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
Primary Dataset: The original, raw, operational/transactional data that comes directly from the source system (OLTP). It contains all details in one big flat table with lots of redundancy, mixed data types, and no optimization for analysis.

Secondary Dataset: Data that has been extracted, cleaned, transformed, and loaded (ETL) into a data warehouse for analytical processing (OLAP). It is split into fact tables + dimension tables (Star/Snowflake/Galaxy schemas).

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

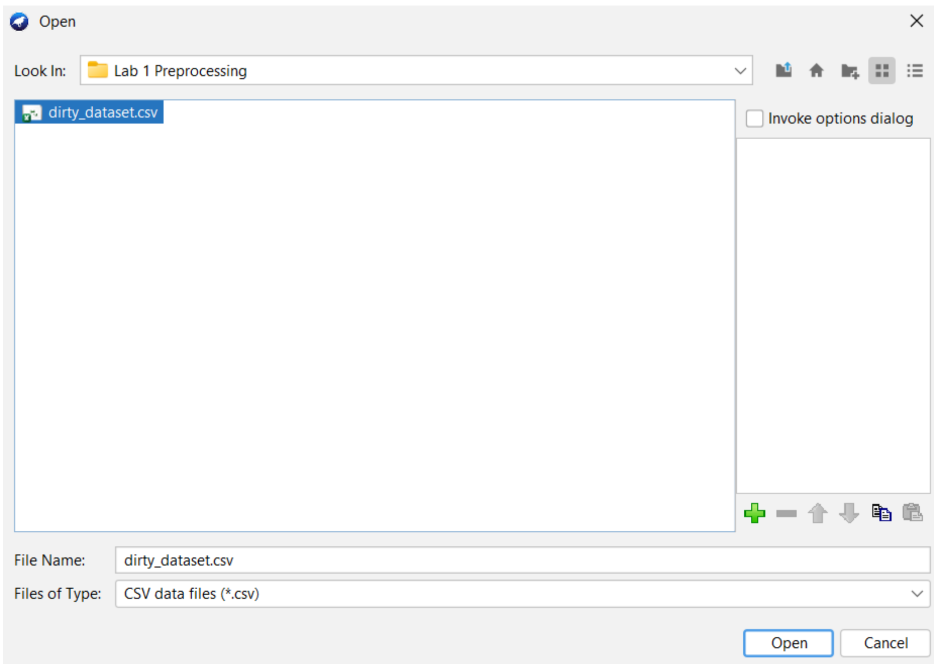


Figure 1: Opening the dataset in weka

2. Visualize the data

Viewer

Relation: dirty_dataset

No.	1: CustomerID String	2: Age Numeric	3: Gender Nominal	4: Income Numeric	5: Region Nominal	6: Spend Numeric	7: SignupDate Nominal	8: LastPurchase Nominal	9: Churn 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

