Experiment No.8
Pre-processing using WEKA tool
Date of Performance:
Date of Submission:

Aim: To implement Pre-processing using WEKA tool

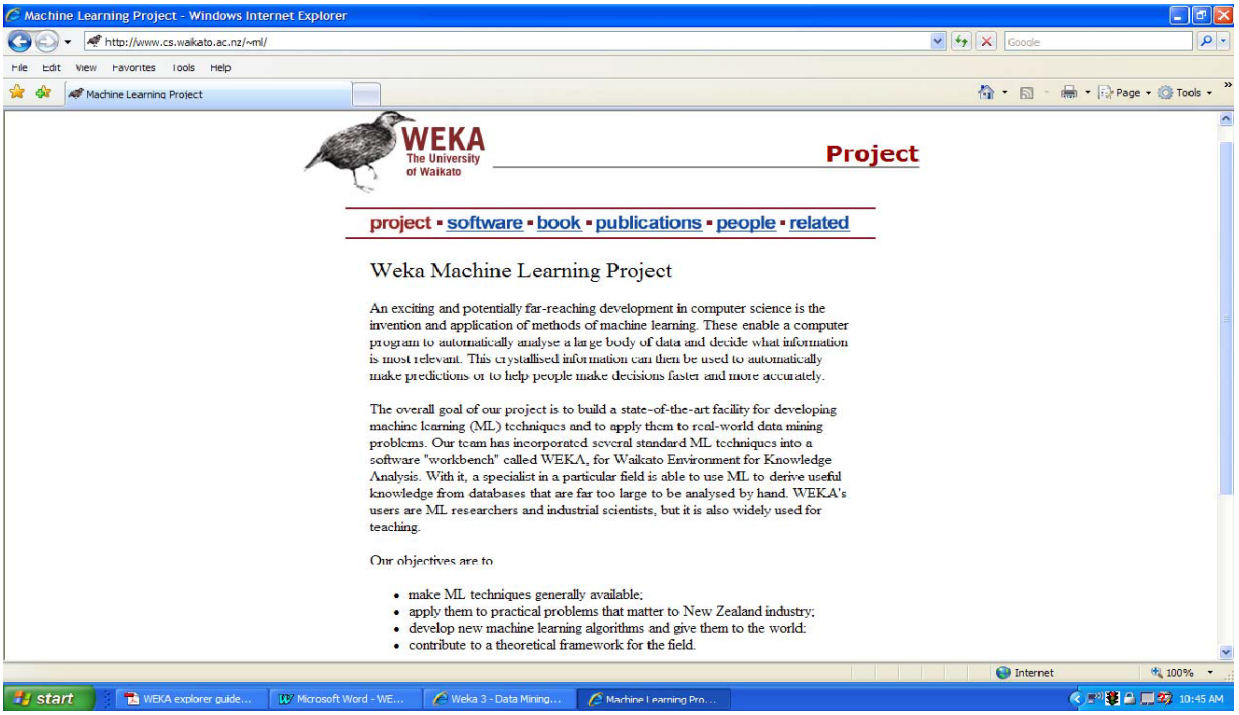
Objective: Pre-processing using WEKA tool

Theory:

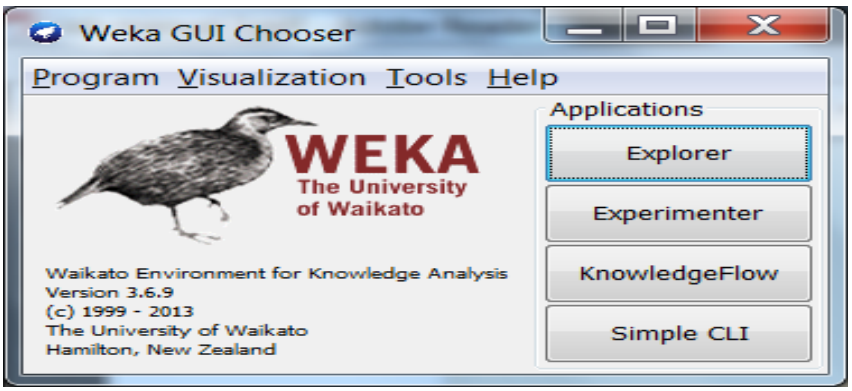
WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

Installation

The program information can be found by conducting a search on the Web for WEKA Data Mining . The site has a very large amount of useful information on the program's benefits and background. New users might find some benefit from investigating the user manual for the program. The main WEKA site has links to this information as well as past experiments for new users to refine the potential uses that might be of particular interest to them. When prepared to download the software it is best to select the latest application from the selection offered on the site. The format for downloading the application is offered in a self installation package and is a simple procedure that provides the complete program on the end users machine that is ready to use when extracted.



