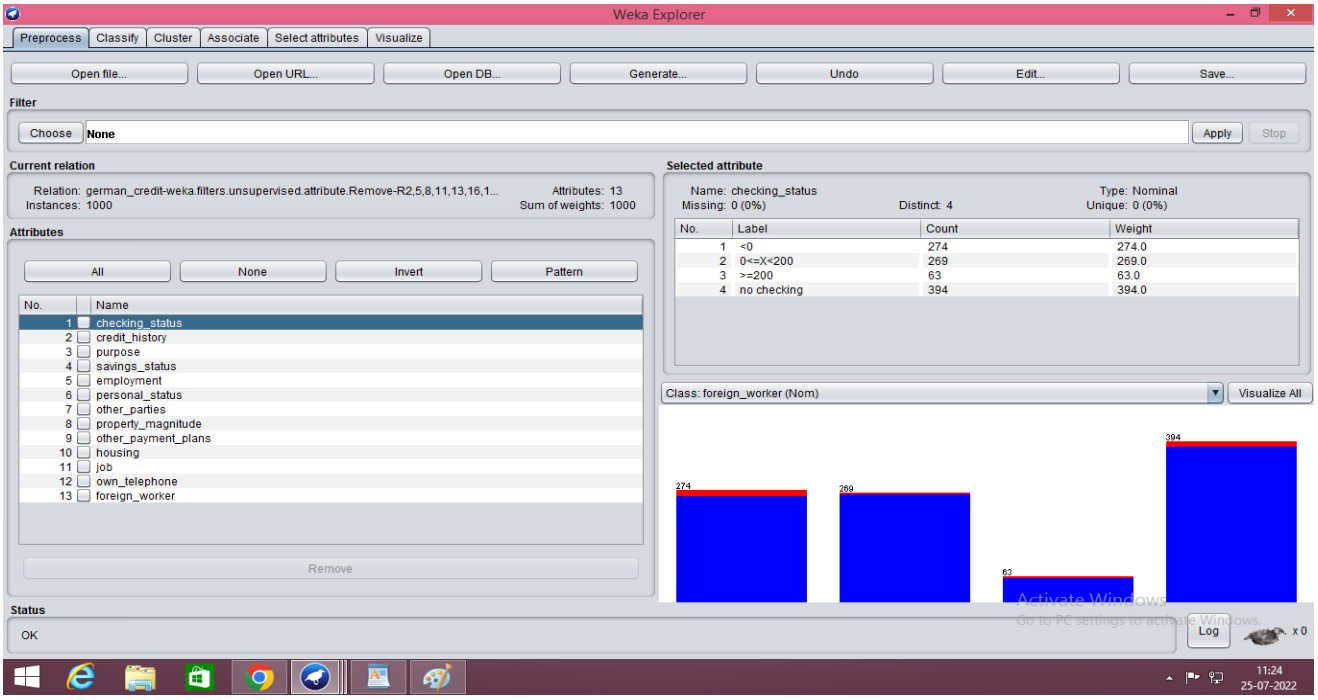
- Double click on credit-g.arff file.
- Select all categorical attributes.