Figure 2: Visualization of the dataset

3. Remove unwanted columns

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

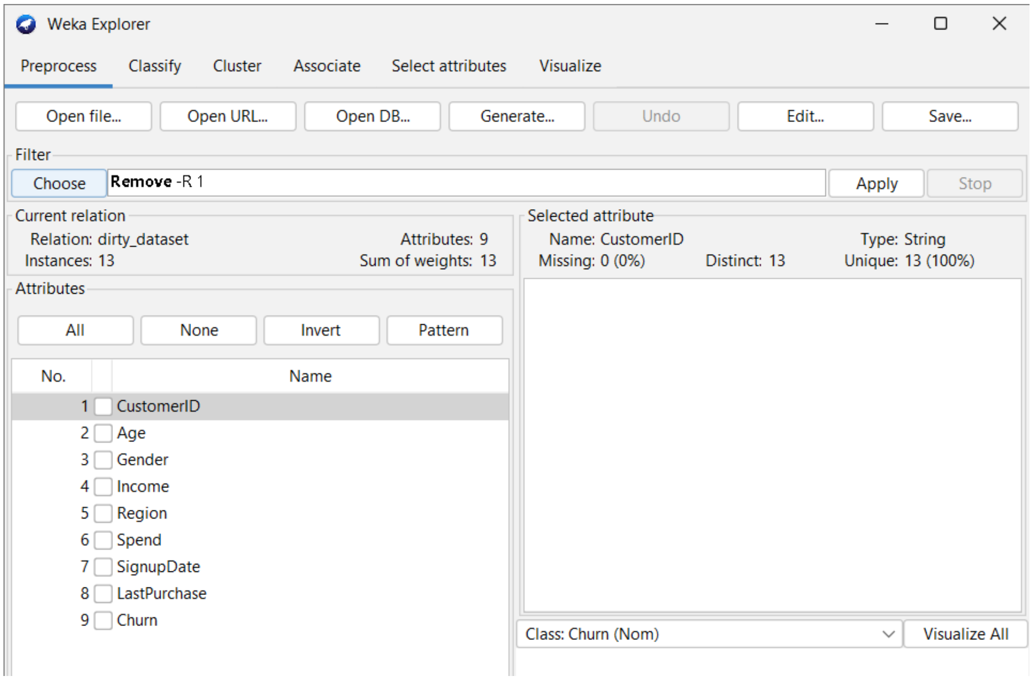


Figure 3: Applying the remove filter

Viewer

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

Right c

Figure 4: Dataset after removing the 1st column

4. Remove any duplicate values

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

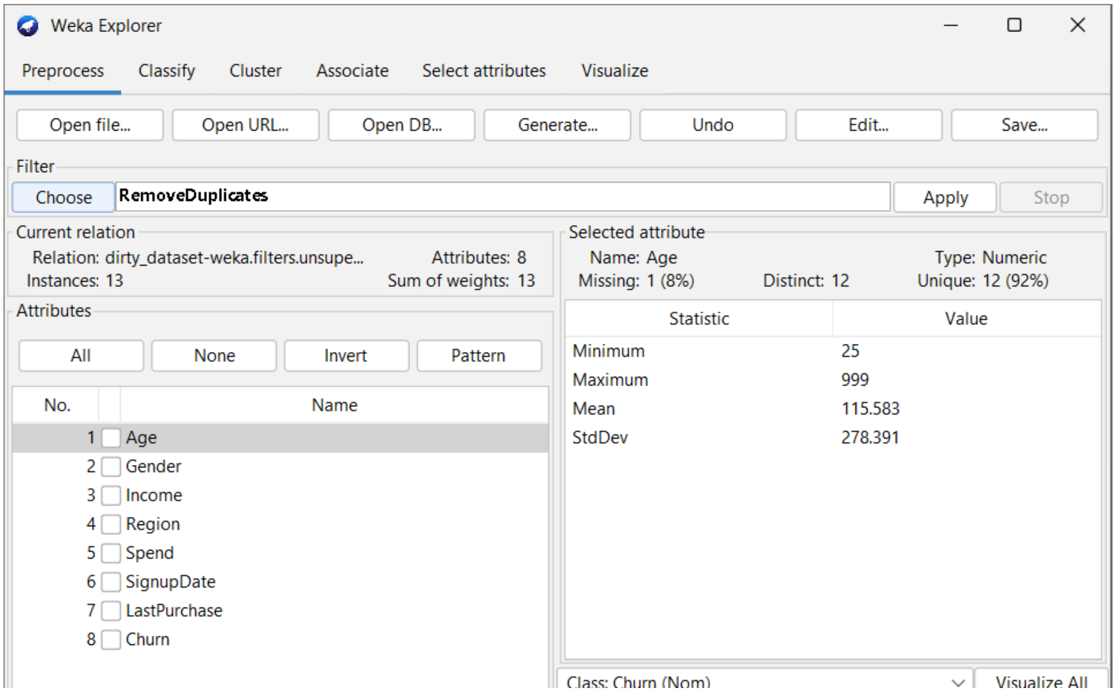


Figure 5: Applying the RemoveDuplicate Filter

5. Replace any missing values

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

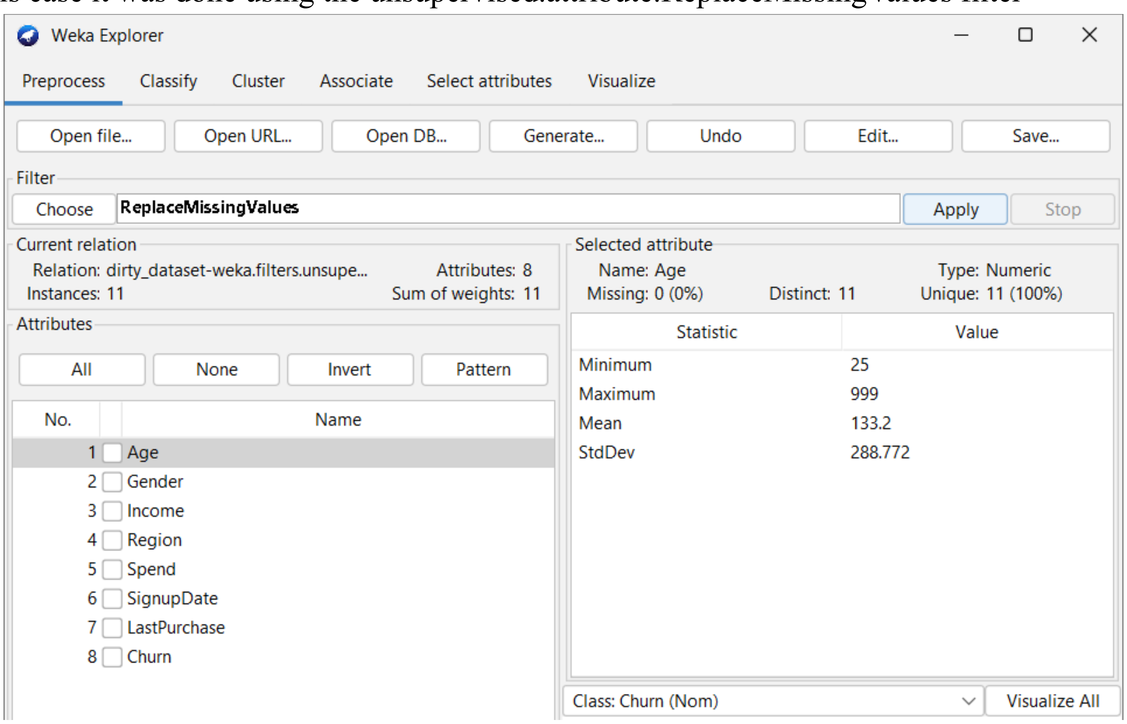


Figure 6: Applying the ReplaceMissingValues filter

6. Convert string into nominal values

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

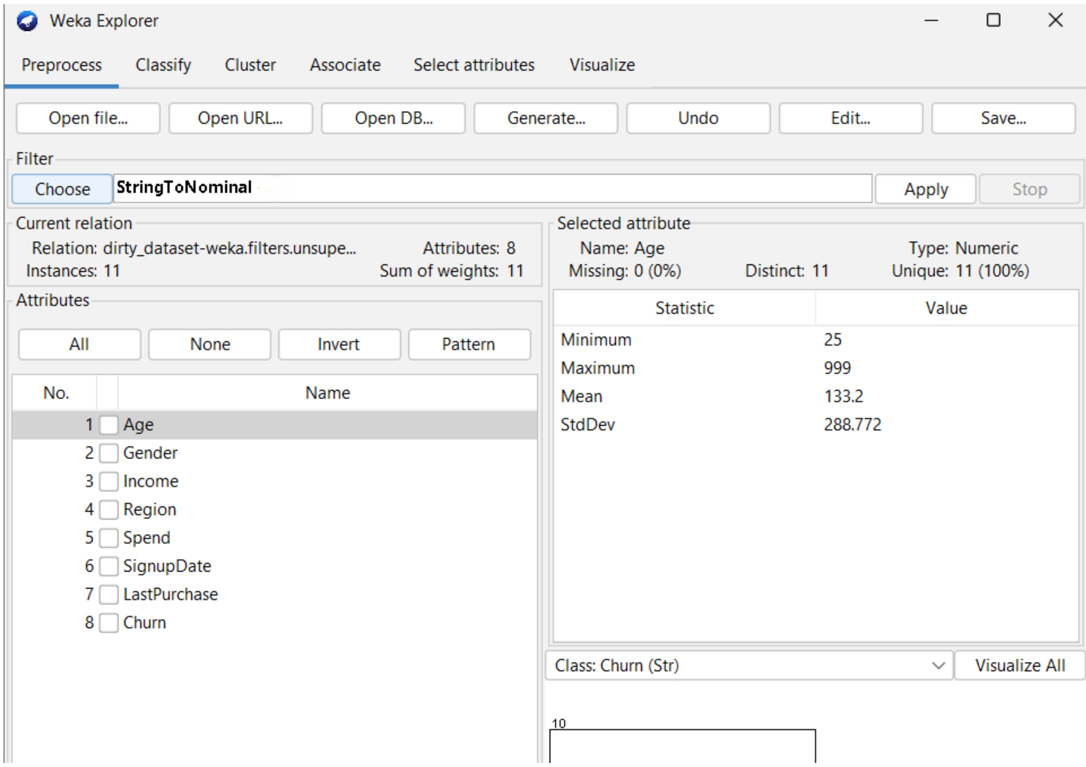


Figure 7: Applying the StringToNominal Filter

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.