The Weka GUI Chooser (class weka.gui.GUIChooser) provides a starting point for launching Weka's main GUI applications and supporting tools. If one prefers a MDI ("multiple document interface") appearance, then this is provided by an alternative launcher called "Main" (class weka.gui.Main).
The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and four menus.



The buttons can be used to start the following applications:

• Explorer An environment for exploring data with WEKA (the rest of this documentation deals with its application in more detail).

• Experimenter An environment for performing experiments and conducting statistical tests between learning schemes.

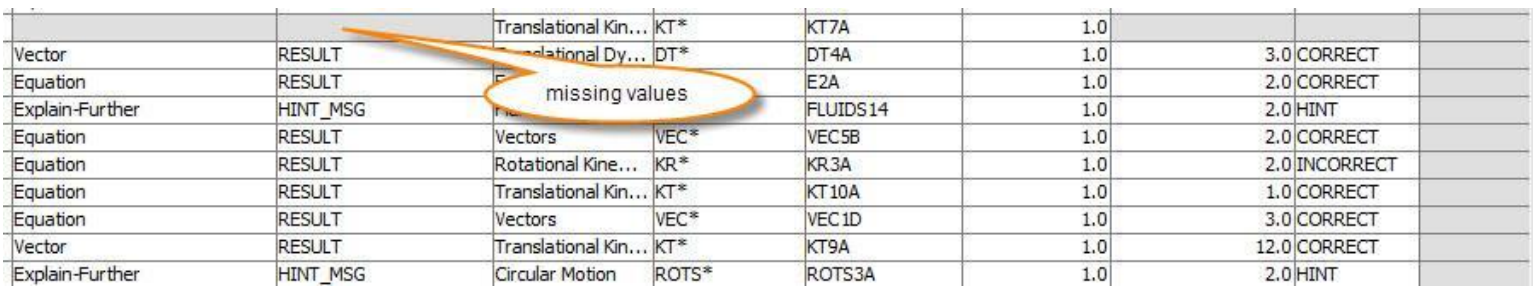
• KnowledgeFlow This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.