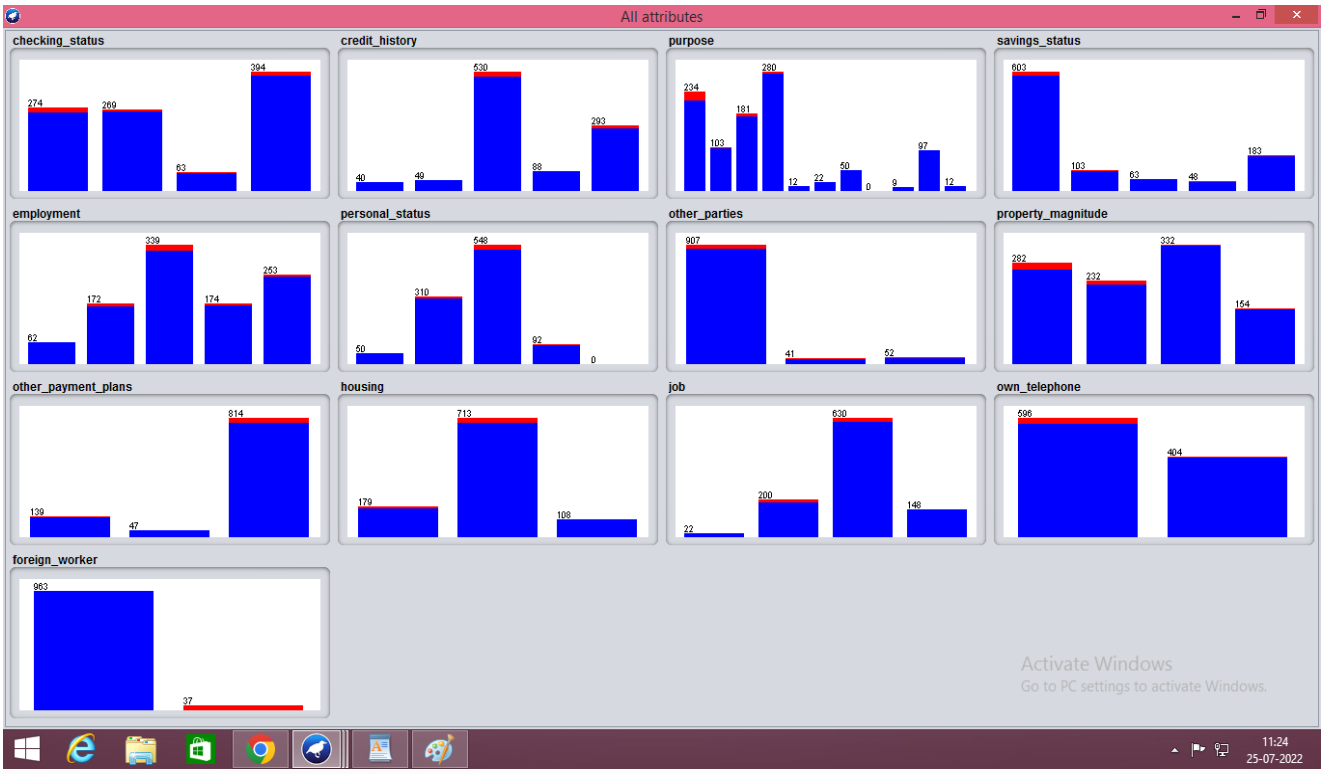
- Click on invert.



- Then we get all real valued