Lab Number: 1

Title

Preprocessing of Primary and Secondary Datasets Containing Dirty Data

Objective

To understand and practically perform data preprocessing on a dirty dataset using Weka Explorer.

IDE/Tools Used

Weka 3.8.6

Theory

Dirty Data: A dirty dataset refers to a collection of data that contains inaccuracies, inconsistencies, and errors, which can compromise its usefulness and reliability for analysis, reporting, or decision-making

Types of Dirty Data

- Missing Data
- Duplicate Records
- Inconsistent Values
- Noise
- Outliers

Data Preprocessing: Data preprocessing is the process of cleaning, transforming, and organizing raw data into a structured format that is ready for analysis or use in machine learning models.

- **Cleaning:** Involves handling missing values, removing duplicates, and correcting errors to make the data accurate and consistent.
- **Transformation:** Involves converting data into a suitable format. Examples include standardizing numerical features, normalizing data, or encoding categorical variables.
- **Integration:** Combines data from multiple sources into a single, unified dataset for analysis.

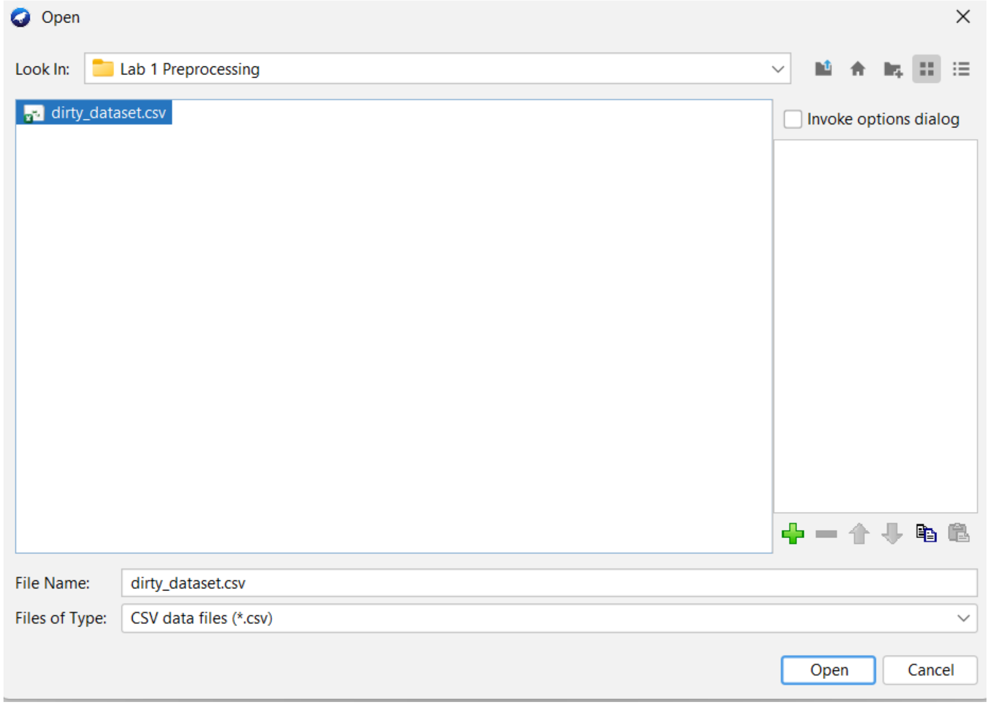
Implementation

A. For Primary Dataset

For primary dataset, a customer churn data was generated.

Steps used to clean the data:

1. Open the dataset in the pre-processor of WEKA

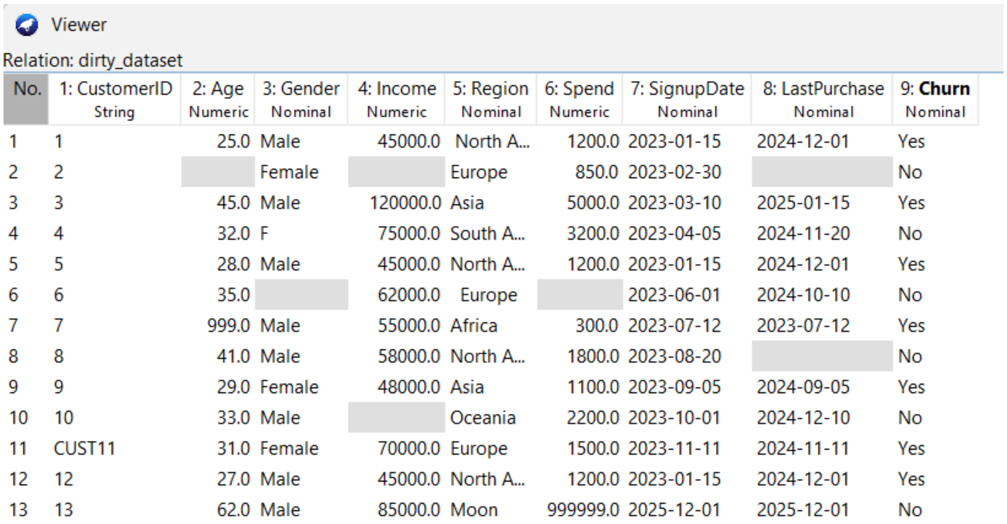


Look In: Lab 1 Preprocessing

File Name: dirty_dataset.csv

Files of Type: CSV data files (*.csv)

2. Visualize the data

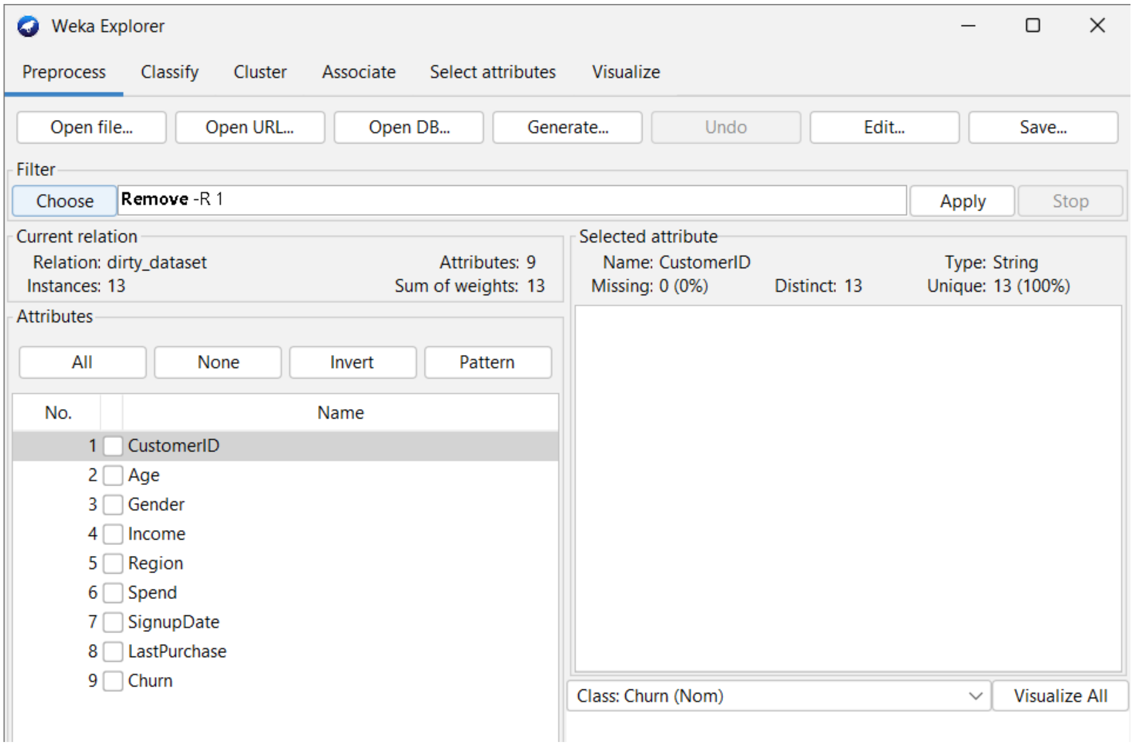


Relation: dirty_dataset

No.	1: CustomerID	2: Age	3: Gender	4: Income	5: Region	6: Spend	7: SignupDate	8: LastPurchase	9: Churn
	String	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Nominal
1	1	25.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	Yes
2	2		Female		Europe	850.0	2023-02-30		No
3	3	45.0	Male	120000.0	Asia	5000.0	2023-03-10	2025-01-15	Yes
4	4	32.0	F	75000.0	South A...	3200.0	2023-04-05	2024-11-20	No
5	5	28.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	Yes
6	6	35.0		62000.0	Europe		2023-06-01	2024-10-10	No
7	7	999.0	Male	55000.0	Africa	300.0	2023-07-12	2023-07-12	Yes
8	8	41.0	Male	58000.0	North A...	1800.0	2023-08-20		No
9	9	29.0	Female	48000.0	Asia	1100.0	2023-09-05	2024-09-05	Yes
10	10	33.0	Male		Oceania	2200.0	2023-10-01	2024-12-10	No
11	CUST11	31.0	Female	70000.0	Europe	1500.0	2023-11-11	2024-11-11	Yes
12	12	27.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	Yes
13	13	62.0	Male	85000.0	Moon	999999.0	2025-12-01	2025-12-01	No

3. Remove unwanted columns

In this case CustomerID was removed using the unsupervised.attribute.Remove filter



Filter: **Remove** -R 1

Current relation

Relation: dirty_dataset Attributes: 9

Instances: 13 Sum of weights: 13

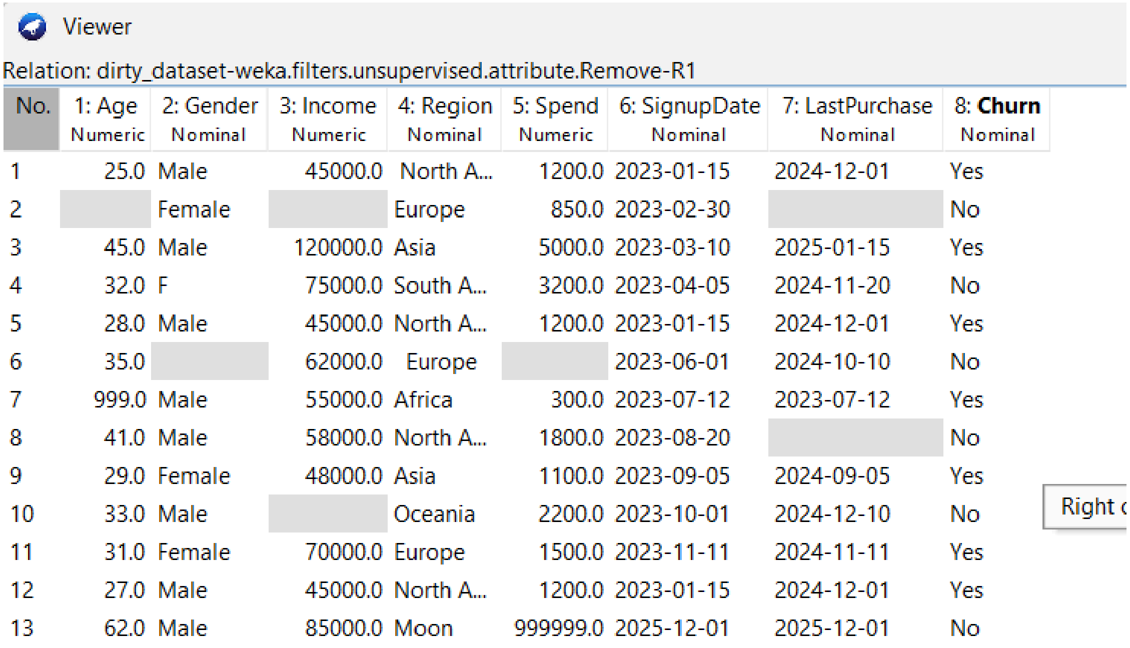
Selected attribute

Name: CustomerID Type: String

Missing: 0 (0%) Distinct: 13 Unique: 13 (100%)