• SimpleCLI Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

Step 1: Data Pre Processing or Cleaning

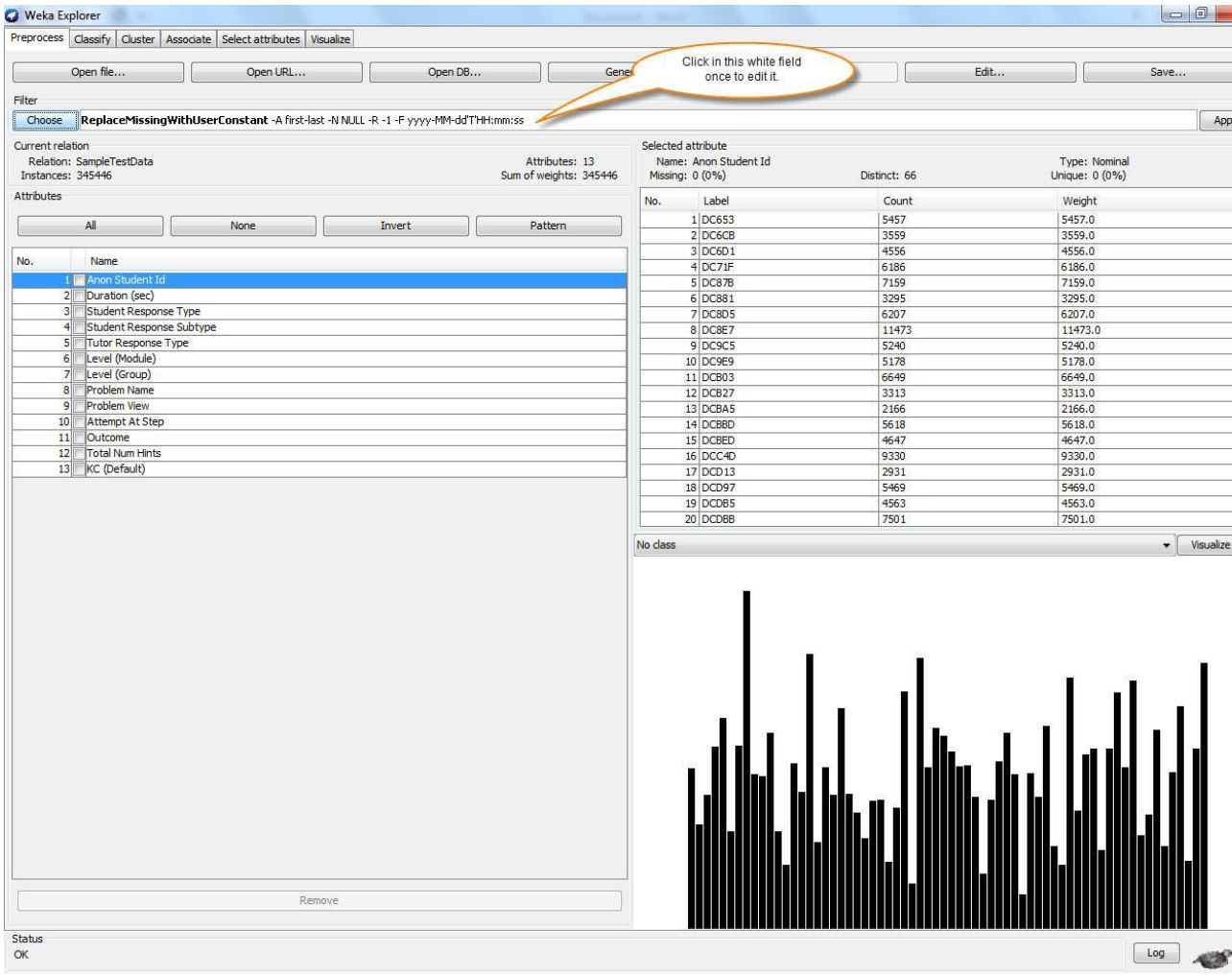
1. Launch Weka-> click on the tab Explorer
2. Load a dataset. (Click on "Open File" & locate the datafile)
3. Click on PreProcess tab & then look at your lower R.H.S. bottom window click on drop down arrow and choose "No Class"
4. Click on "Edit" tab, a new window opens up that will show you the loaded datafile. By looking at your dataset you can also find out if there are missing values in it or not. Also please note the attribute types on the column header. It would either be 'nominal' or 'numeric'.

a. If your data has missing values then its best to clean it first before you apply any forms of mining algorithm to it. In the below figure , you will see the highlighted fields are blank that means the data at hand is dirty and it first needs to be cleaned.