attributes selected
- Click on remove

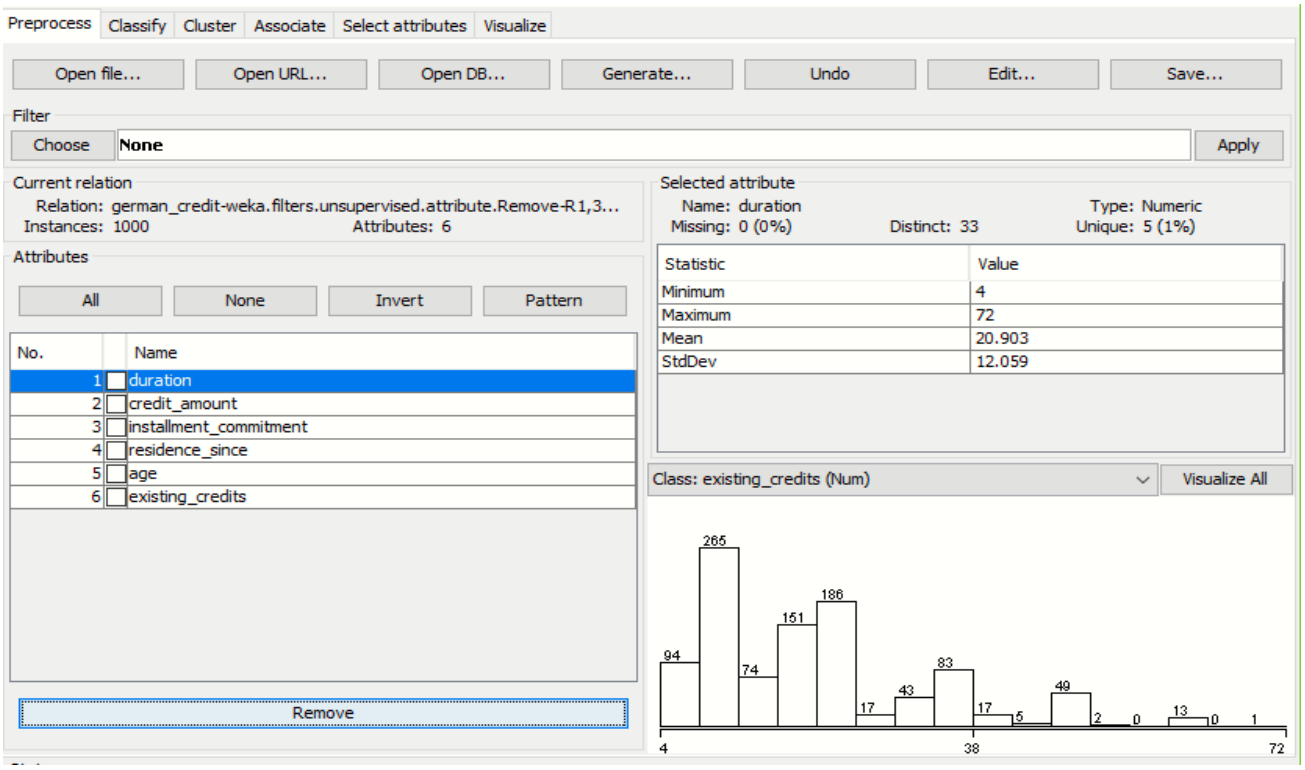