No.		Name
1	□	CustomerID
2	□	Age
3	□	Gender
4	□	Income
5	□	Region
6	□	Spend
7	□	SignupDate
8	□	LastPurchase
9	□	Churn

Class: Churn (Nom)

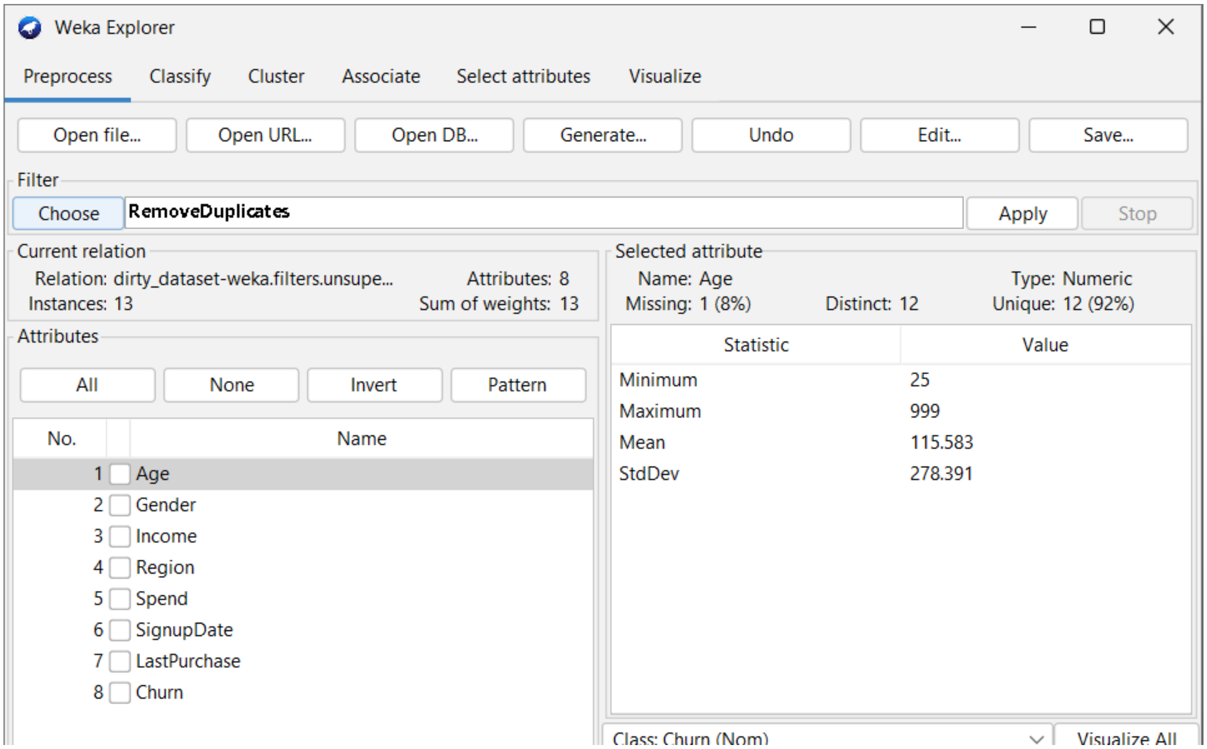


Relation: dirty_dataset-weka.filters.unsupervised.attribute.Remove-R1

No.	1: Age Numeric	2: Gender Nominal	3: Income Numeric	4: Region Nominal	5: Spend Numeric	6: SignupDate Nominal	7: LastPurchase Nominal	8: **Churn** Nominal
1	25.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	Yes
2		Female		Europe	850.0	2023-02-30		No
3	45.0	Male	120000.0	Asia	5000.0	2023-03-10	2025-01-15	Yes
4	32.0	F	75000.0	South A...	3200.0	2023-04-05	2024-11-20	No
5	28.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	Yes
6	35.0		62000.0	Europe		2023-06-01	2024-10-10	No
7	999.0	Male	55000.0	Africa	300.0	2023-07-12	2023-07-12	Yes
8	41.0	Male	58000.0	North A...	1800.0	2023-08-20		No
9	29.0	Female	48000.0	Asia	1100.0	2023-09-05	2024-09-05	Yes
10	33.0	Male		Oceania	2200.0	2023-10-01	2024-12-10	No
11	31.0	Female	70000.0	Europe	1500.0	2023-11-11	2024-11-11	Yes
12	27.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	Yes
13	62.0	Male	85000.0	Moon	999999.0	2025-12-01	2025-12-01	No

4. Remove any duplicate values

In this case this was done using the unsupervised.instance.RemoveDuplicates filter.



Filter: RemoveDuplicates

Current relation
Relation: dirty_dataset-weka.filters.unsupe... Attributes: 8
Instances: 13 Sum of weights: 13

Selected attribute
Name: Age Type: Numeric
Missing: 1 (8%) Distinct: 12 Unique: 12 (92%)

Statistic	Value
Minimum	25
Maximum	999
Mean	115.583
StdDev	278.391

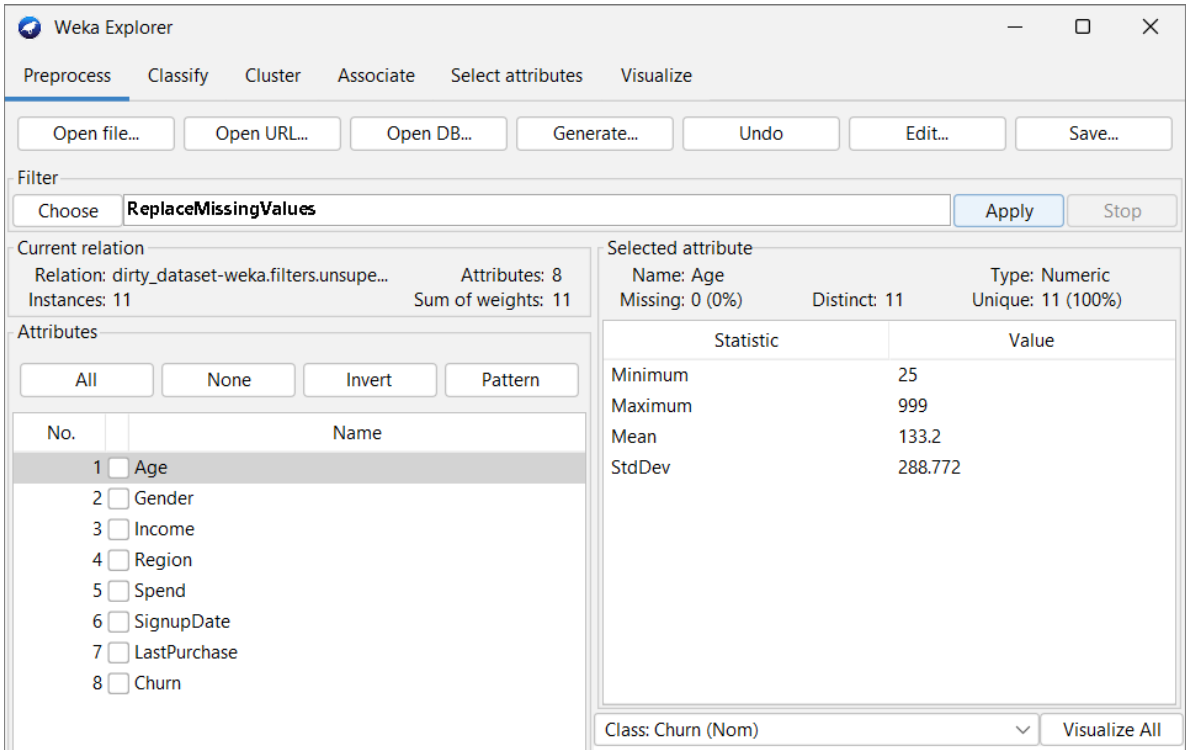
Attributes

No.		Name
1	□	Age
2	□	Gender
3	□	Income
4	□	Region
5	□	Spend
6	□	SignupDate
7	□	LastPurchase
8	□	Churn

Class: Churn (Nom)

5. Replace any missing values

In this case it was done using the unsupervised.attribute.ReplaceMissingValues filter



Filter: ReplaceMissingValues

Current relation
Relation: dirty_dataset-weka.filters.unsupe... Attributes: 8
Instances: 11 Sum of weights: 11

Selected attribute
Name: Age Type: Numeric
Missing: 0 (0%) Distinct: 11 Unique: 11 (100%)

Statistic	Value
Minimum	25
Maximum	999
Mean	133.2
StdDev	288.772

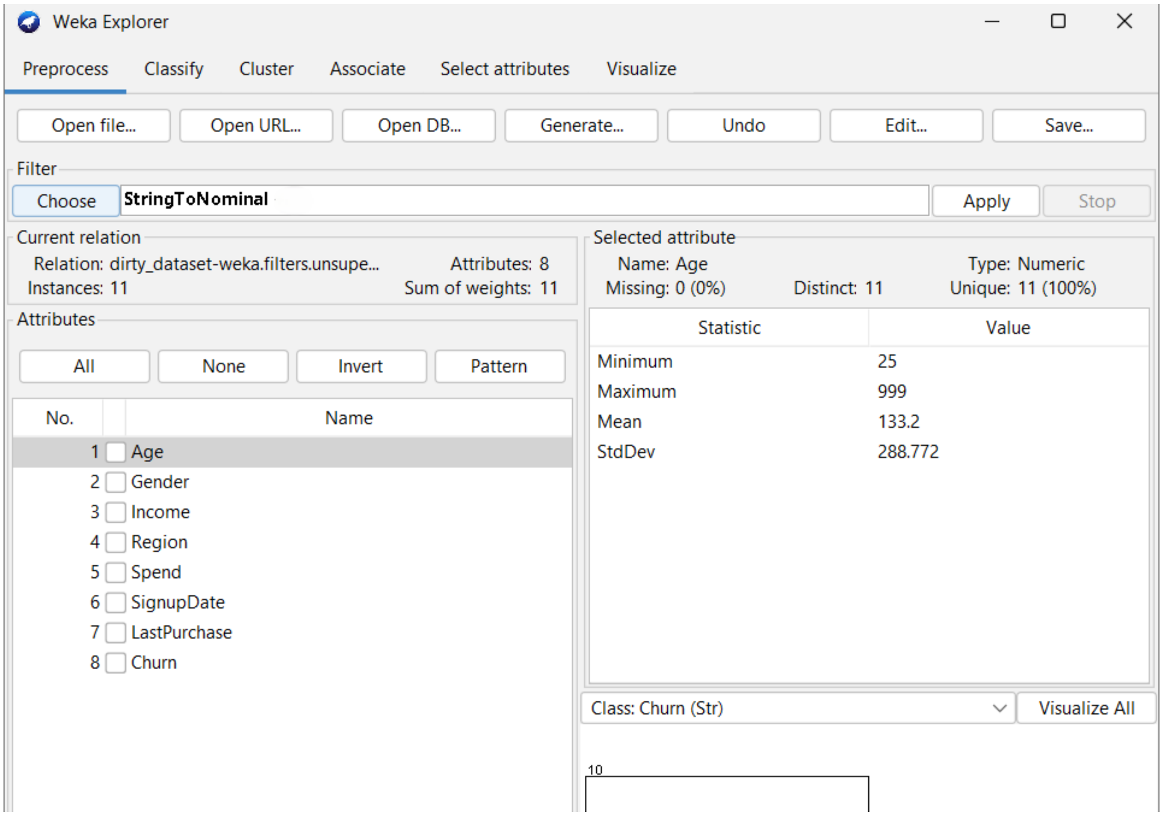
Attributes

No.		Name
1	□	Age
2	□	Gender
3	□	Income
4	□	Region
5	□	Spend
6	□	SignupDate
7	□	LastPurchase
8	□	Churn

Class: Churn (Nom)

6. Convert string into nominal values

This is done using the unsupervised.attribute.StringToNominal filter.



