Weka Project Data Preparation Tips

by Data Analytics Tutor Riyad Hussein

Acknowledgement from the instructor: Thanks to Riyad for creating the following notes.

Weka resources:
Working with data in Weka: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab3-WorkingWithDataInWeka.pdf
Classification with WEKA Explorer: http://storm.cis.fordham.edu/~yli/documents/CISC4631Spring16/Preprocess.pdf

Data Preparation
The first and foremost step of data mining process is to understand the data and identify the research question(s). Here are some suggestions to explore and understand datasets:
- Look at the attribute type; e.g., nominal, ordinal or quantitative.

Open the file german_credit..arff in text editor and see the attribute part on top of the file

```
@relation german_credit

@attribute Creditability {0,1}
@attribute 'Account Balance' {1,2,3,4}
@attribute 'Duration of Credit (month)' numeric
@attribute 'Payment Status of Previous Credit' {0,1,2,3,4}
@attribute Purpose {0,1,2,3,4,5,6,8,9,10}
@attribute 'Credit Amount' Numeric
@attribute 'Value Savings/Stocks' {1,2,3,4,5}
@attribute 'Length of current employment' {1,2,3,4,5}
@attribute 'Instalment per cent' real
@attribute 'Sex & Marital Status' {1,2,3,4}
@attribute Guarantors {1,2,3}
@attribute 'Duration in Current address' {1,2,3,4}
@attribute 'Most valuable available asset' {1,2,3,4}
@attribute 'Age (years)' numeric
@attribute 'Concurrent Credits' {1,2,3}
@attribute 'Type of apartment' {1,2,3}
@attribute 'No of Credits at this Bank' numeric
@attribute Occupation {1,2,3,4}
@attribute 'No of dependents' numeric
@attribute Telephone {1,2}
@attribute 'Foreign Worker' {1,2}
```

- Find max, min, mean and standard deviation of attributes.

Use R-Studio:

```
gc <- read.csv("D:/Users/rhusein/Documents/german_credit_card/german_credit.csv",header = T,
stringsAsFactors = F, na.strings = c("","NA"))
str(gc)
summary(gc)
```

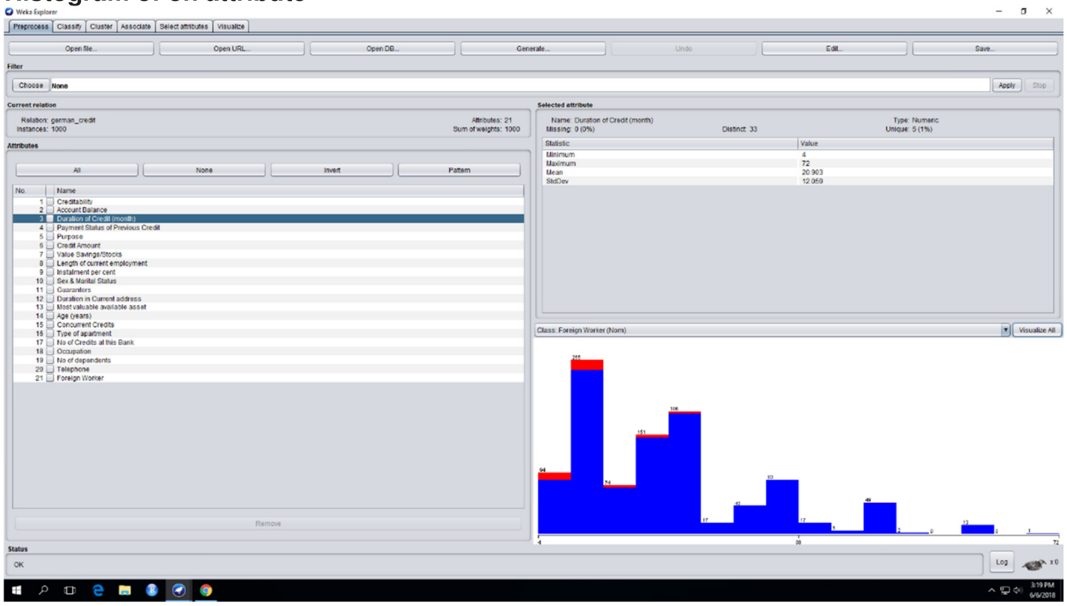
- Determine any outlier values (records) for each of the attributes or attributes under consideration (min, max, std. dev, scatter plots, box plots or others can be used).