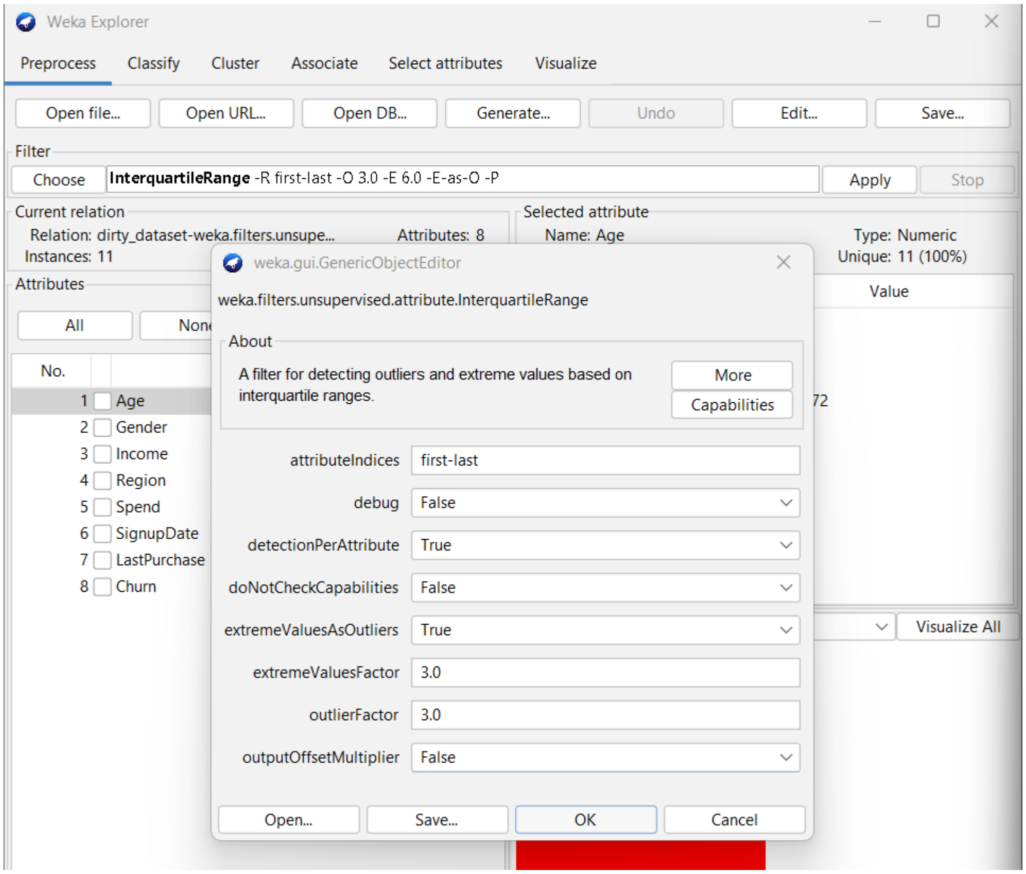
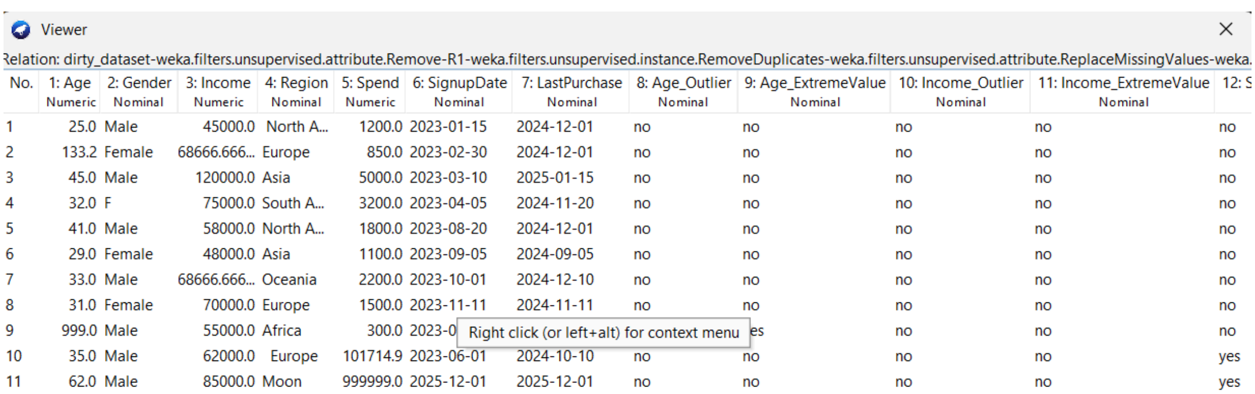


Figure 8: Selecting appropriate setting for InterquartileRange



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	[illegible]	[illegible]	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

Figure 9: Table showing newly added columns after applying InterquartileRange Filter

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.