Filter: StringToNominal

Current relation

Relation: dirty_dataset-weka.filters.unsupe... Attributes: 8

Instances: 11 Sum of weights: 11

Selected attribute

Name: Age Type: Numeric

Missing: 0 (0%) Distinct: 11 Unique: 11 (100%)

Statistic	Value
Minimum	25
Maximum	999
Mean	133.2
StdDev	288.772

Attributes

No.		Name
1	□	Age
2	□	Gender
3	□	Income
4	□	Region
5	□	Spend
6	□	SignupDate
7	□	LastPurchase
8	□	Churn

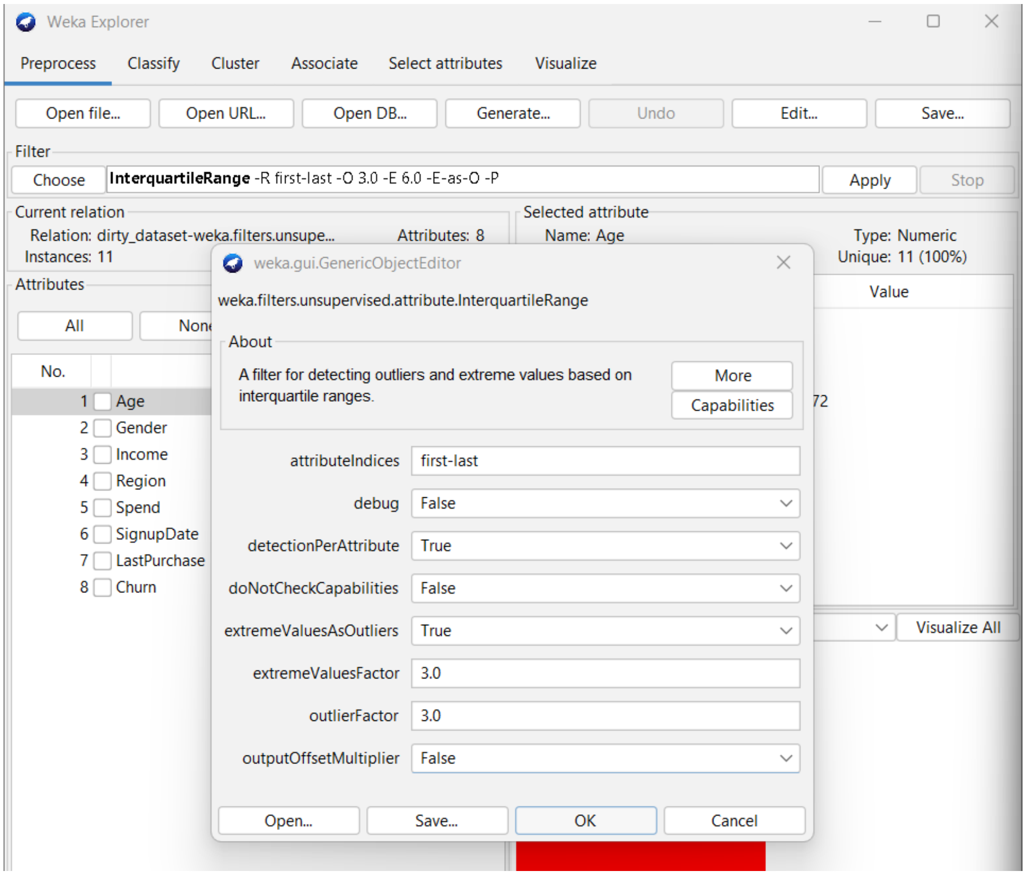
Class: Churn (Str)

7. Removing Outliers

To remove Outliers, we perform the following steps:

7.1. Interquartile Range

Choose Interquartile Range filter from unsupervised.attribute.InterquartileRange and select the following settings. This will give outlier and extreme values in the dataset.



Filter: **InterquartileRange** -R first-last -O 3.0 -E 6.0 -E-as-O -P

Relation: dirty_dataset-weka.filters.unsupe... Attributes: 8
Instances: 11

Name: Age Type: Numeric
Unique: 11 (100%)

1 □ Age
2 □ Gender
3 □ Income
4 □ Region
5 □ Spend
6 □ SignupDate
7 □ LastPurchase
8 □ Churn

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.InterquartileRange

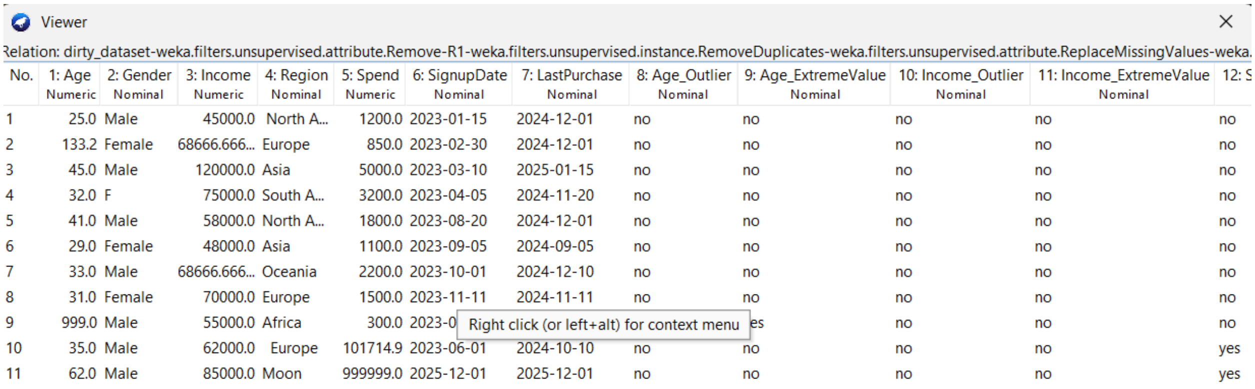
About

A filter for detecting outliers and extreme values based on interquartile ranges.

More
Capabilities

attributeIndices	first-last
debug	False
detectionPerAttribute	True
doNotCheckCapabilities	False
extremeValuesAsOutliers	True
extremeValuesFactor	3.0
outlierFactor	3.0
outputOffsetMultiplier	False

Open... Save... OK Cancel



Viewer

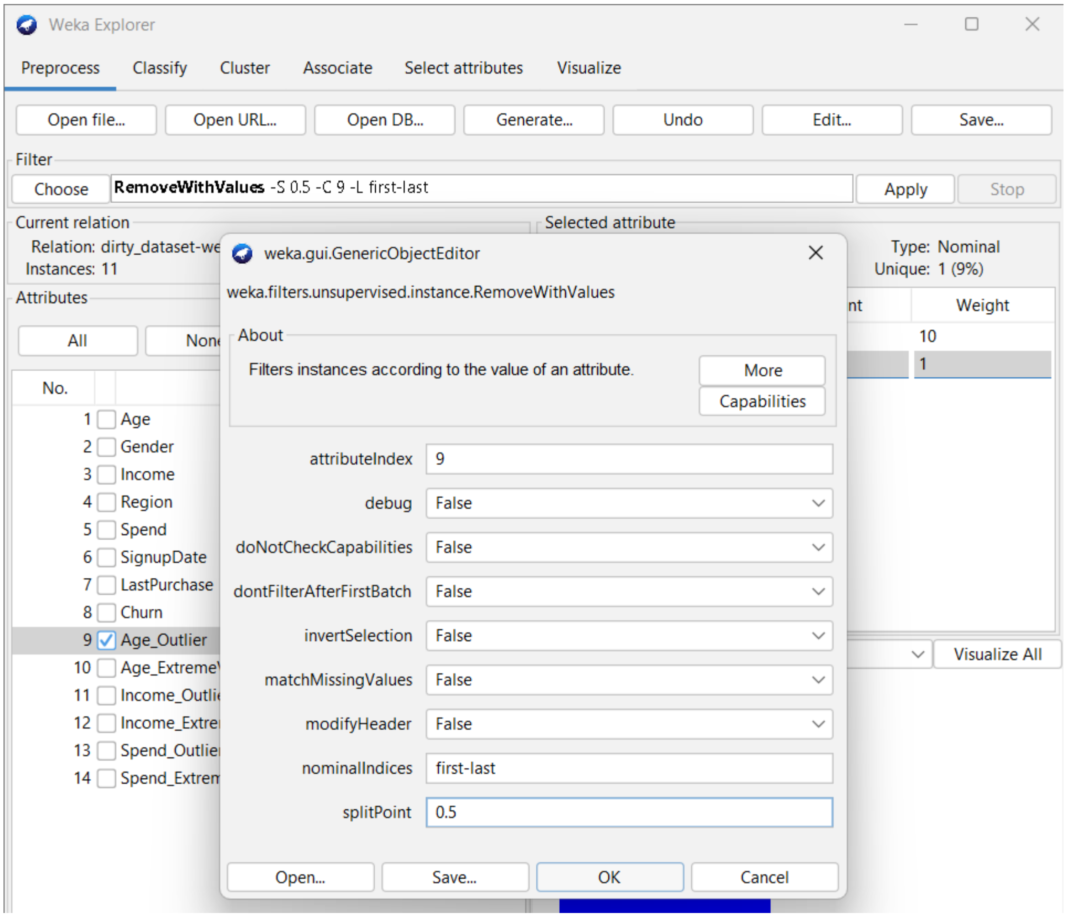
Relation: dirty_dataset-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.