Using R-Studio. Use the functions boxplot() and boxplot.stats() to get the outliers. The attribute gc$Credit.Amount as an example

```
boxplot(gc$Credit.Amount)

> boxplot.stats(gc$Credit.Amount)
$`stats`
[1]  250.0 1365.0 2319.5 3972.5 7882.0

$n
[1] 1000

$conf
[1] 2189.219 2449.781

$out
 [1] 10875  8858 12749  8072  8487 12169 10722  8613  8588 10366  8133  9436 10477 13756 11760 14179
10974  9566  8358  9857 10222  9055  7966 12204  8229 10623
[27]  9277 15857 10144 15653  8335  8471  8947 11054  9157  9283 14555  9271  8386 14318 15672 10961  7980 11560
11328 11938 14782 12612  9398  9572  8065  9034
[53] 14027  9629 12976 10297 14421  8086 10127 12389 11590 15945  9960  8648  8318 11816 11998 18424
14896  8978 12579 12680

>
```

- Analyze the distribution of numeric attributes (normal or other). Plot histograms for attributes of concern and analyze whether they have any influence on the class attribute.

Histogram of on attribute



Filter: None

Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

No.	Name
1	Creditability
2	Account Balance
3	Duration of Credit (month)
4	Payment Status of Previous Credit
5	Purpose
6	Credit Amount
7	Value Savings/Stocks
8	Length of current employment
9	Instalment per cent
10	Sex & Marital Status
11	Guarantors
12	Duration in Current address
13	Most valuable available asset
14	Age (years)
15	Concurrent Credits
16	Type of apartment
17	No of Credits at this Bank
18	Occupation
19	No of dependents
20	Telephone
21	Foreign Worker

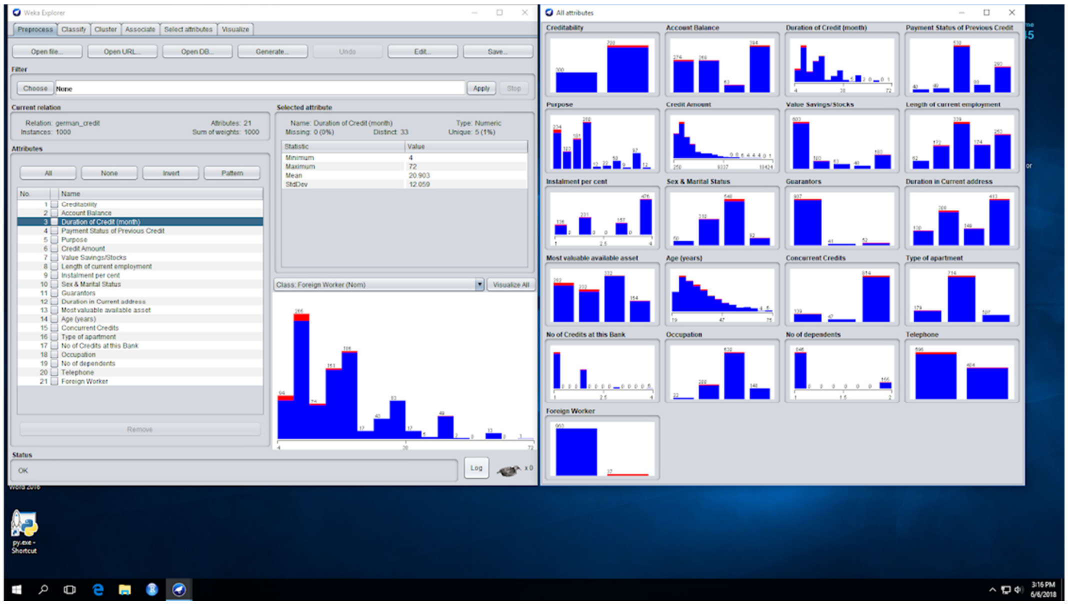
Name: Duration of Credit (month)
Missing: 0 (0%)
Distinct: 33
Type: Numeric
Unique: 5 (1%)

Statistic	Value
Minimum	4
Maximum	72
Mean	20.903
StdDev	12.059

Class: Foreign Worker (Nom)