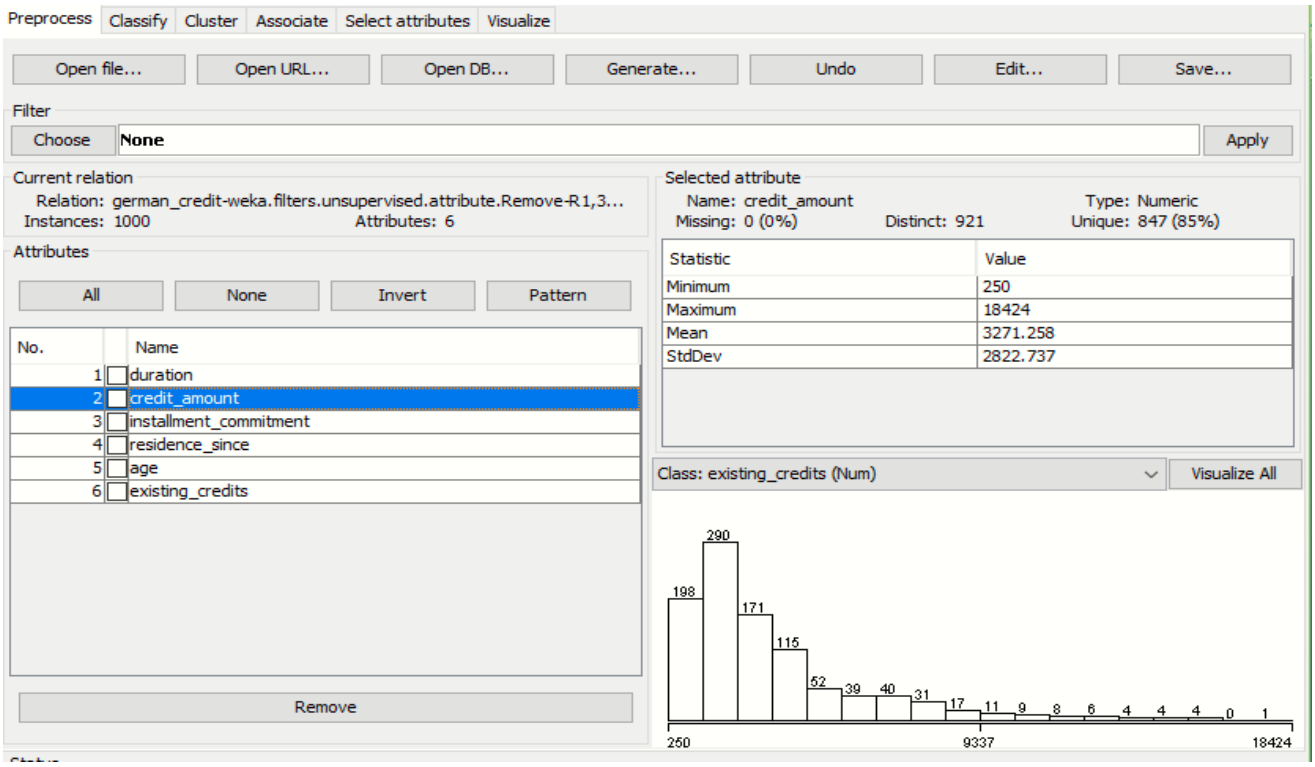
- Click on visualize all.

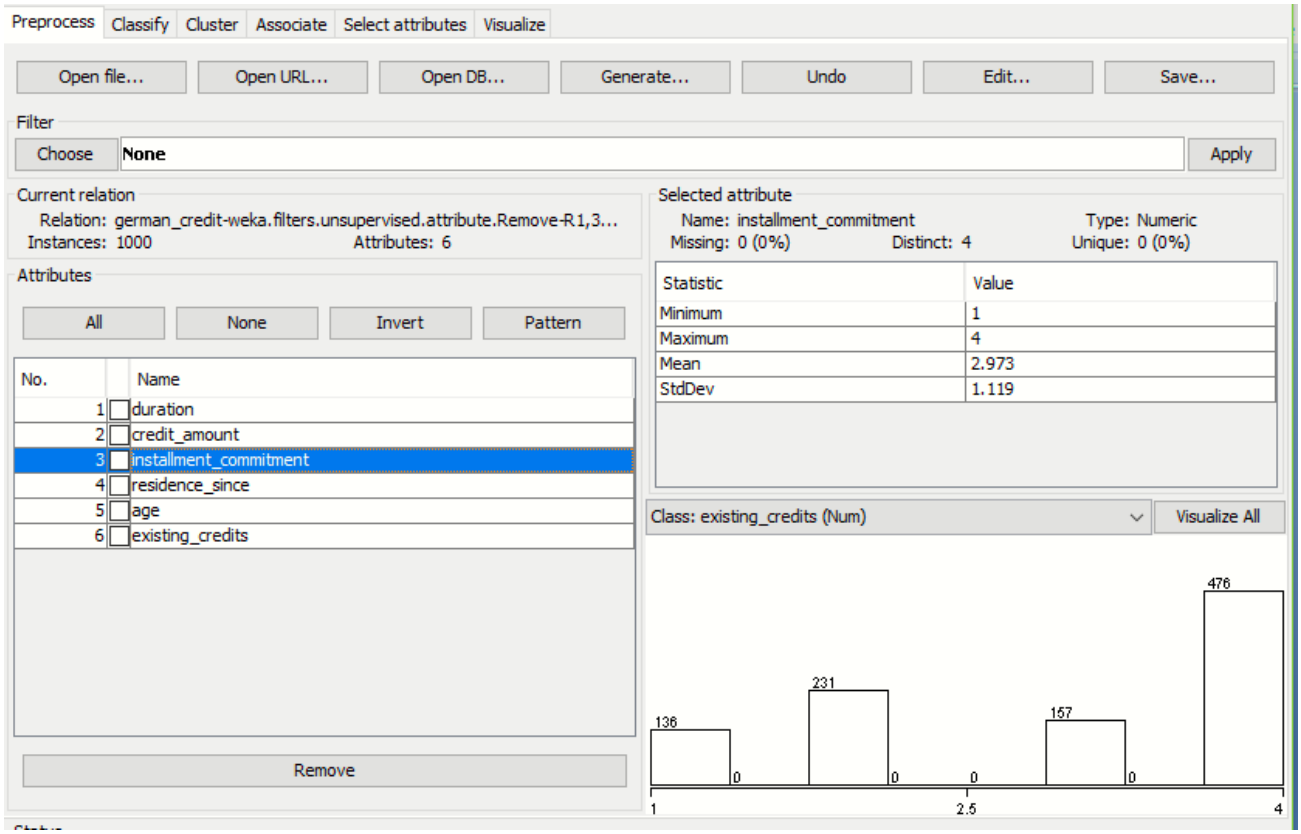