No.	1: Age Numeric	2: Gender Nominal	3: Income Numeric	4: Region Nominal	5: Spend Numeric	6: SignupDate Nominal	7: LastPurchase Nominal	8: Age_Outlier Nominal	9: Age_ExtremeValue Nominal	10: Income_Outlier Nominal	11: Income_ExtremeValue Nominal	12: S
1	25.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	no	no	no	no	no
2	133.2	Female	68666.666...	Europe	850.0	2023-02-30	2024-12-01	no	no	no	no	no
3	45.0	Male	120000.0	Asia	5000.0	2023-03-10	2025-01-15	no	no	no	no	no
4	32.0	F	75000.0	South A...	3200.0	2023-04-05	2024-11-20	no	no	no	no	no
5	41.0	Male	58000.0	North A...	1800.0	2023-08-20	2024-12-01	no	no	no	no	no
6	29.0	Female	48000.0	Asia	1100.0	2023-09-05	2024-09-05	no	no	no	no	no
7	33.0	Male	68666.666...	Oceania	2200.0	2023-10-01	2024-12-10	no	no	no	no	no
8	31.0	Female	70000.0	Europe	1500.0	2023-11-11	2024-11-11	no	no	no	no	no
9	999.0	Male	55000.0	Africa	300.0	2023-0			es	no	no	no
10	35.0	Male	62000.0	Europe	101714.9	2023-06-01	2024-10-10	no	no	no	no	yes
11	62.0	Male	85000.0	Moon	999999.0	2025-12-01	2025-12-01	no	no	no	no	yes

Right click (or left+alt) for context menu

7.2. Remove rows with outliers

To remove the outlier, we use unsupervised.instance.RemoveWithValues filter and apply the following preferences and repeat for attribute indices Age_Outlier, Income_Outlier, and Spend_Outlier (i.e. 9, 11, 13). Here splitPoint is 0.5 because, “No” = 0 and “Yes” = 1, so anything beside “No” will be deleted.



Filter: RemoveWithValues -S 0.5 -C 9 -L first-last

Current relation
Relation: dirty_dataset-we
Instances: 11

Selected attribute
Type: Nominal
Unique: 1 (9%)

No.	
1	□ Age
2	□ Gender
3	□ Income
4	□ Region
5	□ Spend
6	□ SignupDate
7	□ LastPurchase
8	□ Churn
9	☑ Age_Outlier
10	□ Age_Extreme
11	□ Income_Outli
12	□ Income_Extre
13	□ Spend_Outlie
14	□ Spend_Extrem

nt	Weight
	10
	1

weka.gui.GenericObjectEditor

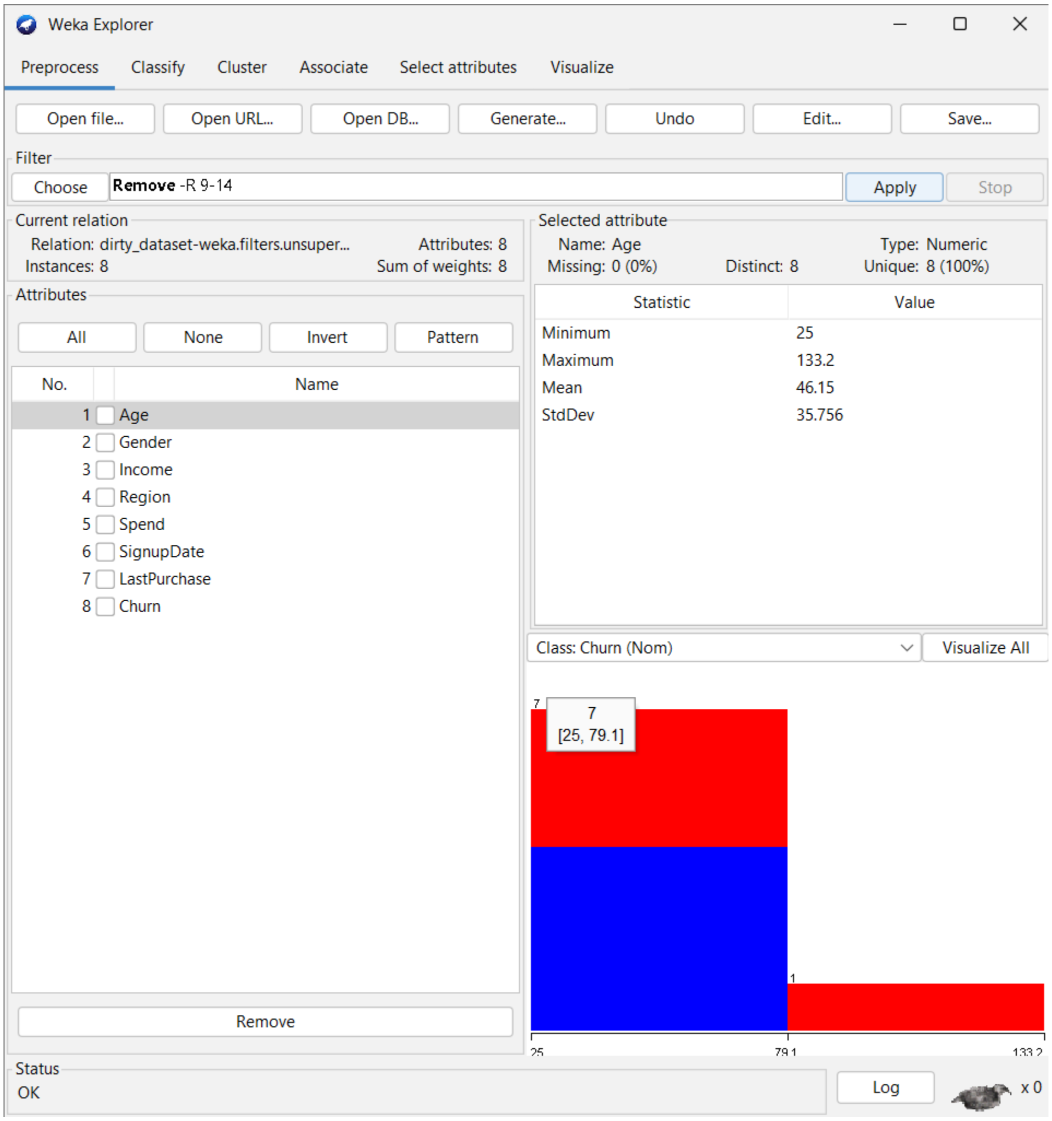
weka.filters.unsupervised.instance.RemoveWithValues

About
Filters instances according to the value of an attribute.

attributeIndex	9
debug	False
doNotCheckCapabilities	False
dontFilterAfterFirstBatch	False
invertSelection	False
matchMissingValues	False
modifyHeader	False
nominalIndices	first-last
splitPoint	0.5

7.3. Remove the columns created

Finally remove the columns from 9 to 14 by unsupervised.attribute.Remove filter.



Filter: **Remove** -R 9-14

Current relation

Relation: dirty_dataset-weka.filters.unsuper... Attributes: 8
Instances: 8 Sum of weights: 8

Selected attribute

Name: Age Type: Numeric
Missing: 0 (0%) Distinct: 8 Unique: 8 (100%)

Attributes

No.	Name
1	□ Age
2	□ Gender
3	□ Income
4	□ Region
5	□ Spend
6	□ SignupDate
7	□ LastPurchase
8	□ Churn

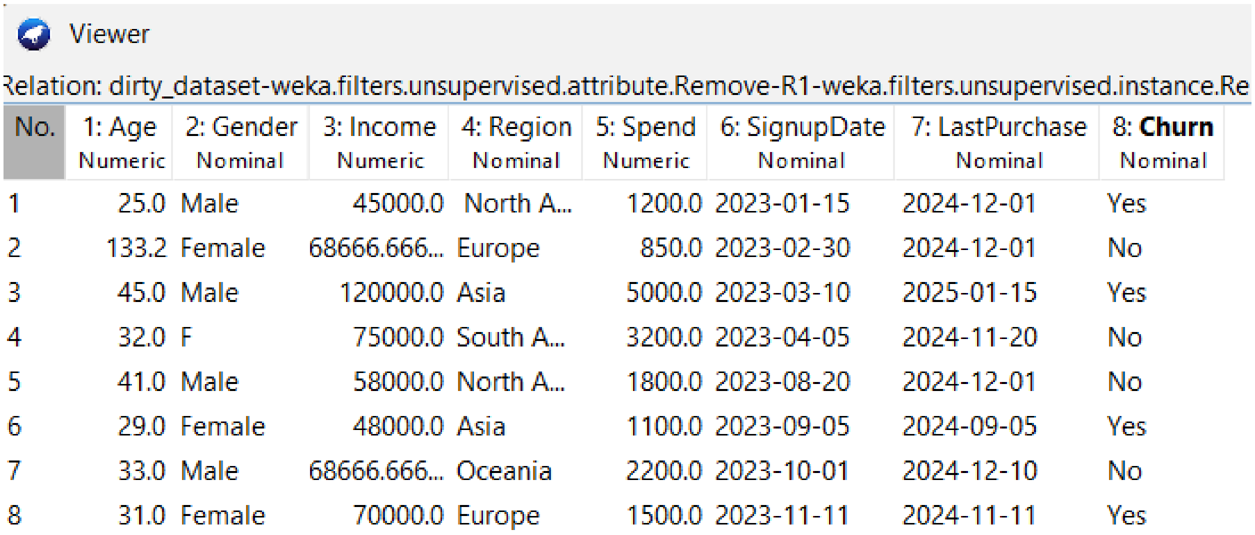
Statistic	Value
Minimum	25
Maximum	133.2
Mean	46.15
StdDev	35.756

Class: Churn (Nom)