Status
OK

Histograms/bar-char for all attributes



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Filter: None

Current relation

Relation: german_credit Attributes: 21
Instances: 1000 Sum of weights: 1000

Selected attribute

Name: Duration of Credit (month) Type: Numeric
Missing: 0 (0%) Distinct: 33 Unique: 5 (1%)

Statistic	Value
Minimum	4
Maximum	72
Mean	20.903
StdDev	12.059

Attributes

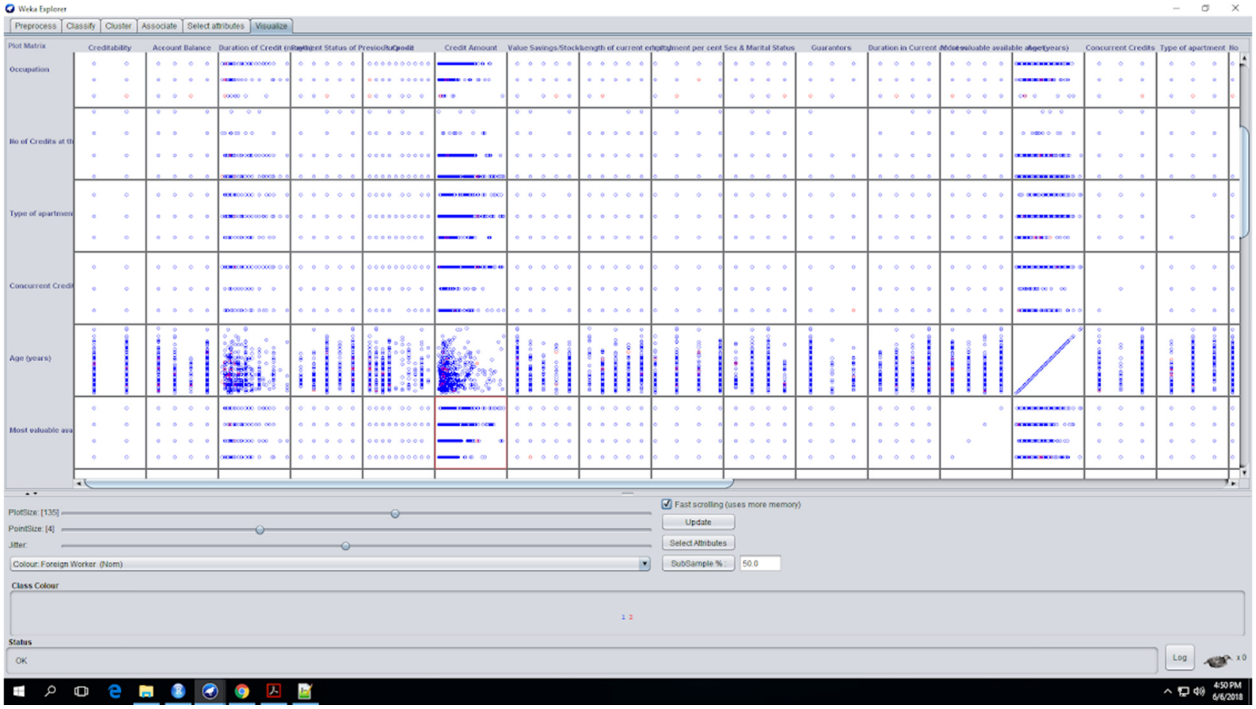
No.	Name
1	Creditability
2	Account Balance
3	Duration of Credit (month)
4	Payment Status of Previous Credit
5	Purpose
6	Credit Amount
7	Value Savings/Stocks
8	Length of current employment
9	Instalment per cent
10	Sex & Marital Status
11	Guarantors
12	Duration in Current address
13	Most valuable available asset
14	Age (years)
15	Concurrent Credits
16	Type of apartment
17	No of Credits at this Bank
18	Occupation
19	No of dependents
20	Telephone
21	Foreign Worker

Class: Foreign Worker (Nom)

Status: OK

- Load the dataset in Weka and click on visualization tab. Which attributes seem to be correlated? Which attributes seem to be most linked to the class attribute?