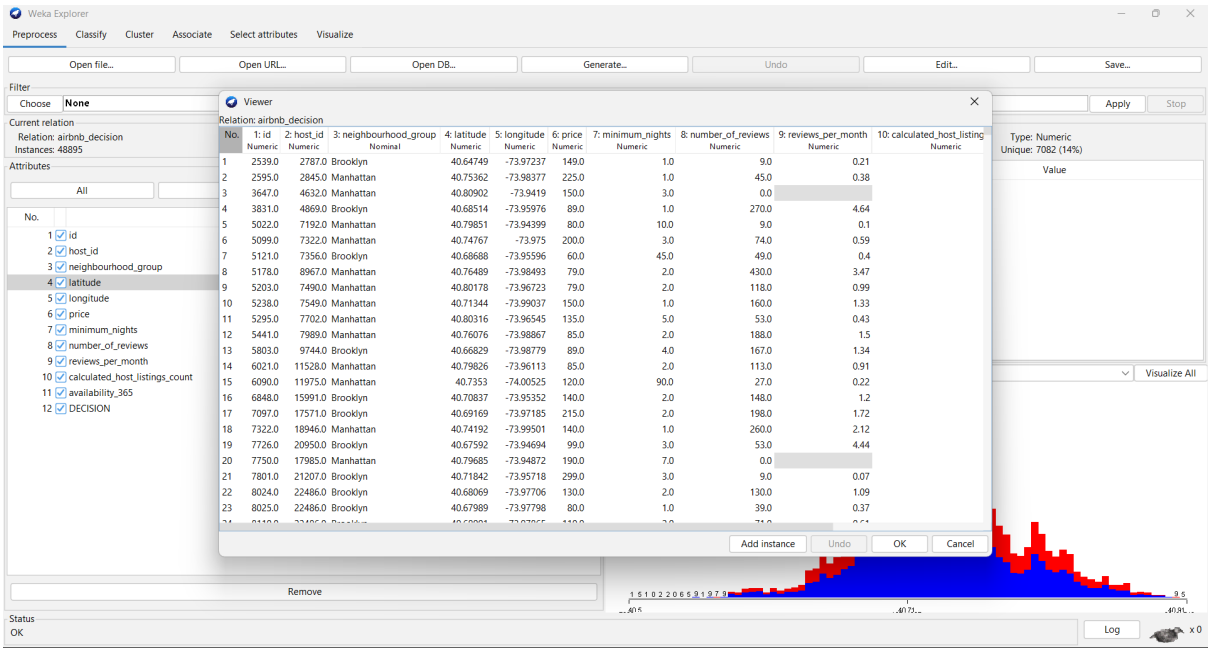


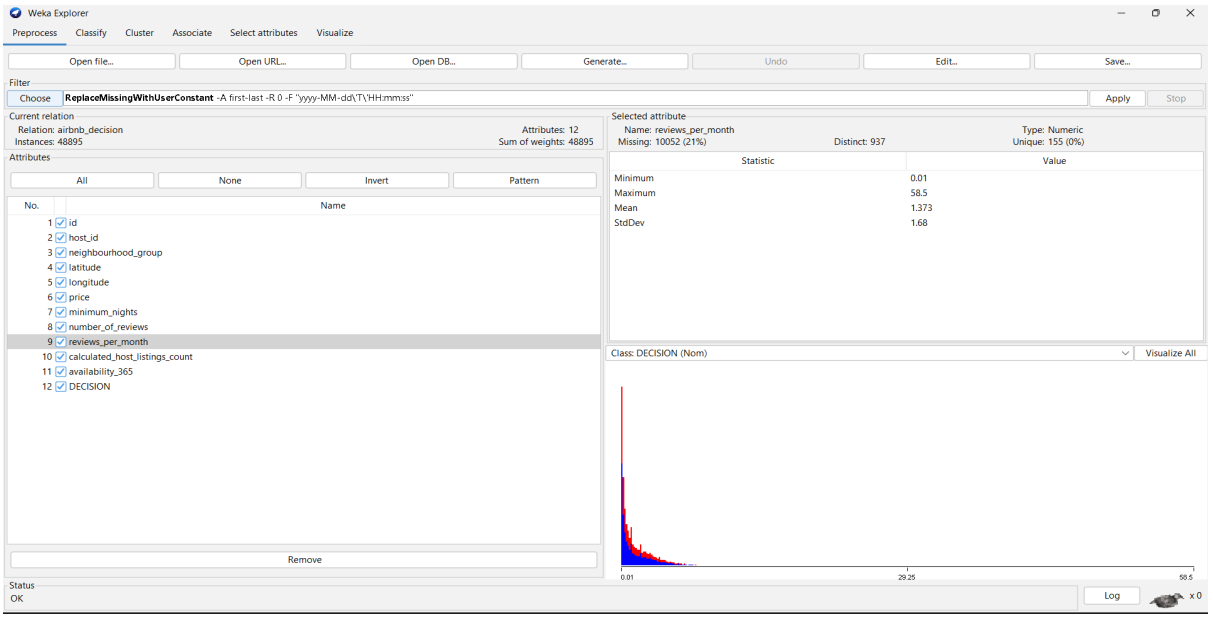
		Translational Kin...	KT*	KT7A	1.0		
Vector	RESULT	[illegible]lational Dy...	DT*	DT4A	1.0	3.0	CORRECT
Equation	RESULT	[illegible]	[illegible]	E2A	1.0	2.0	CORRECT
Explain-Further	HINT_MSG	[illegible]	[illegible]	FLUIDS14	1.0	2.0	HINT
Equation	RESULT	Vectors	VEC*	VEC5B	1.0	2.0	CORRECT
Equation	RESULT	Rotational Kine...	KR*	KR3A	1.0	2.0	INCORRECT
Equation	RESULT	Translational Kin...	KT*	KT10A	1.0	1.0	CORRECT
Equation	RESULT	Vectors	VEC*	VEC1D	1.0	3.0	CORRECT
Vector	RESULT	Translational Kin...	KT*	KT9A	1.0	12.0	CORRECT
Explain-Further	HINT_MSG	Circular Motion	ROTS*	ROTS3A	1.0	2.0	HINT