Status
OK

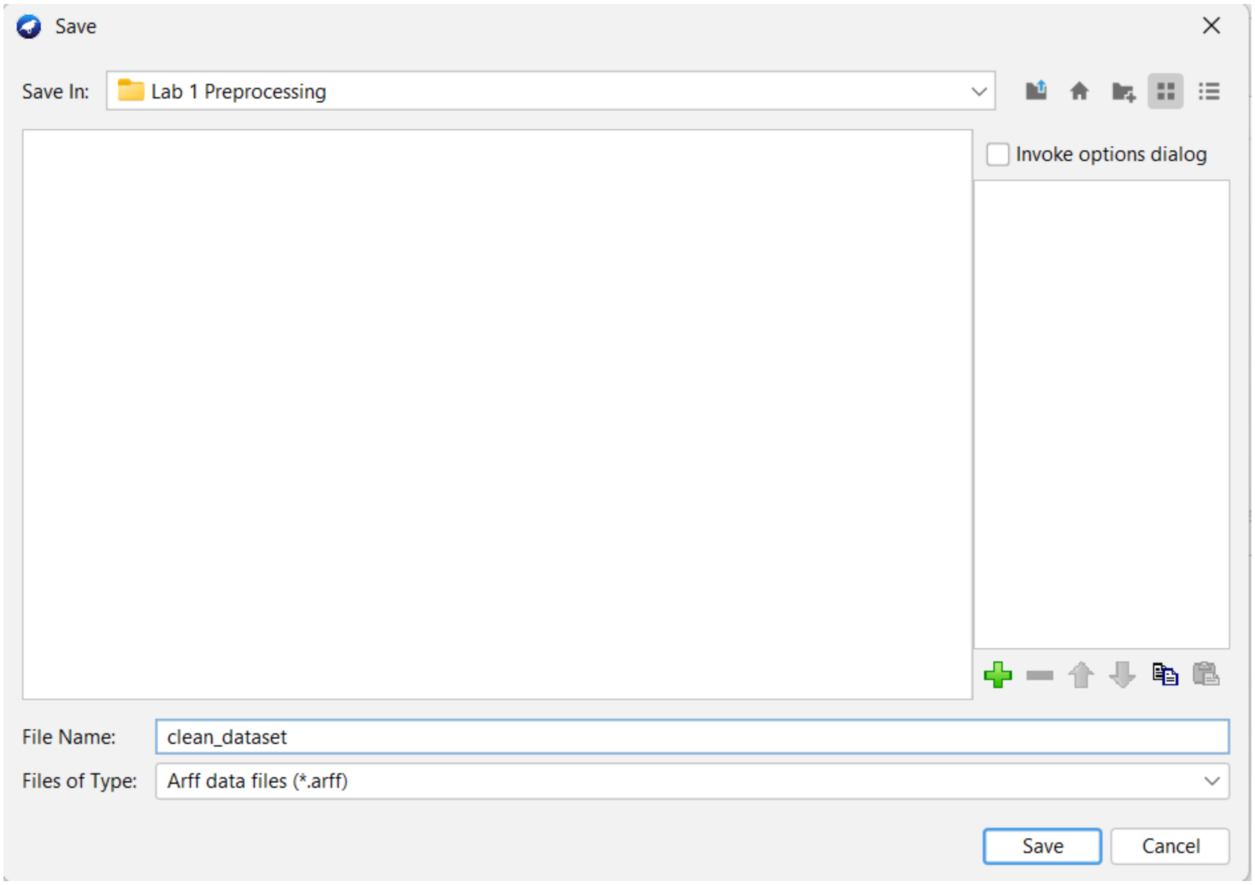
8. Finalize

Data Cleaning Process is done. Visualize and save the clean data.



Relation: dirty_dataset-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.instance.Re

No.	1: Age Numeric	2: Gender Nominal	3: Income Numeric	4: Region Nominal	5: Spend Numeric	6: SignupDate Nominal	7: LastPurchase Nominal	8: Churn Nominal
1	25.0	Male	45000.0	North A...	1200.0	2023-01-15	2024-12-01	Yes
2	133.2	Female	68666.666...	Europe	850.0	2023-02-30	2024-12-01	No
3	45.0	Male	120000.0	Asia	5000.0	2023-03-10	2025-01-15	Yes
4	32.0	F	75000.0	South A...	3200.0	2023-04-05	2024-11-20	No
5	41.0	Male	58000.0	North A...	1800.0	2023-08-20	2024-12-01	No
6	29.0	Female	48000.0	Asia	1100.0	2023-09-05	2024-09-05	Yes
7	33.0	Male	68666.666...	Oceania	2200.0	2023-10-01	2024-12-10	No
8	31.0	Female	70000.0	Europe	1500.0	2023-11-11	2024-11-11	Yes



Save In: Lab 1 Preprocessing

□ Invoke options dialog

File Name: clean_dataset

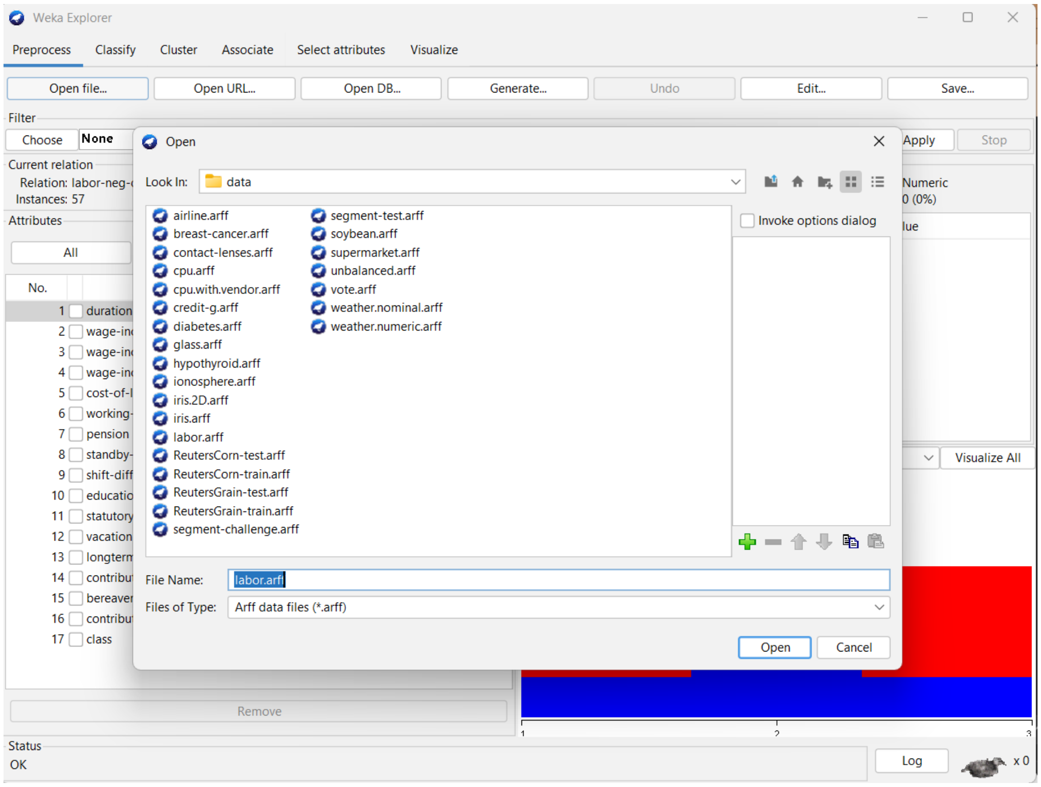
Files of Type: Arff data files (*.arff)

B. For Secondary Dataset

For secondary dataset, default data provided by the Weka, labor.arff was selected.

Steps used to clean the data:

1. Open the dataset in the pre-processor of WEKA



Relation: labor-neg-
Instances: 57

No.	
1	☐ duration
2	☐ wage-in
3	☐ wage-in
4	☐ wage-in
5	☐ cost-of-l
6	☐ working-
7	☐ pension
8	☐ standby-
9	☐ shift-diff
10	☐ educatio
11	☐ statutory
12	☐ vacation
13	☐ longterm
14	☐ contribu
15	☐ bereaver
16	☐ contribu
17	☐ class

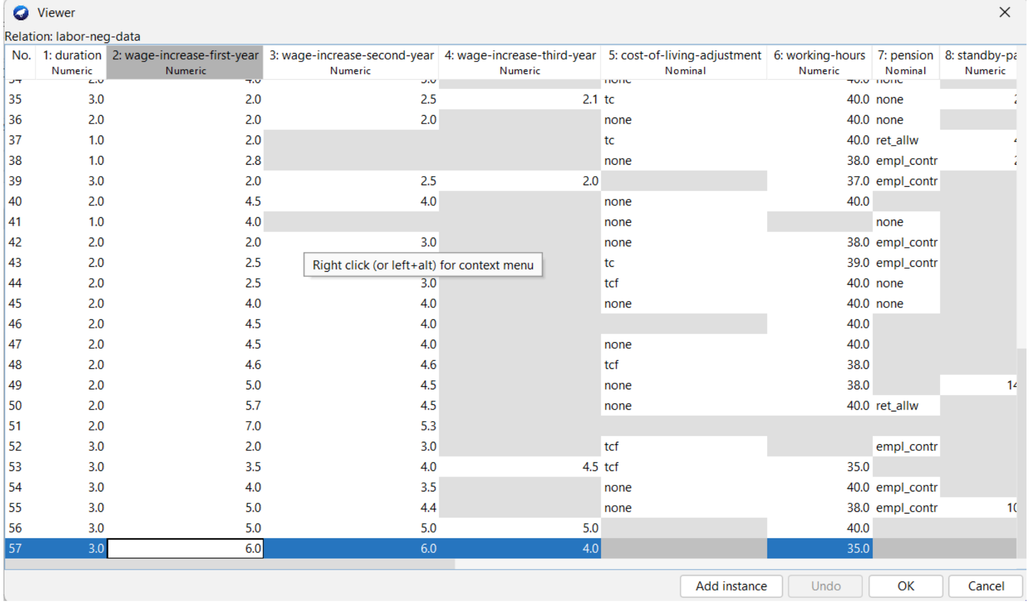
Look In: data

File Name: labor.arff

Files of Type: Arff data files (*.arff)

Status
OK

2. Visualize the data



Relation: labor-neg-data

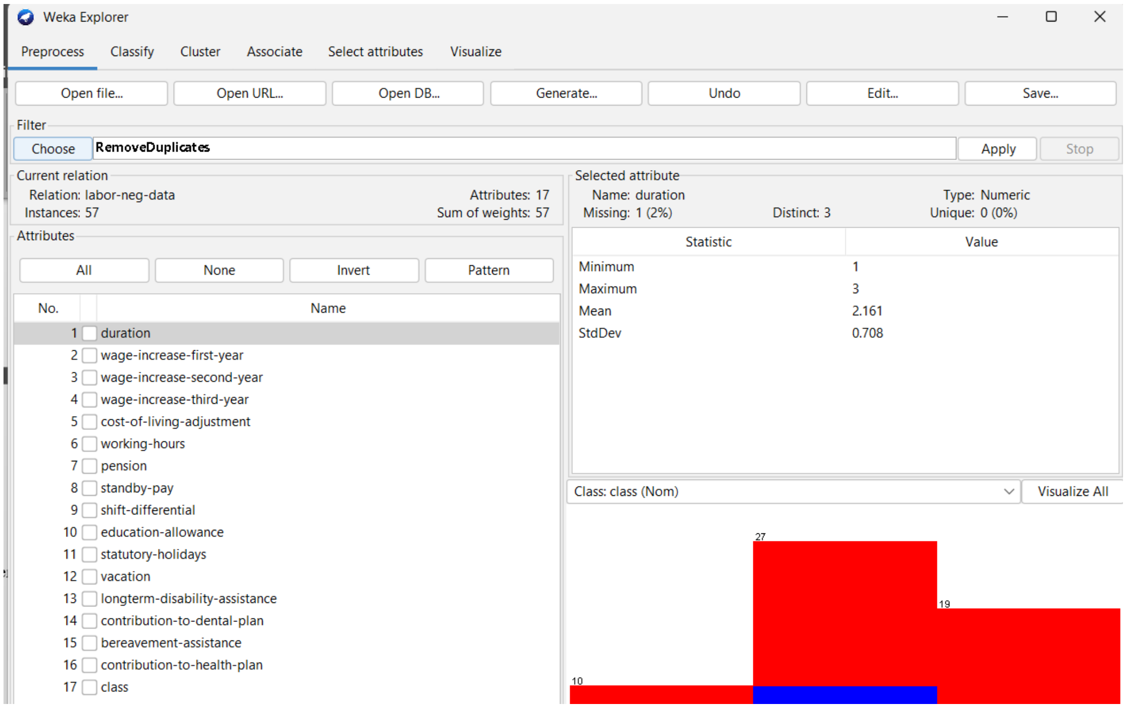
No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second-year Numeric	4: wage-increase-third-year Numeric	5: cost-of-living-adjustment Nominal	6: working-hours Numeric	7: pension Nominal	8: standby-pa Numeric
35	3.0	2.0	2.5	2.1	tc	40.0	none	2
36	2.0	2.0	2.0		none	40.0	none	
37	1.0	2.0			tc	40.0	ret_allw	4
38	1.0	2.8			none	38.0	empl_contr	2
39	3.0	2.0	2.5	2.0		37.0	empl_contr	
40	2.0	4.5	4.0		none	40.0		
41	1.0	4.0			none		none	
42	2.0	2.0	3.0		none	38.0	empl_contr	
43	2.0	2.5			tc	39.0	empl_contr	
44	2.0	2.5	3.0		tcf	40.0	none	
45	2.0	4.0	4.0		none	40.0	none	
46	2.0	4.5	4.0			40.0		
47	2.0	4.5	4.0		none	40.0		
48	2.0	4.6	4.6		tcf	38.0		
49	2.0	5.0	4.5		none	38.0		14
50	2.0	5.7	4.5		none	40.0	ret_allw	
51	2.0	7.0	5.3					
52	3.0	2.0	3.0		tcf		empl_contr	
53	3.0	3.5	4.0	4.5	tcf	35.0		
54	3.0	4.0	3.5		none	40.0	empl_contr	
55	3.0	5.0	4.4		none	38.0	empl_contr	10
56	3.0	5.0	5.0	5.0		40.0		
57	3.0	6.0	6.0	4.0		35.0		

3. Remove unwanted columns

This step was not necessary as all columns were needed.