Use Visualize tab on Weka



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

PlotSize: [135]
PointSize: [4]
Jitter:
Colour: Foreign Worker (Nom)

☑ Fast scrolling (uses more memory)

SubSample %: 50.0

Class Colour

Status: OK

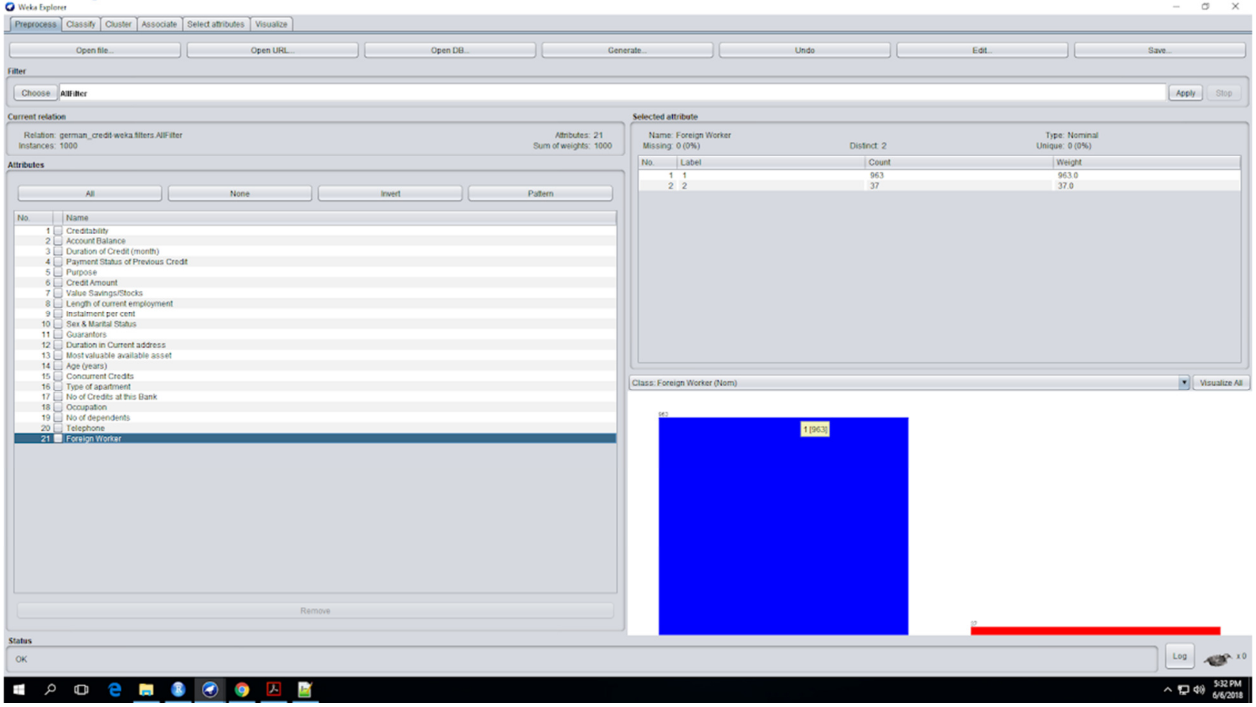
On R-Studio run this command to find correlation between attributes
cor(gc)

- Which attributes do you think can be eliminated or included in the analysis?

How to Perform Feature Selection With Machine Learning Data in Weka: https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/

- Determine whether the dataset has an imbalanced class distribution (same proportion of records of different types or not).

How to handle Imbalanced Classification Problems in machine learning? https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/



Filter: AllFilter

Current relation
Relation: german_credit-weka.filters.AllFilter   Attributes: 21
Instances: 1000   Sum of weights: 1000

Attributes

No.		Name
1	□	Creditability
2	□	Account Balance
3	□	Duration of Credit (month)
4	□	Payment Status of Previous Credit
5	□	Purpose
6	□	Credit Amount
7	□	Value Savings/Stocks
8	□	Length of current employment
9	□	Instalment per cent
10	□	Sex & Marital Status
11	□	Guarantors
12	□	Duration in Current address
13	□	Most valuable available asset
14	□	Age (years)
15	□	Concurrent Credits
16	□	Type of apartment
17	□	No of Credits at this Bank
18	□	Occupation
19	□	No of dependents
20	□	Telephone
21	□	Foreign Worker

Selected attribute
Name: Foreign Worker   Type: Nominal
Missing: 0 (0%)   Distinct: 2   Unique: 0 (0%)

No.	Label	Count	Weight
1	1	963	963.0
2	2	37	37.0

Class: Foreign Worker (Nom)

Status
OK