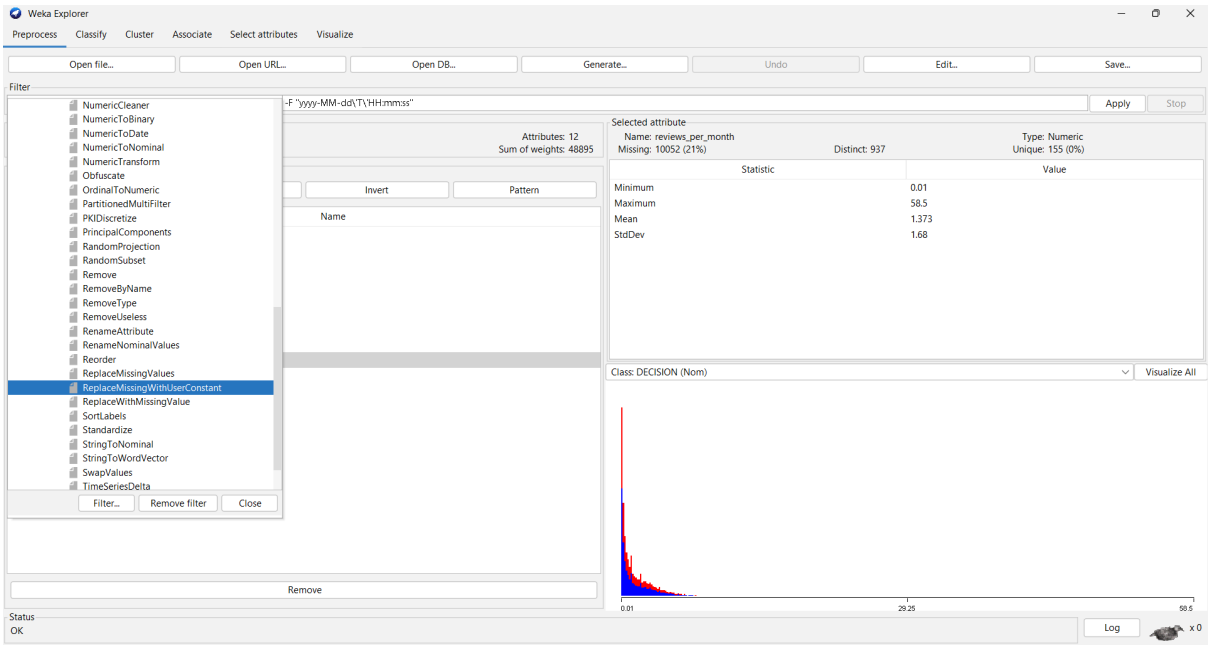
b. Data Cleaning: To clean the data, you apply "Filters" to it. Generally the data will be missing with values, so the filter to apply is "ReplaceMissingWithUserConstant" (the filter choice may vary according to your need, for more information on it please consult the resources).Click on Choose button below Filters-> Unsupervised->attribute————————> ReplaceMissingWithUserConstant