4. Remove any duplicate values

In this case this was done using the unsupervised.instance.RemoveDuplicates filter.



Filter: RemoveDuplicates

Current relation
Relation: labor-neg-data Attributes: 17
Instances: 57 Sum of weights: 57

Selected attribute
Name: duration Type: Numeric
Missing: 1 (2%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	3
Mean	2.161
StdDev	0.708

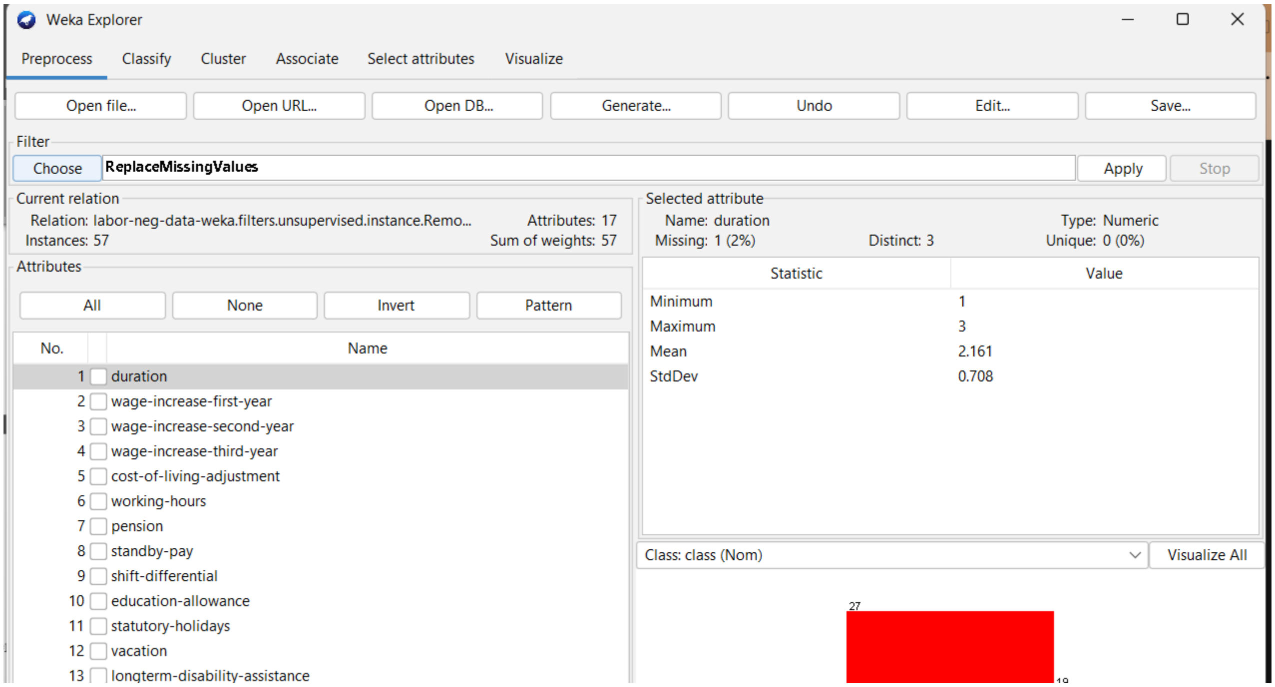
Class: class (Nom)

Attributes

No.	Name
1	duration
2	wage-increase-first-year
3	wage-increase-second-year
4	wage-increase-third-year
5	cost-of-living-adjustment
6	working-hours
7	pension
8	standby-pay
9	shift-differential
10	education-allowance
11	statutory-holidays
12	vacation
13	longterm-disability-assistance
14	contribution-to-dental-plan
15	bereavement-assistance
16	contribution-to-health-plan
17	class

5. Replace any missing values

In this case it was done using the unsupervised.attribute.ReplaceMissingValues filter



Filter: ReplaceMissingValues

Current relation
Relation: labor-neg-data-weka.filters.unsupervised.instance.Remo... Attributes: 17
Instances: 57 Sum of weights: 57

Selected attribute
Name: duration Type: Numeric
Missing: 1 (2%) Distinct: 3 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	3
Mean	2.161
StdDev	0.708

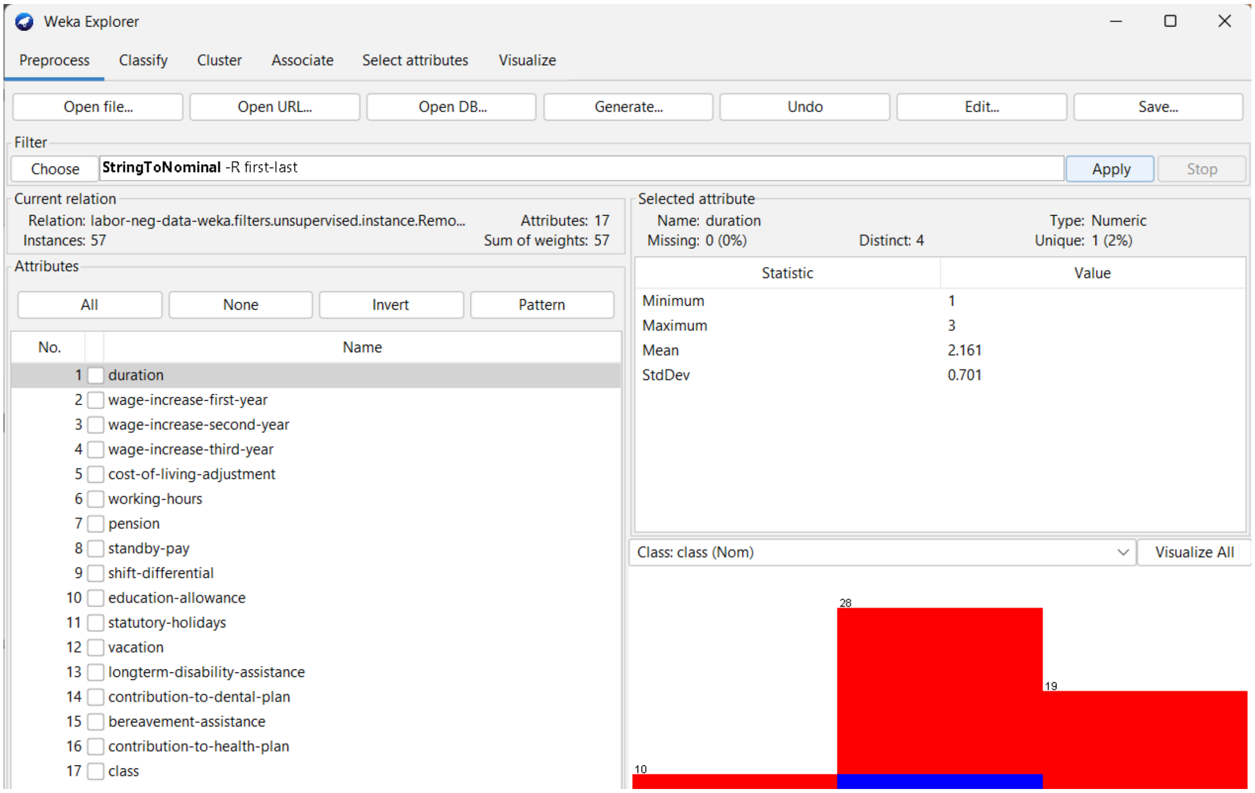
Class: class (Nom)

Attributes

No.	Name
1	duration
2	wage-increase-first-year
3	wage-increase-second-year
4	wage-increase-third-year
5	cost-of-living-adjustment
6	working-hours
7	pension
8	standby-pay
9	shift-differential
10	education-allowance
11	statutory-holidays
12	vacation
13	longterm-disability-assistance

6. Convert string into nominal values

This is done using the unsupervised.attribute.StringToNominal filter.



Filter

StringToNominal -R first-last

Current relation

Relation: labor-neg-data-weka.filters.unsupervised.instance.Remo... Attributes: 17
Instances: 57 Sum of weights: 57

Attributes

No.	Name
1	□ duration
2	□ wage-increase-first-year
3	□ wage-increase-second-year
4	□ wage-increase-third-year
5	□ cost-of-living-adjustment
6	□ working-hours
7	□ pension
8	□ standby-pay
9	□ shift-differential
10	□ education-allowance
11	□ statutory-holidays
12	□ vacation
13	□ longterm-disability-assistance
14	□ contribution-to-dental-plan
15	□ bereavement-assistance
16	□ contribution-to-health-plan
17	□ class

Selected attribute

Name: duration Type: Numeric
Missing: 0 (0%) Distinct: 4 Unique: 1 (2%)