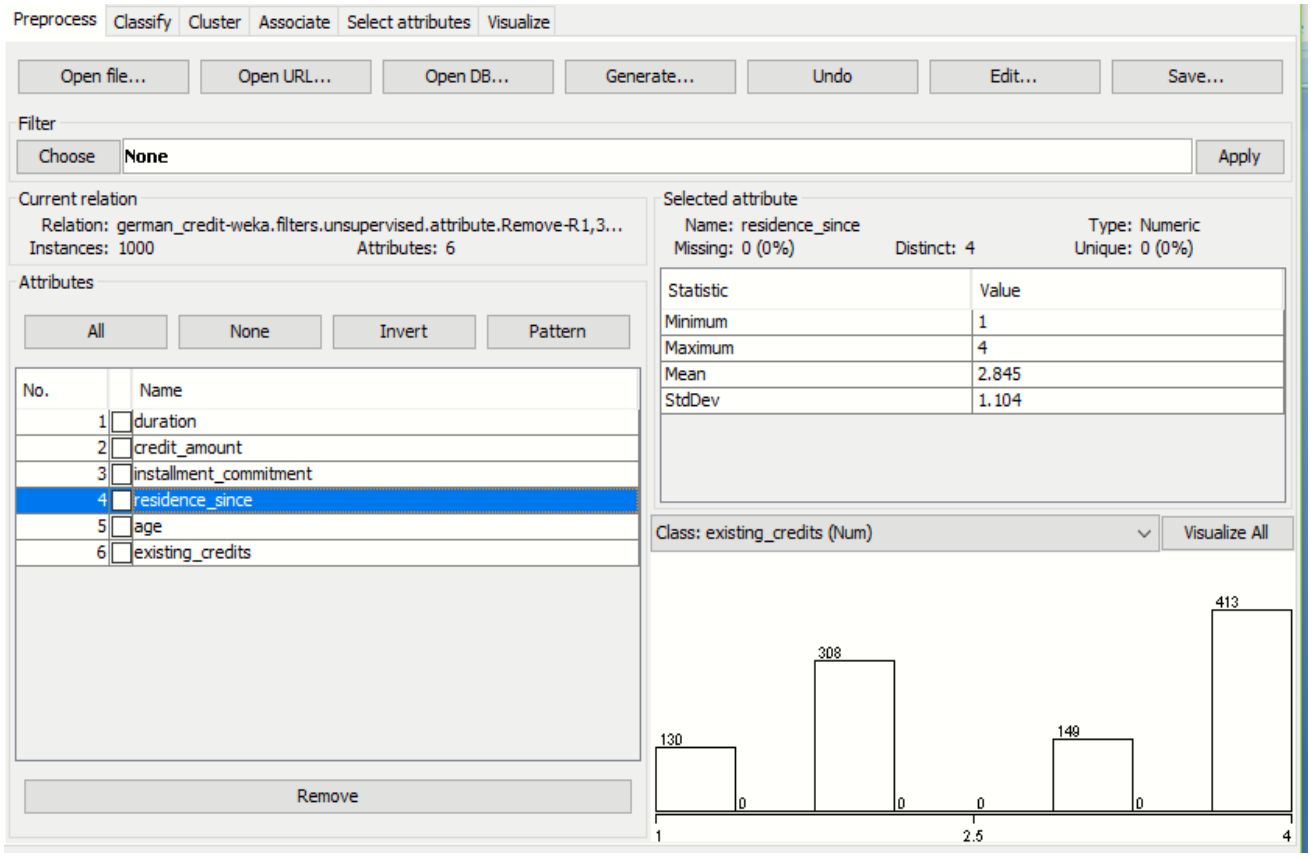
3. Now to identify the real valued attributes repeat the above