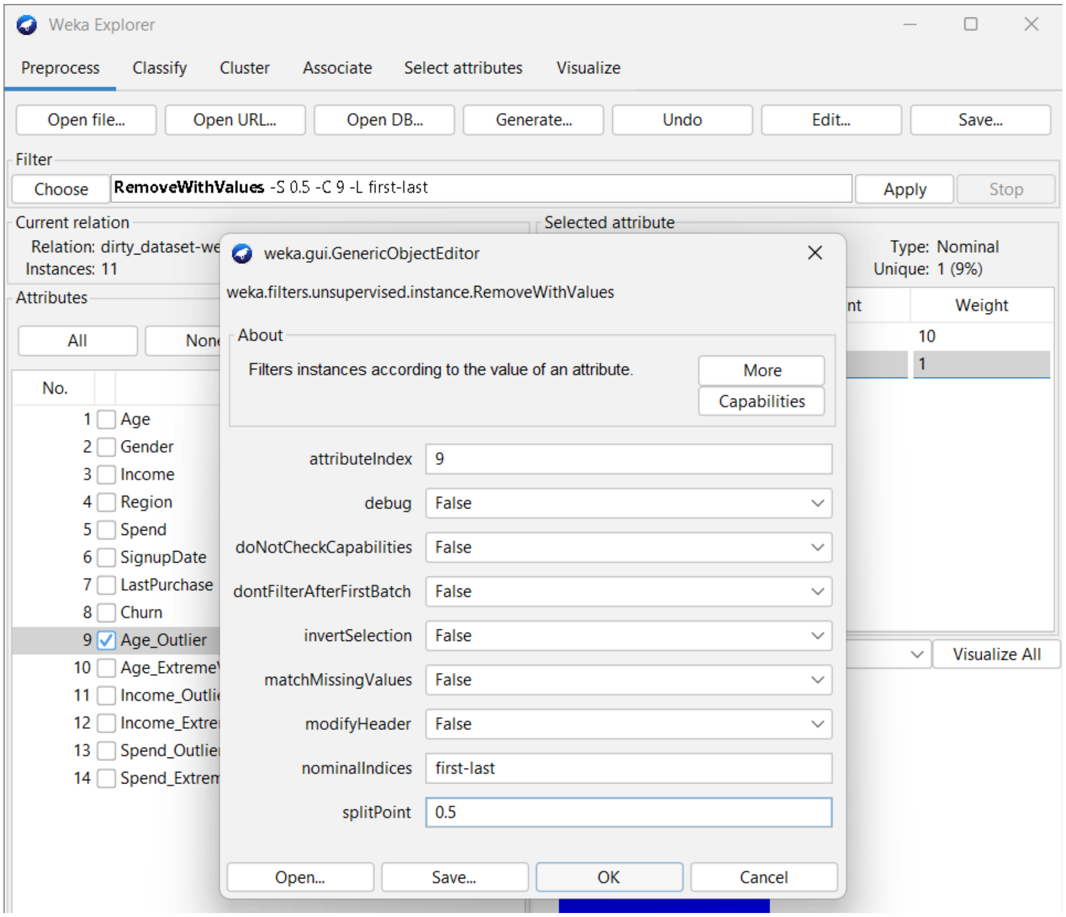


Figure 10: Removing row with outliers for column 9

7.3. Remove the columns created

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

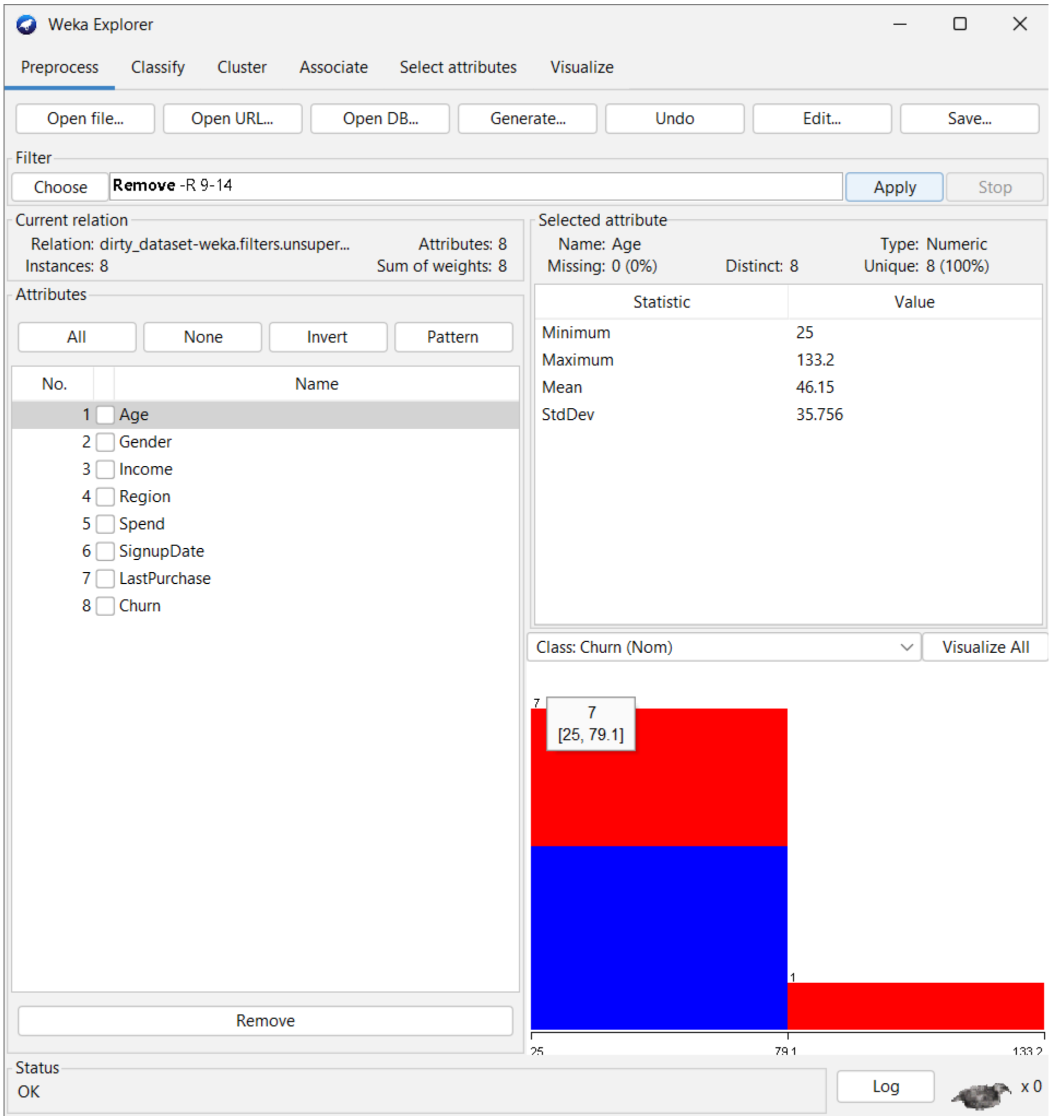
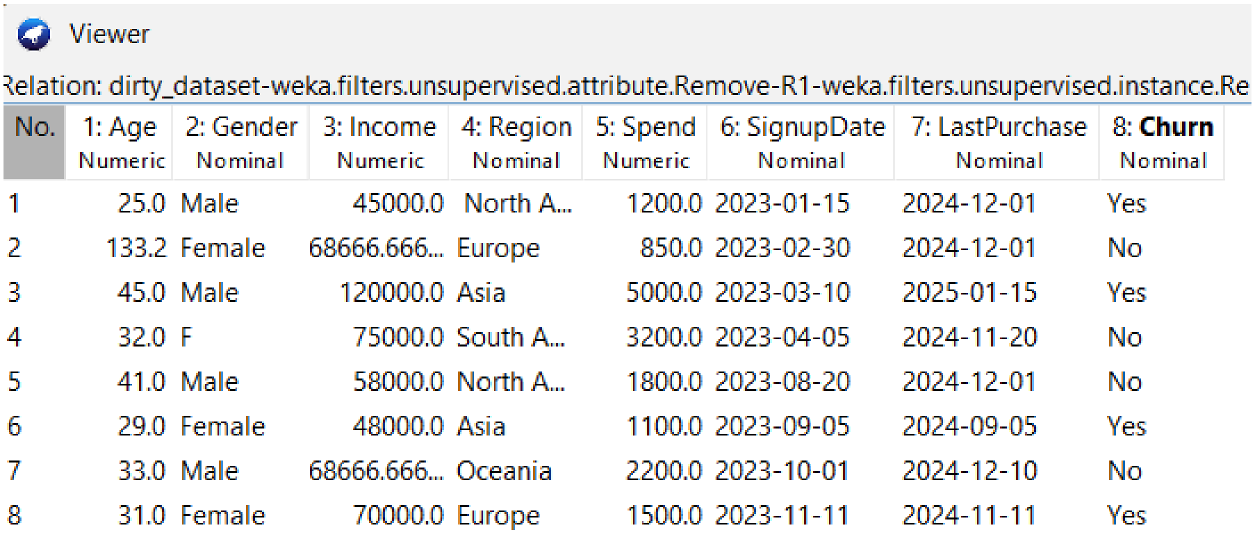


Figure 11: Removing the columns that were added previously after deleting rows with outliers