Statistic	Value
Minimum	1
Maximum	3
Mean	2.161
StdDev	0.701

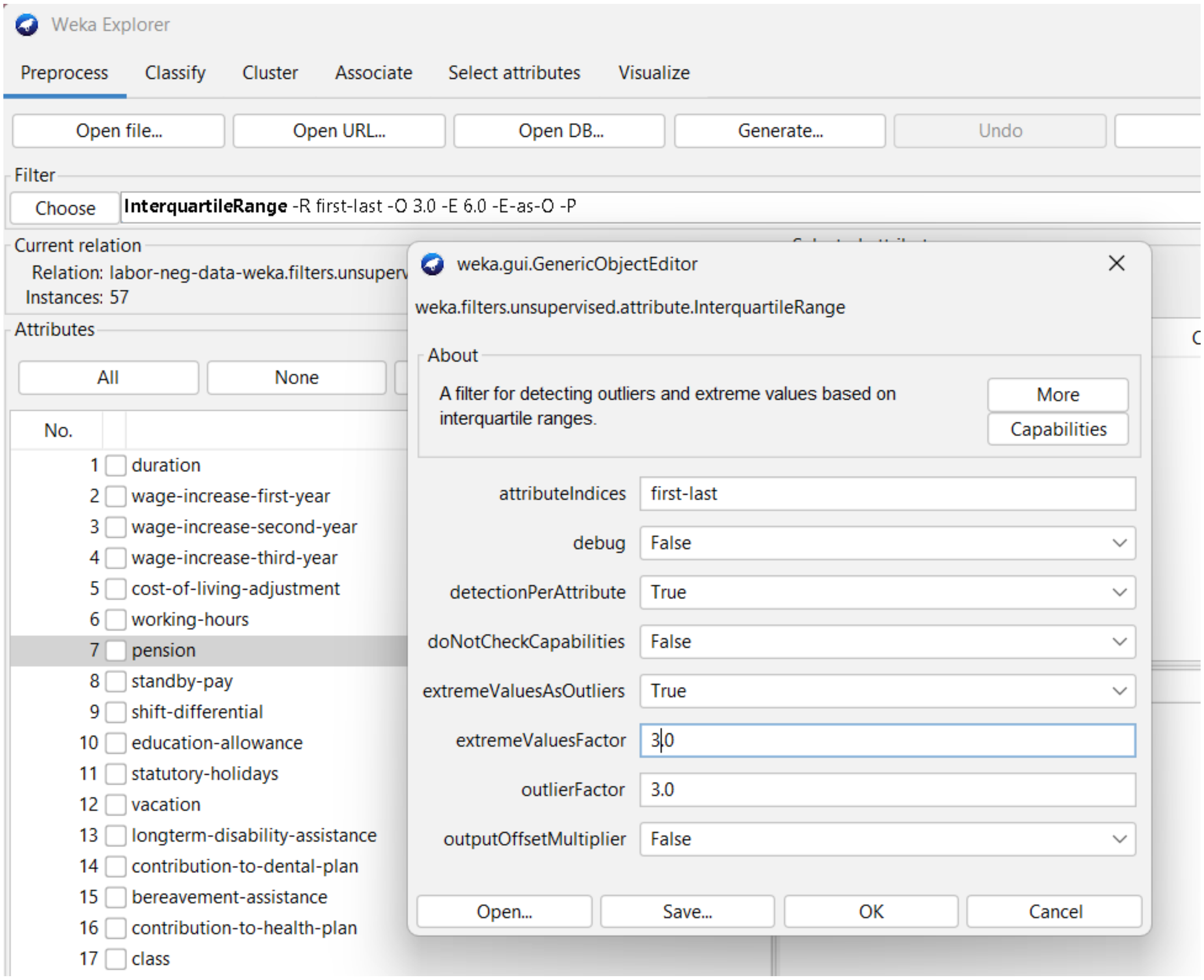
Class: class (Nom)

7. Removing Outliers

To remove Outliers, we perform the following steps:

7.1. Interquartile Range

Choose Interquartile Range filter from unsupervised.attribute.InterquartileRange and select the following settings. This will give outlier and extreme values in the dataset.



Filter: **InterquartileRange** -R first-last -O 3.0 -E 6.0 -E-as-O -P

Relation: labor-neg-data-weka.filters.unsuperv

Instances: 57

No.	
1	duration
2	wage-increase-first-year
3	wage-increase-second-year
4	wage-increase-third-year
5	cost-of-living-adjustment
6	working-hours
7	pension
8	standby-pay
9	shift-differential
10	education-allowance
11	statutory-holidays
12	vacation
13	longterm-disability-assistance
14	contribution-to-dental-plan
15	bereavement-assistance
16	contribution-to-health-plan
17	class

weka.filters.unsupervised.attribute.InterquartileRange

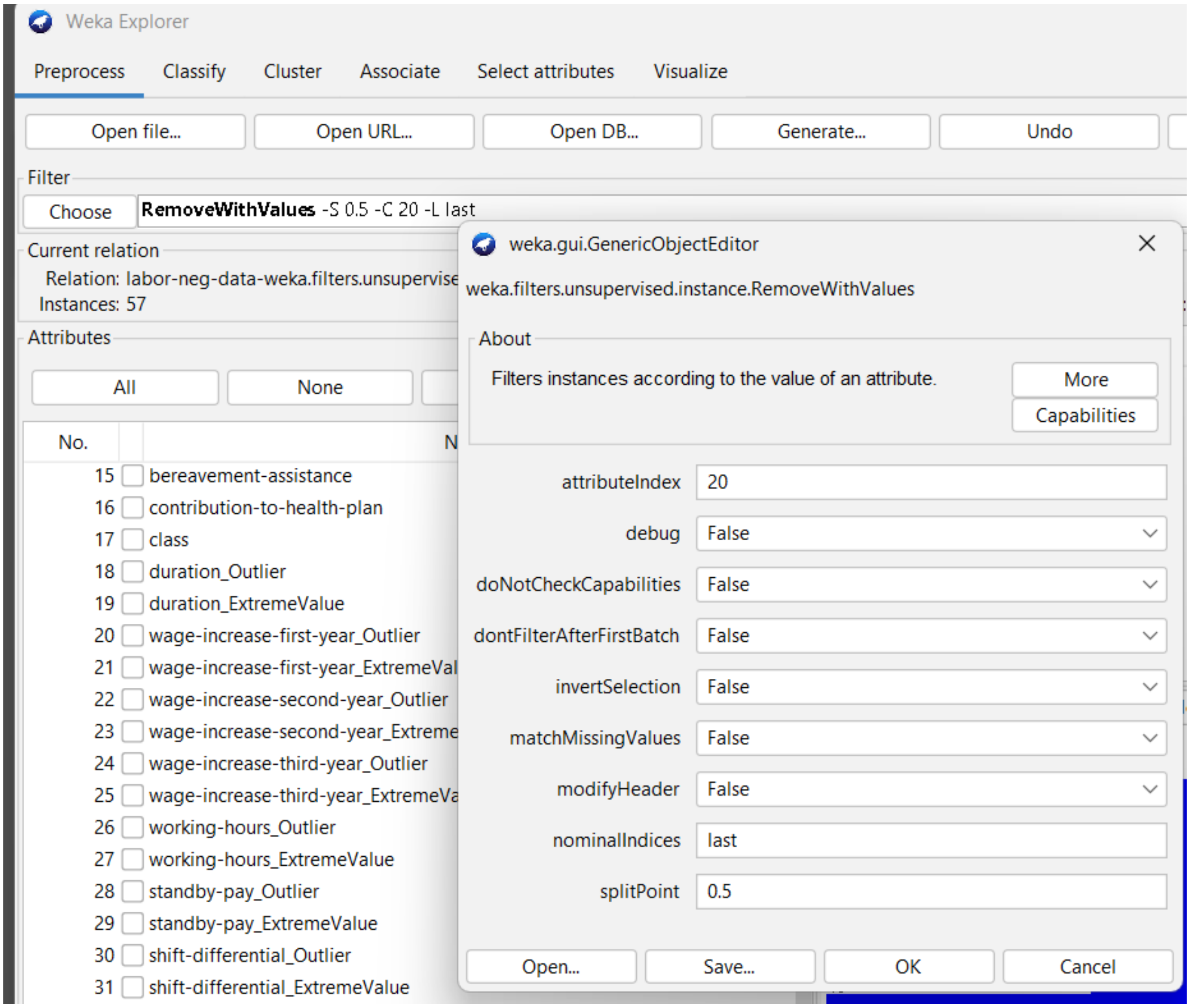
About

A filter for detecting outliers and extreme values based on interquartile ranges.

attributeIndices	first-last
debug	False
detectionPerAttribute	True
doNotCheckCapabilities	False
extremeValuesAsOutliers	True
extremeValuesFactor	3.0
outlierFactor	3.0
outputOffsetMultiplier	False

7.2. Remove rows with outliers

To remove the rows with outliers, we use unsupervised.instance.RemoveWithValues filter and apply the following preferences and repeat for attribute indices 18 to 32. Here, splitPoint is 0.5 because, “No” = 0 and “Yes” = 1, so anything besides “No” will be deleted.



RemoveWithValues -S 0.5 -C 20 -L last

Relation: labor-neg-data-weka.filters.unsupervise

Instances: 57

No.		N
15	□	bereavement-assistance
16	□	contribution-to-health-plan
17	□	class
18	□	duration_Outlier
19	□	duration_ExtremeValue
20	□	wage-increase-first-year_Outlier
21	□	wage-increase-first-year_ExtremeVal
22	□	wage-increase-second-year_Outlier
23	□	wage-increase-second-year_Extreme
24	□	wage-increase-third-year_Outlier
25	□	wage-increase-third-year_ExtremeVa
26	□	working-hours_Outlier
27	□	working-hours_ExtremeValue
28	□	standby-pay_Outlier
29	□	standby-pay_ExtremeValue
30	□	shift-differential_Outlier
31	□	shift-differential_ExtremeValue

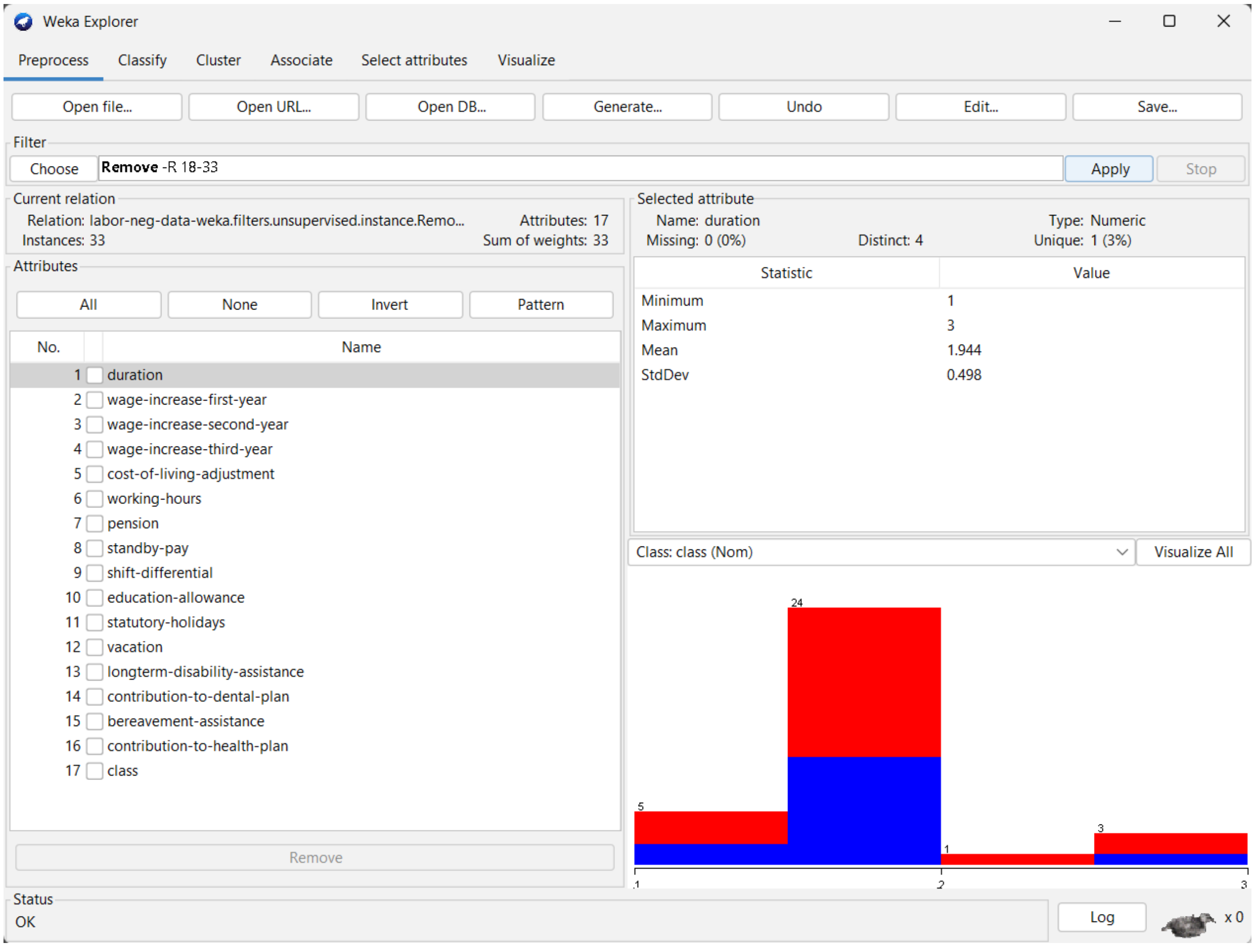
weka.filters.unsupervised.instance.RemoveWithValues