steps for real values.

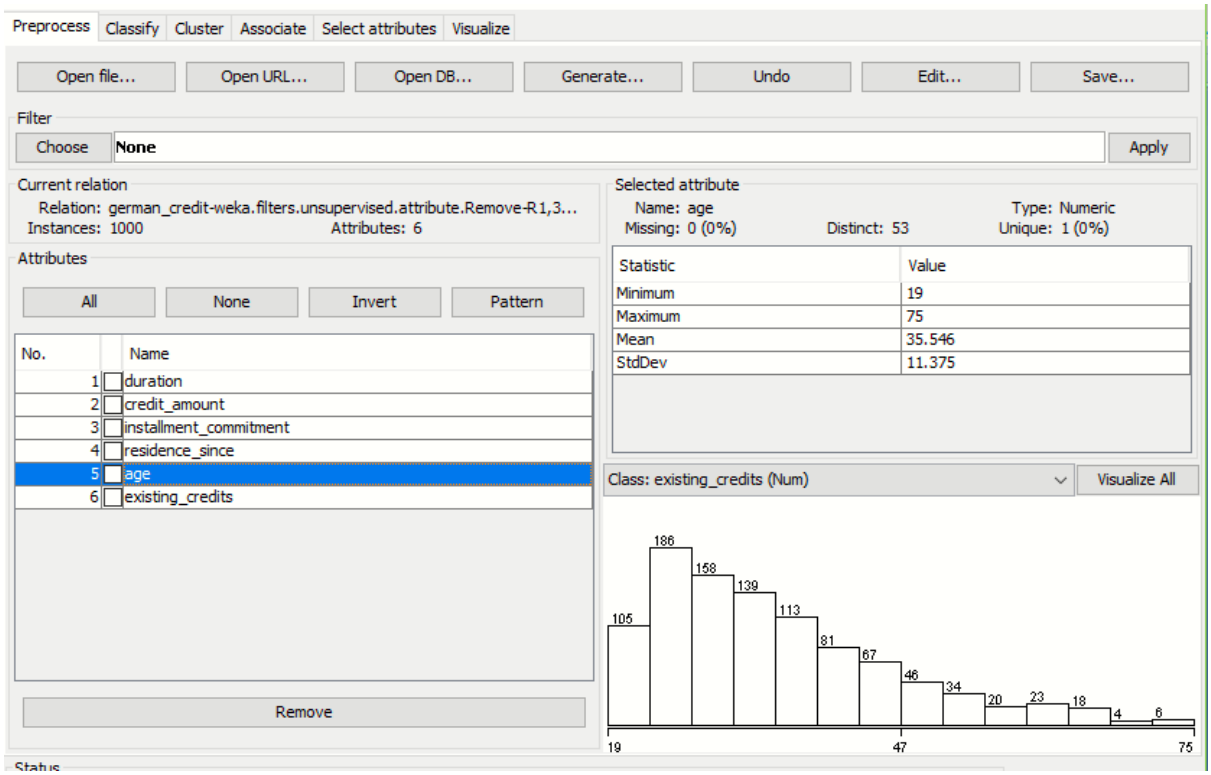
- Double click on credit-g.arff file.