8. Finalize

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



Viewer

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

Figure 12: Final resulting dataset after preprocessing

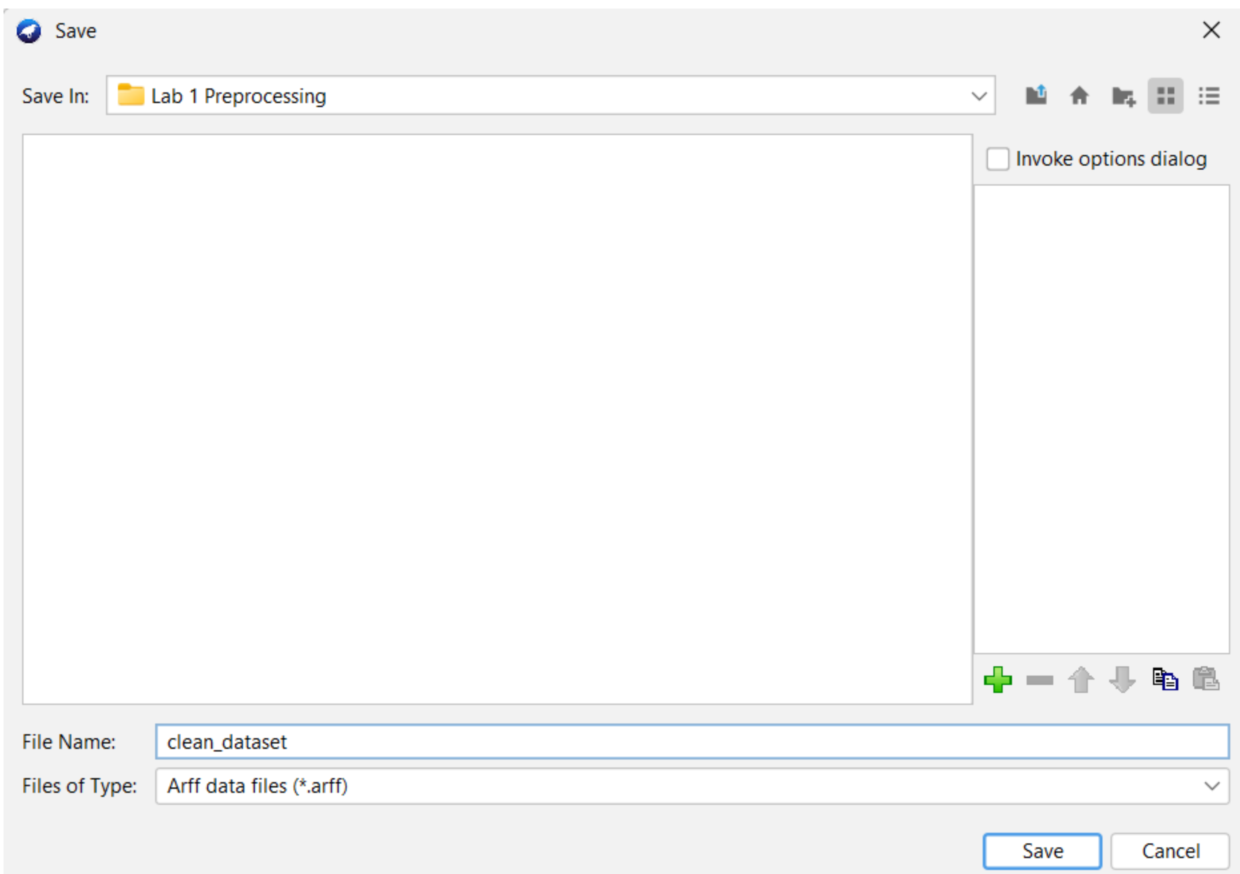


Figure 13: Saving the dataset as clean_dataset in .arff format

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

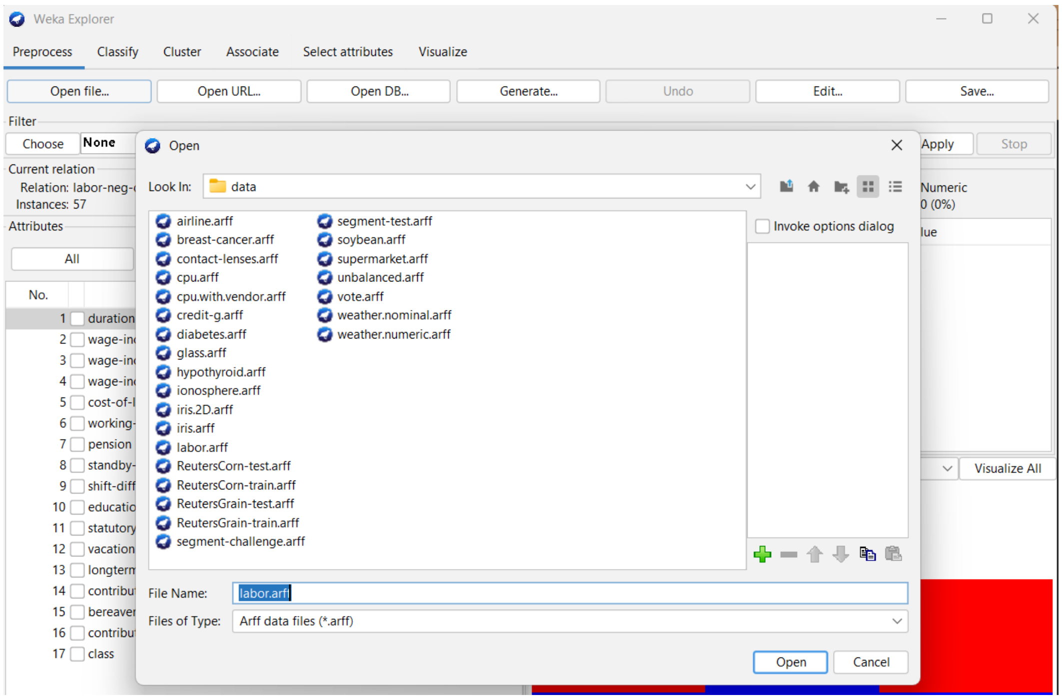


Figure 14: Opening the secondary dataset available in WEKA data

2. Visualize the data

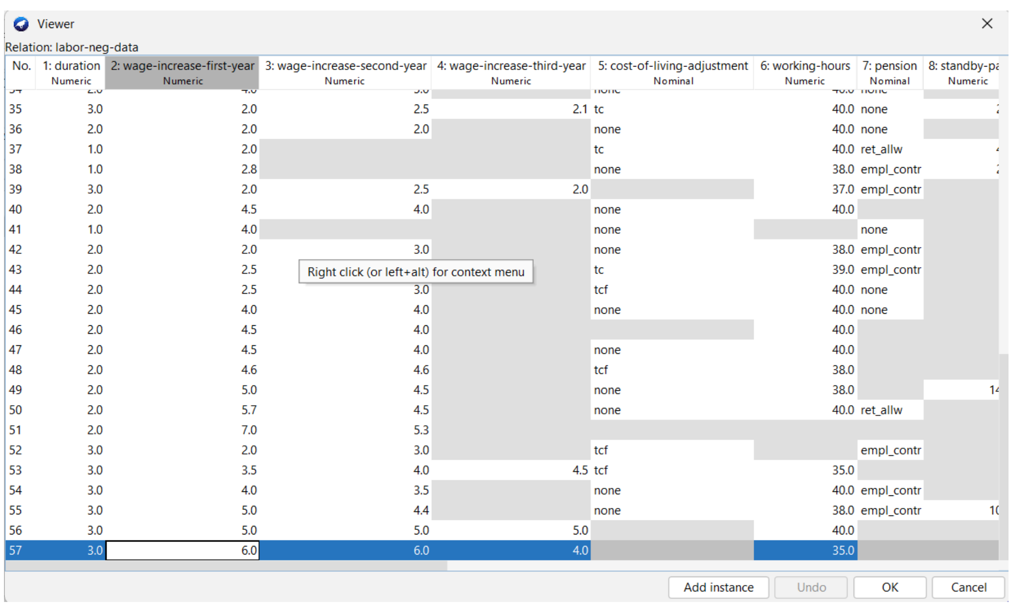


Figure 15: Visualizing the data

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.