Filters instances according to the value of an attribute.

attributeIndex	20
debug	False
doNotCheckCapabilities	False
dontFilterAfterFirstBatch	False
invertSelection	False
matchMissingValues	False
modifyHeader	False
nominalIndices	last
splitPoint	0.5

7.3. Remove the columns created

Finally remove the columns from 9 to 14 by unsupervised.attribute.Remove filter.



Filter: **Remove** -R 18-33

Current relation
Relation: labor-neg-data-weka.filters.unsupervised.instance.Remo... Attributes: 17
Instances: 33 Sum of weights: 33

Selected attribute
Name: duration Type: Numeric
Missing: 0 (0%) Distinct: 4 Unique: 1 (3%)

Statistic	Value
Minimum	1
Maximum	3
Mean	1.944
StdDev	0.498

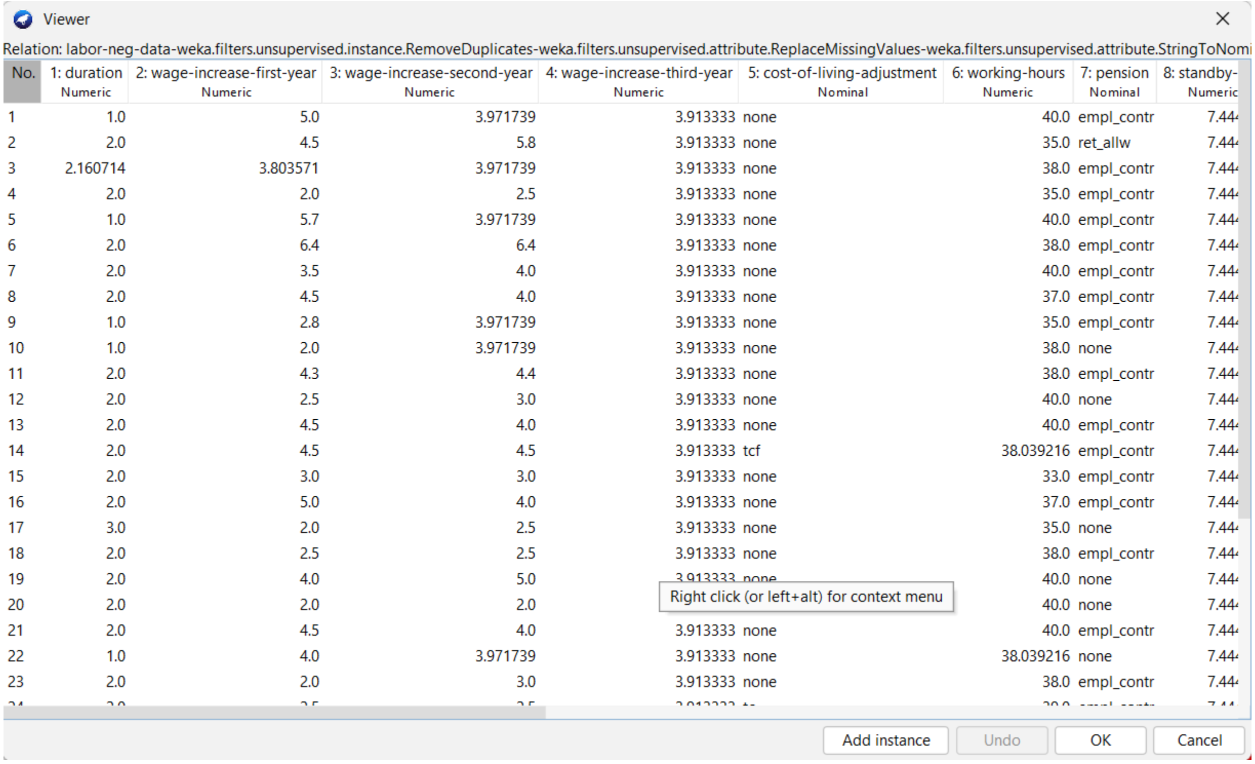
No.	Name
1	duration
2	wage-increase-first-year
3	wage-increase-second-year
4	wage-increase-third-year
5	cost-of-living-adjustment
6	working-hours
7	pension
8	standby-pay
9	shift-differential
10	education-allowance
11	statutory-holidays
12	vacation
13	longterm-disability-assistance
14	contribution-to-dental-plan
15	bereavement-assistance
16	contribution-to-health-plan
17	class

Class: class (Nom)

Status
OK

8. Finalize

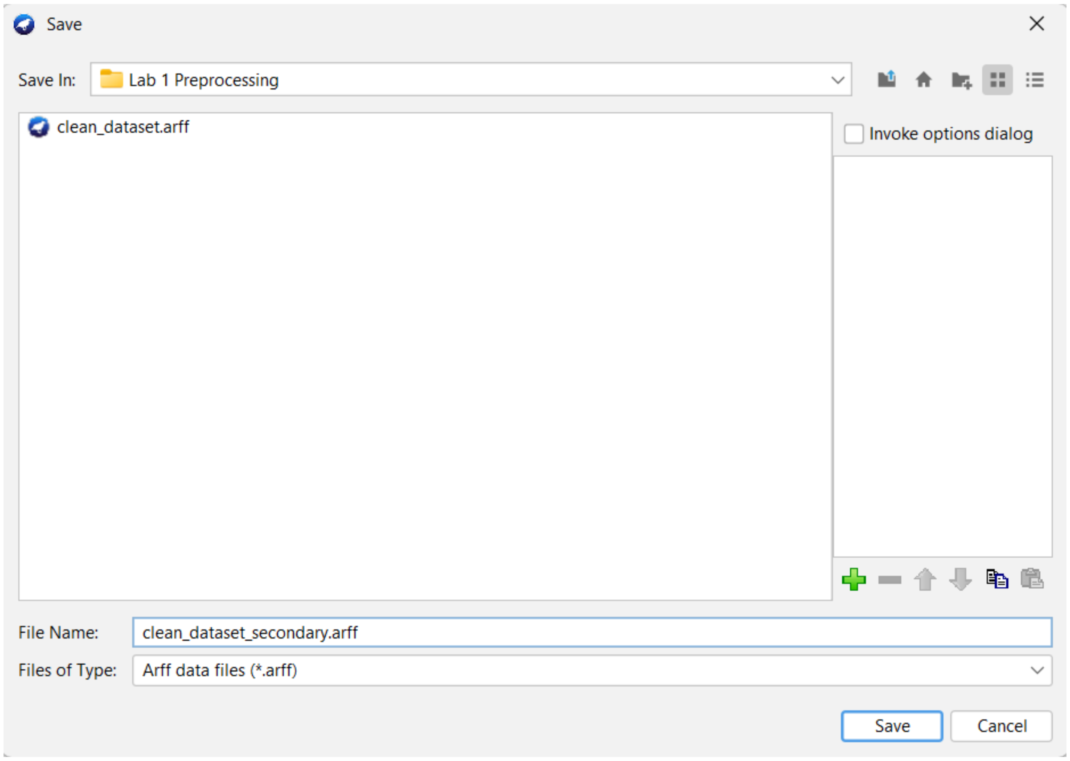
Data Cleaning Process is done. Visualize and save the clean data.



Relation: labor-neg-data-weka.filters.unsupervised.instance.RemoveDuplicates-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.StringToNomi

No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second-year Numeric	4: wage-increase-third-year Numeric	5: cost-of-living-adjustment Nominal	6: working-hours Numeric	7: pension Nominal	8: standby- Numeric
1	1.0	5.0	3.971739	3.913333	none	40.0	empl_contr	7.44
2	2.0	4.5	5.8	3.913333	none	35.0	ret_allw	7.44
3	2.160714	3.803571	3.971739	3.913333	none	38.0	empl_contr	7.44
4	2.0	2.0	2.5	3.913333	none	35.0	empl_contr	7.44
5	1.0	5.7	3.971739	3.913333	none	40.0	empl_contr	7.44
6	2.0	6.4	6.4	3.913333	none	38.0	empl_contr	7.44
7	2.0	3.5	4.0	3.913333	none	40.0	empl_contr	7.44
8	2.0	4.5	4.0	3.913333	none	37.0	empl_contr	7.44
9	1.0	2.8	3.971739	3.913333	none	35.0	empl_contr	7.44
10	1.0	2.0	3.971739	3.913333	none	38.0	none	7.44
11	2.0	4.3	4.4	3.913333	none	38.0	empl_contr	7.44
12	2.0	2.5	3.0	3.913333	none	40.0	none	7.44
13	2.0	4.5	4.0	3.913333	none	40.0	empl_contr	7.44
14	2.0	4.5	4.5	3.913333	tcf	38.039216	empl_contr	7.44
15	2.0	3.0	3.0	3.913333	none	33.0	empl_contr	7.44
16	2.0	5.0	4.0	3.913333	none	37.0	empl_contr	7.44
17	3.0	2.0	2.5	3.913333	none	35.0	none	7.44
18	2.0	2.5	2.5	3.913333	none	38.0	empl_contr	7.44
19	2.0	4.0	5.0	3.913333	none	40.0	none	7.44
20	2.0	2.0	2.0			40.0	none	7.44
21	2.0	4.5	4.0	3.913333	none	40.0	empl_contr	7.44
22	1.0	4.0	3.971739	3.913333	none	38.039216	none	7.44
23	2.0	2.0	3.0	3.913333	none	38.0	empl_contr	7.44

Right click (or left+alt) for context menu



Save In: Lab 1 Preprocessing

clean_dataset.arff

☐ Invoke options dialog

File Name: clean_dataset_secondary.arff

Files of Type: Arff data files (*.arff)