- Select all real valued attributes.
- Click on invert.
- Then we get all categorial attributes selected
- Click on remove
- Visualize each real valued attribute

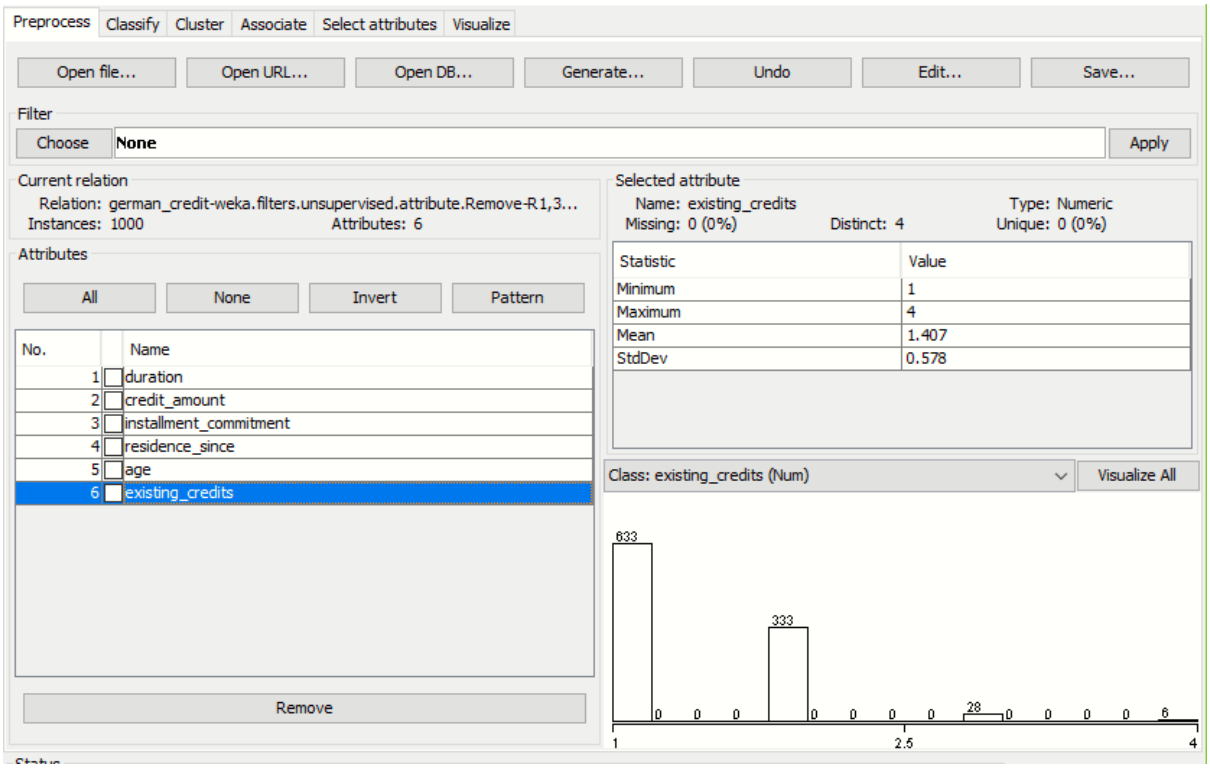




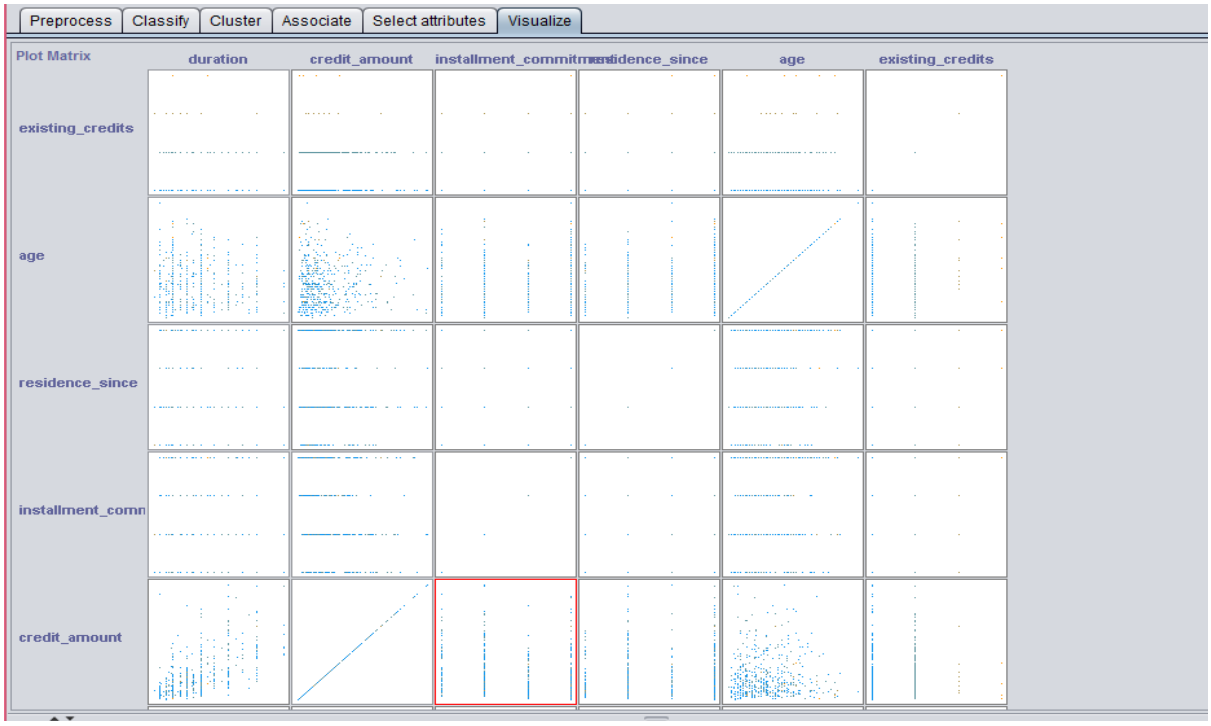

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: german_credit-weka.filters.unsupervised.attribute.Remove-R1,3...
Instances: 1000
Attributes: 6
Selected attribute
Name: installment_commitment
Type: Numeric
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)
Attributes
All
None
Invert
Pattern
No.
Name
1 duration
2 credit_amount
3 installment_commitment
4 residence_since
5 age
6 existing_credits
Statistic
Value
Minimum 1
Maximum 4
Mean 2.973
StdDev 1.119
Class: existing_credits (Num)
Visualize All
476
231
157
136
0
0
0
0
1
2.5
4
Remove


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: german_credit-weka.filters.unsupervised.attribute.Remove-R1,3...
Instances: 1000
Attributes: 6
Selected attribute
Name: residence_since
Type: Numeric
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)
Attributes
All
None
Invert
Pattern
No.
Name
1 duration
2 credit_amount
3 installment_commitment
4 residence_since
5 age
6 existing_credits
Statistic
Value
Minimum 1
Maximum 4
Mean 2.845
StdDev 1.104
Class: existing_credits (Num)
Visualize All
413
308
149
130
0
0
0
0
1
2.5
4
Remove





- Click on visualize all.



3.5 Results and Discussion:

To List all the categorical attributes and the real-valued attributes separately in 'German credit' data set is executed successfully by identifying categorical attributes and the real-valued attributes separately and by visualizing each attribute.