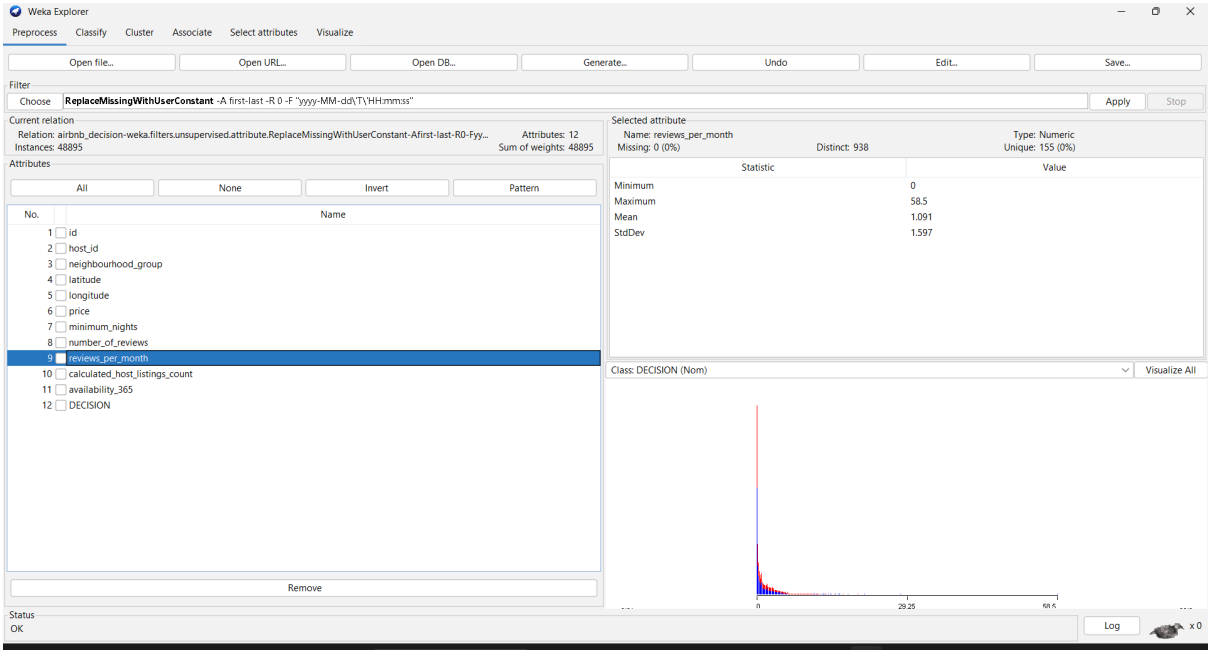


Code and output:




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
ReplaceMissingWithUserConstant -A first-last -R 0 -F "yyyy-MM-dd\T\HH:mm:ss"
Apply
Stop
Current relation
Relation: airbnb_decision
Instances: 48895
Attributes: 12
Sum of weights: 48895
Attributes
All
None
Invert
Pattern
No.
Name
1 id
2 host_id
3 neighbourhood_group
4 latitude
5 longitude
6 price
7 minimum_nights
8 number_of_reviews
9 reviews_per_month
10 calculated_host_listings_count
11 availability_365
12 DECISION
Remove
Selected attribute
Name: reviews_per_month
Missing: 10052 (21%)
Distinct: 937
Type: Numeric
Unique: 155 (0%)
Statistic
Value
Minimum 0.01
Maximum 58.5
Mean 1.373
StdDev 1.68
Class: DECISION (Nom)
Visualize All
0.01
29.25
58.5
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
-F "yyyy-MM-dd\T\HH:mm:ss"
Apply
Stop
NumericCleaner
NumericToBinary
NumericToDate
NumericToNominal
NumericTransform
Obfuscate
OrdinalToNumeric
PartitionedMultiFilter
PKIDiscretize
PrincipalComponents
RandomProjection
RandomSubset
Remove
RemoveByName
RemoveType
RemoveUseless
RenameAttribute
RenameNominalValues
Reorder
ReplaceMissingValues
ReplaceMissingWithUserConstant
ReplaceWithMissingValue
SortLabels
Standardize
StringToNominal
StringToWordVector
SwapValues
TimeSeriesDelta
Filter...
Remove filter
Close
Invert
Pattern
Name
Remove
Attributes: 12
Sum of weights: 48895
Selected attribute
Name: reviews_per_month
Missing: 10052 (21%)
Distinct: 937
Type: Numeric
Unique: 155 (0%)
Statistic
Value
Minimum 0.01
Maximum 58.5
Mean 1.373
StdDev 1.68
Class: DECISION (Nom)
Visualize All
0.01
29.25
58.5
Status
OK
Log
x 0



Conclusion:In conclusion, WEKA stands out as a highly valuable tool in the realm of data mining and machine learning. Its user-friendly interface makes it accessible to users with diverse skill levels, while the extensive range of algorithms and robust data preprocessing capabilities cater to various analytical needs. The generated output, rich in evaluations, metrics, and visualizations, provides a comprehensive understanding of model performance. WEKA's openness, encouraging collaboration and customization, further enhances its utility. Whether for educational purposes or real-world applications, WEKA's versatility makes it a powerful asset for users seeking effective and efficient solutions in the realm of machine learning.