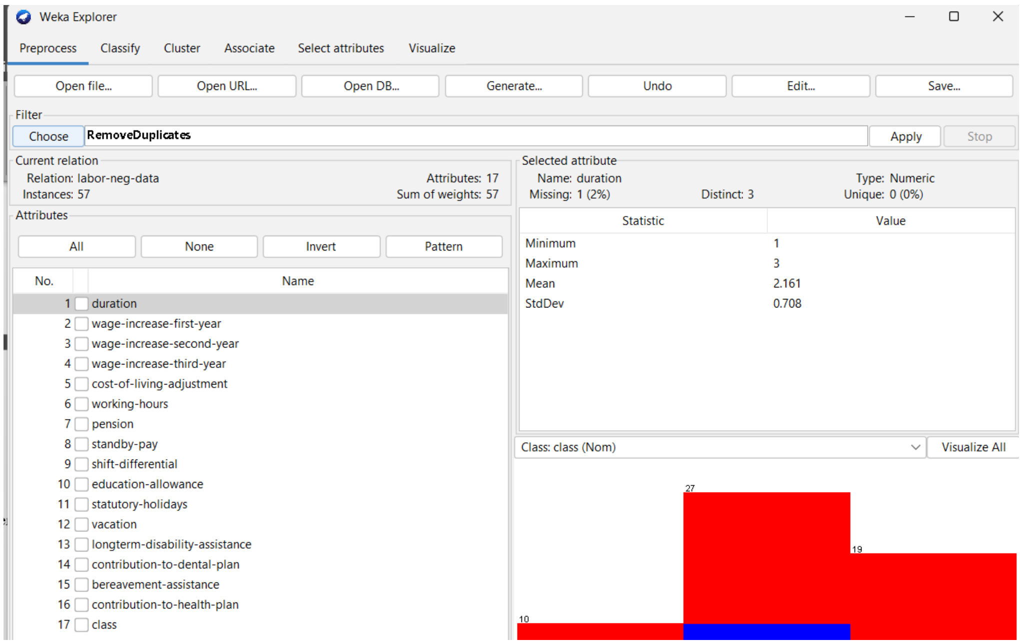


Figure 16: Applying the RemoveDuplicate filter

5. Replace any missing values

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

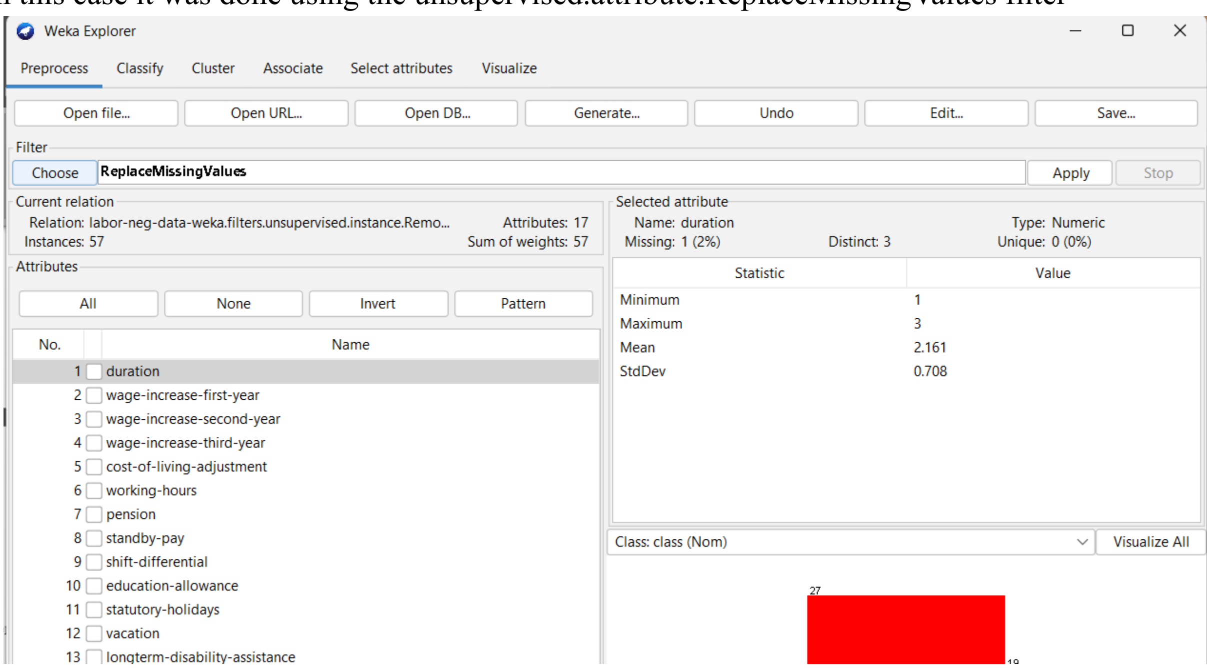


Figure 17: Applying the ReplaceMissingValues filter

6. Convert string into nominal values

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

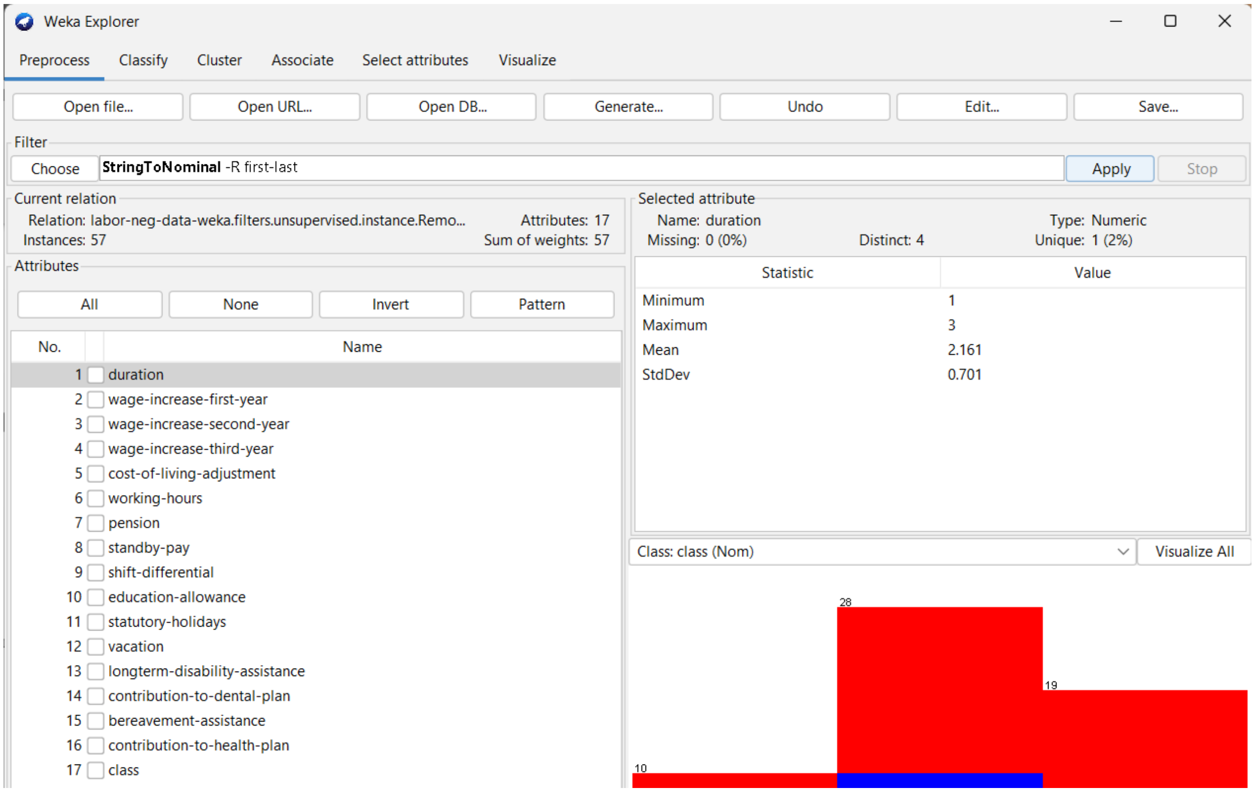


Figure 18: Conversion of String to Nominal Values

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.

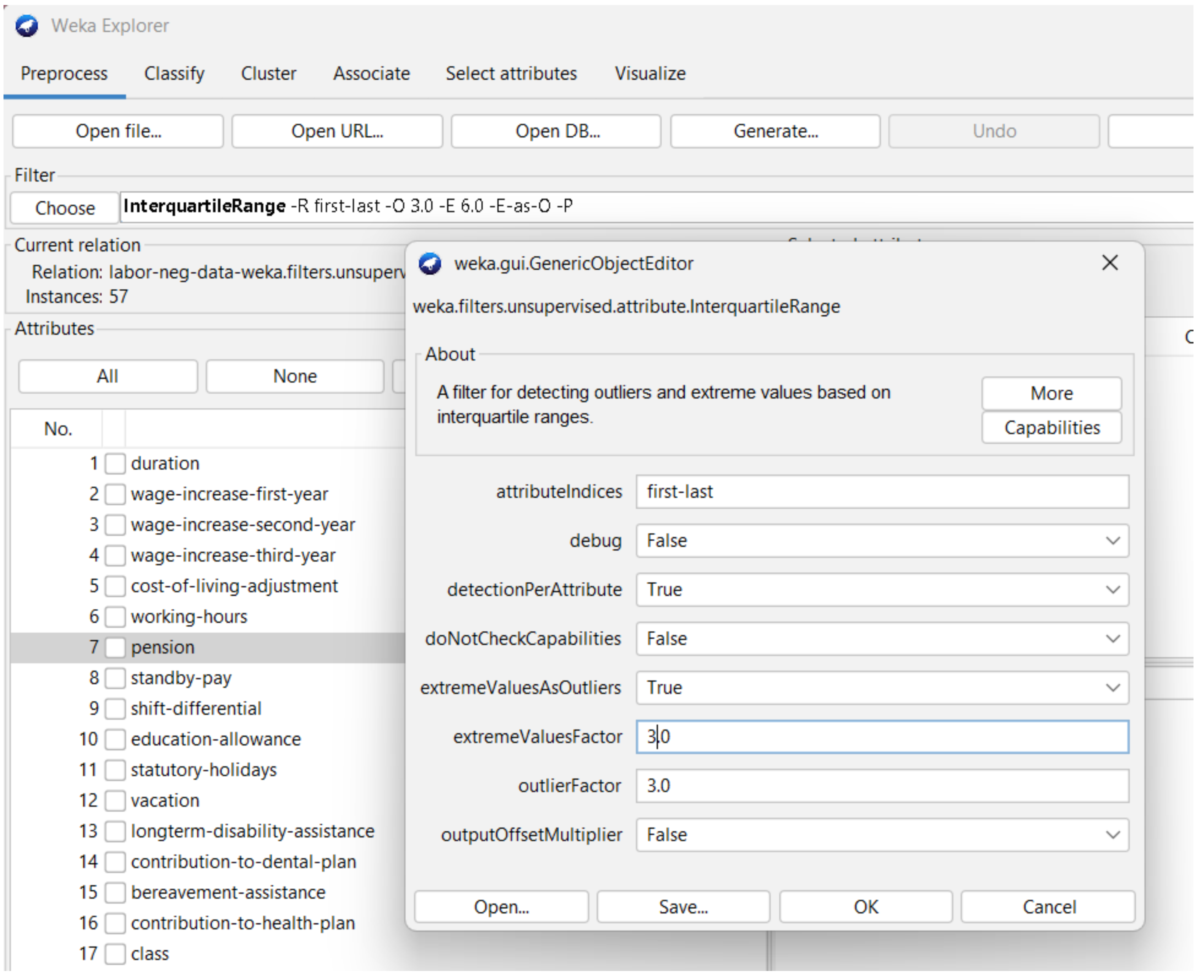


Figure 19: Selecting appropriate setting for InterquartileRange

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.

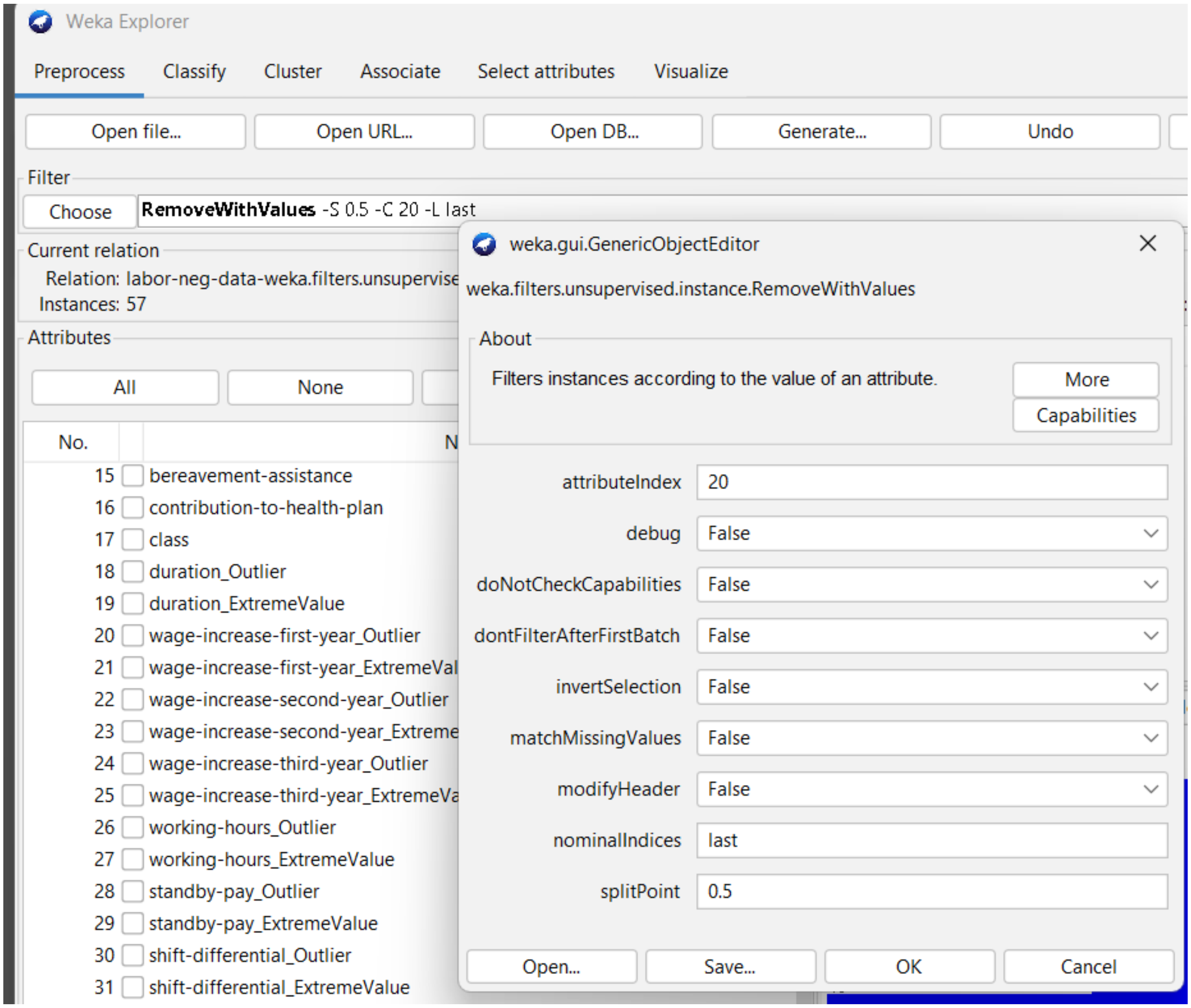


Figure 20: Removing rows containing outliers

7.3. Remove the columns created

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

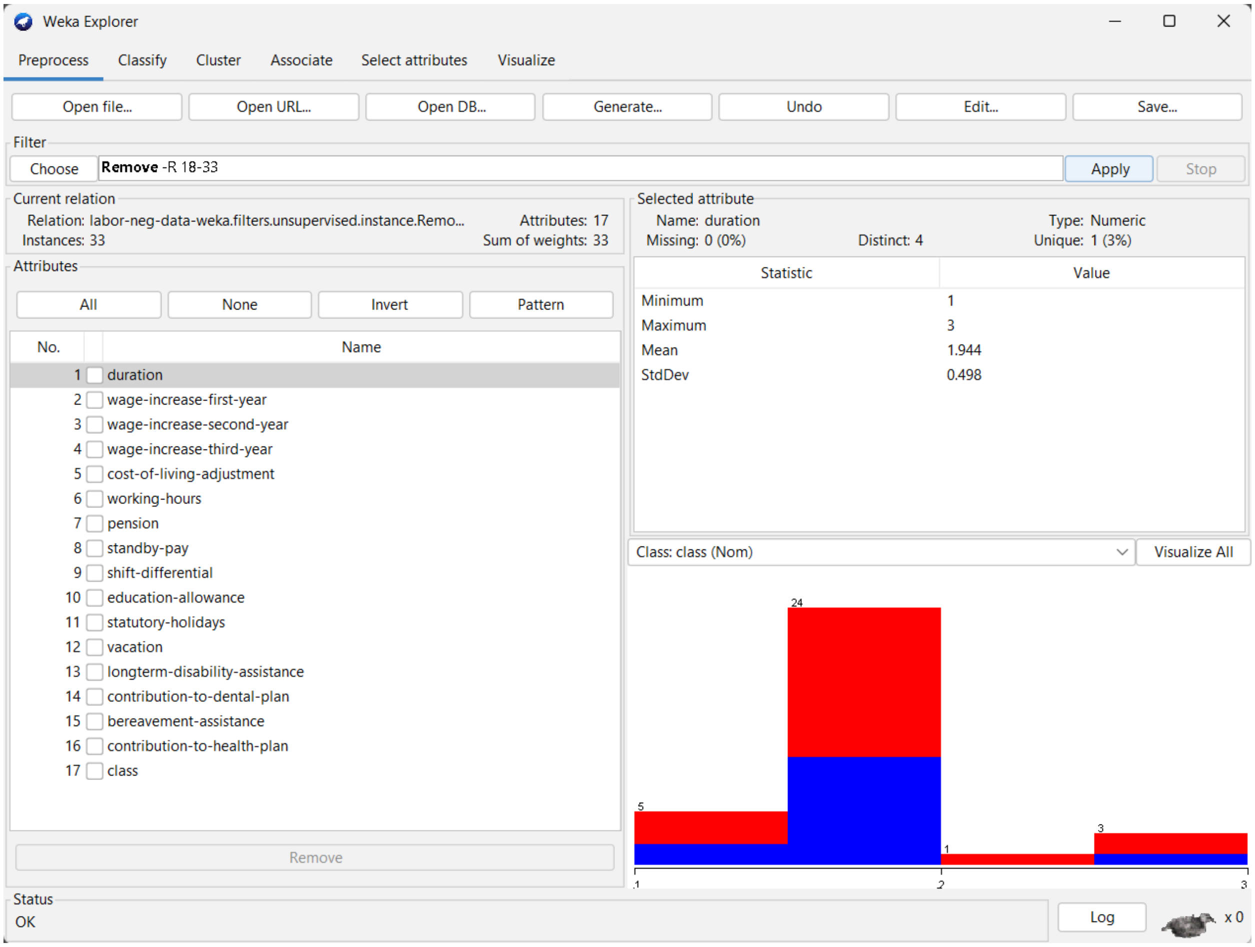
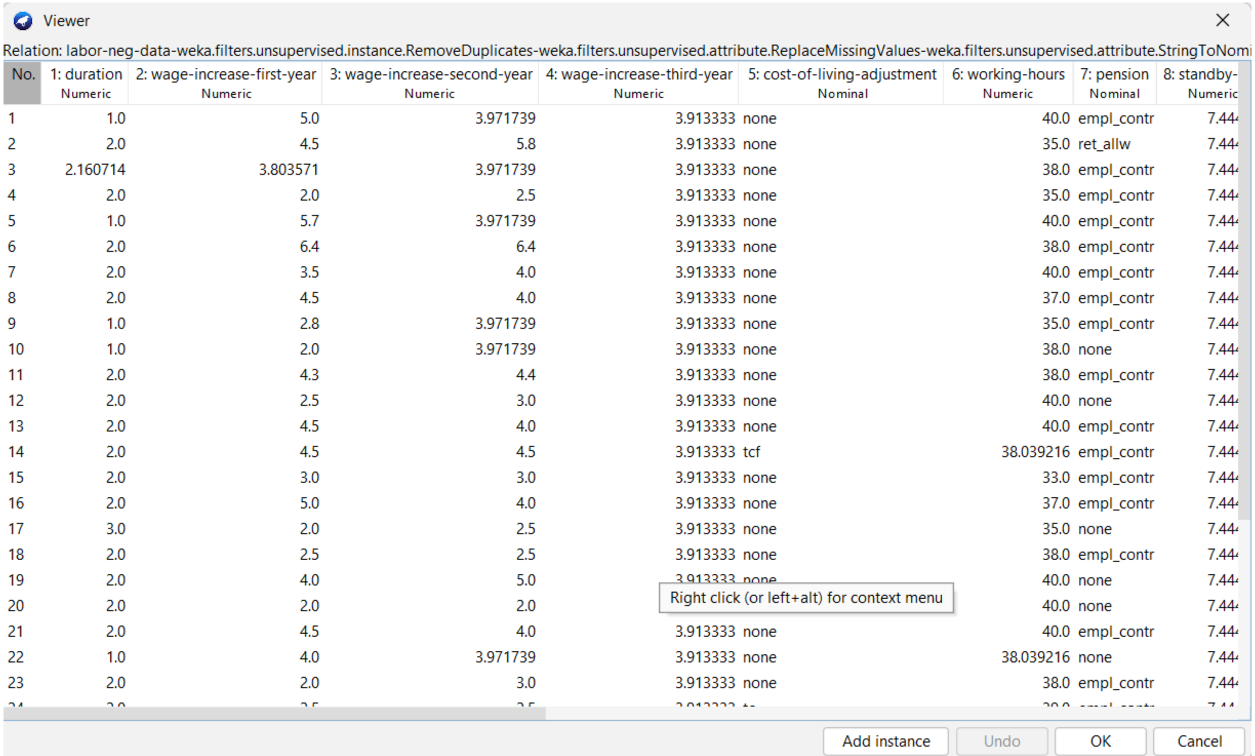


Figure 21: Removal of previously created unwanted columns

8. Finalize

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

Figure 22: Final view of the data

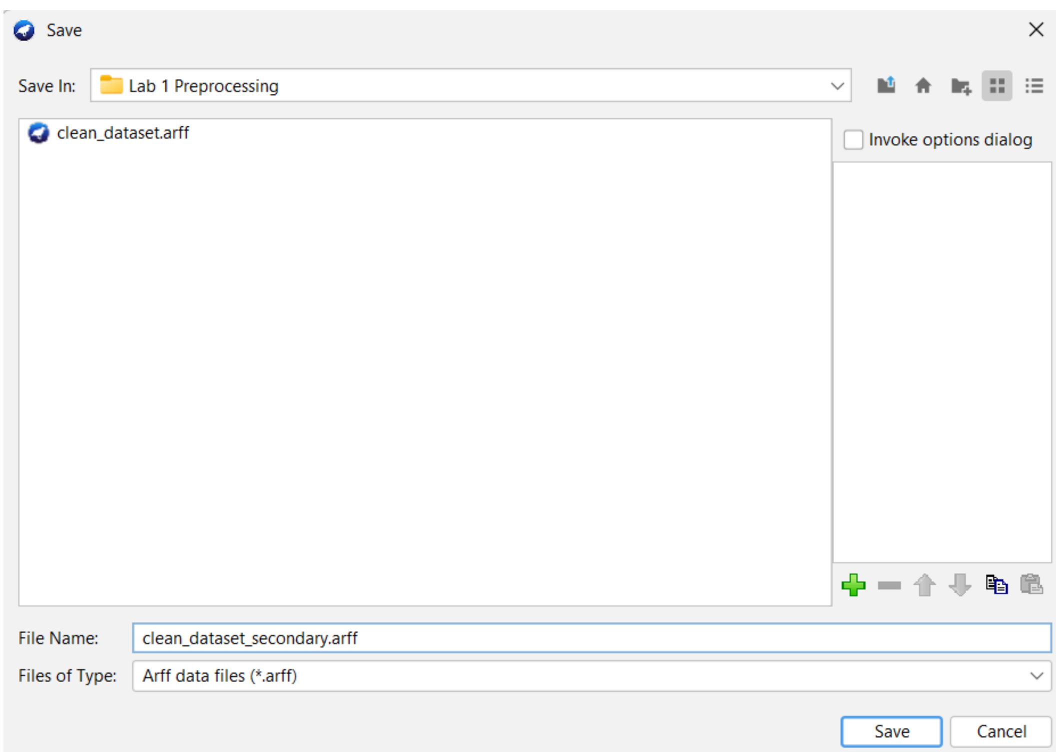


Figure 23: Saving the file as clean dataset in .arff format

Discussion

The preprocessing performed in this lab showed how essential data cleaning is before any data mining or analytical task. The primary dataset contained several typical forms of dirty data such as missing values, duplicated entries, string attributes, and outliers, which required multiple WEKA filters to handle. The removal of unnecessary attributes (like CustomerID) helped reduce noise, while ReplaceMissingValues ensured completeness. Converting strings to nominal values was necessary to prepare the dataset for algorithms that only accept categorical attributes.

Outliers were an important focus of discussion, as WEKA's InterquartileRange filter first tagged them and subsequent RemoveWithValues operations eliminated them. This illustrated how outliers can be systematically detected rather than manually assumed. The secondary dataset required fewer cleaning steps, which highlighted the difference between raw operational data and structured warehouse-ready data. The process demonstrated that preprocessing is often the most time-consuming but impactful stage of data preparation.

Conclusion

This lab successfully demonstrated the full workflow of data preprocessing using WEKA. The primary dataset required extensive cleaning, while the secondary dataset required minimal adjustments, confirming the intended differences between the two. After applying filters for missing values, duplicates, inconsistent formats, and outliers, both datasets were transformed into clean, analysis-ready .arff files. The lab reinforced the importance of preprocessing for accuracy, consistency, and meaningful downstream analysis.