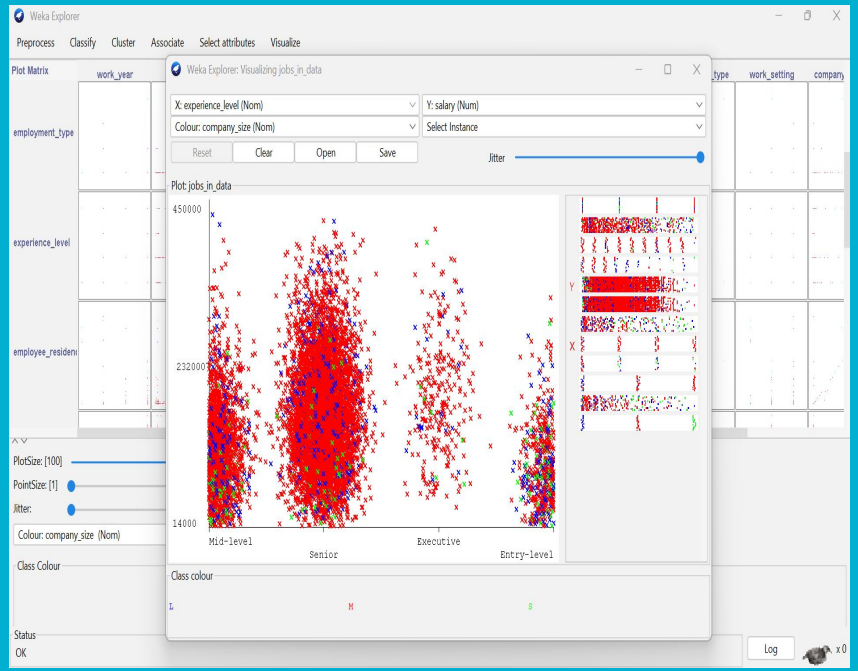
Pre-processing: Jobs and Salaries in Data Science

—

Team Members: Puja Shah & Sanjida Chowdhury

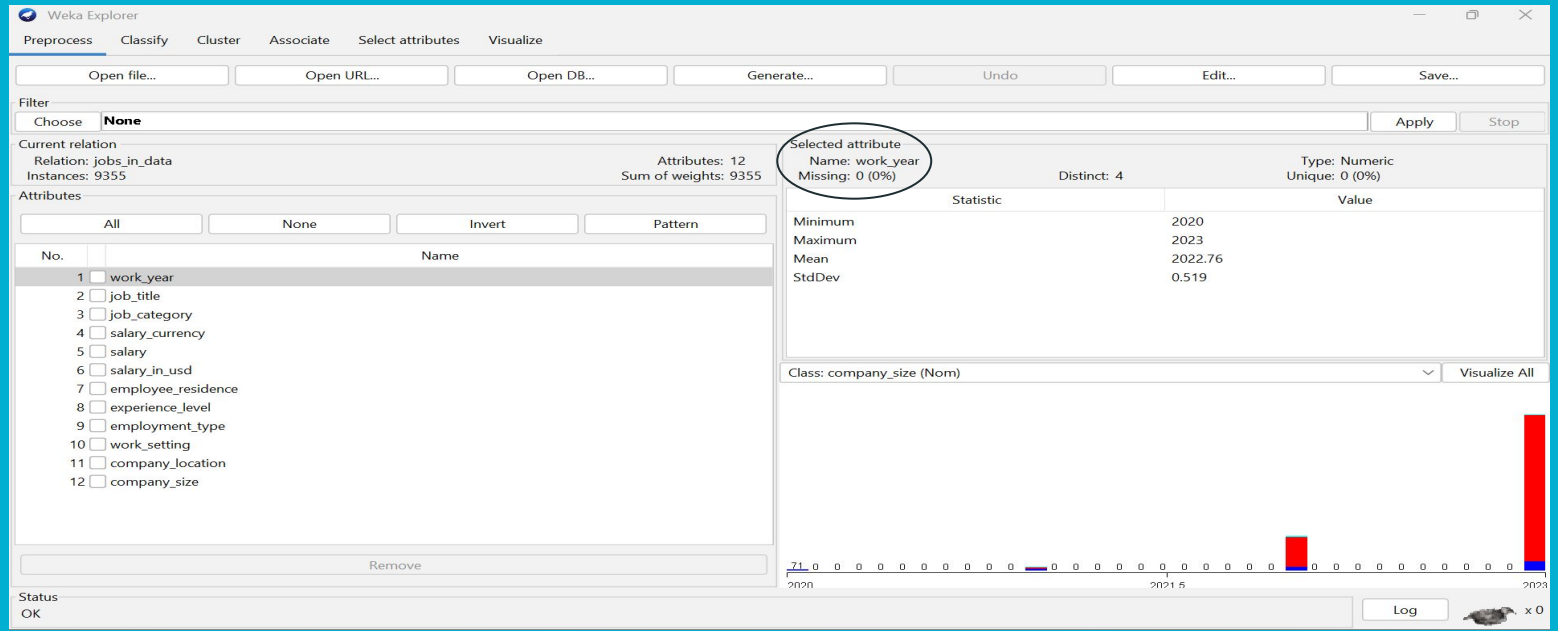
Dataset: “Jobs and Salaries in Data Science”

- The dataset captures a comprehensive snapshot of experienced level in jobs and salaries in the data science filed for year 2023.
- The goal is to explore relationships, identify patterns, and understand factors influencing salary levels in data science.
- Attributes: work year, job title, job category, salary currency, salary, salary in USD, employee residence, experience level, employment type, work setting, company location, and company size.



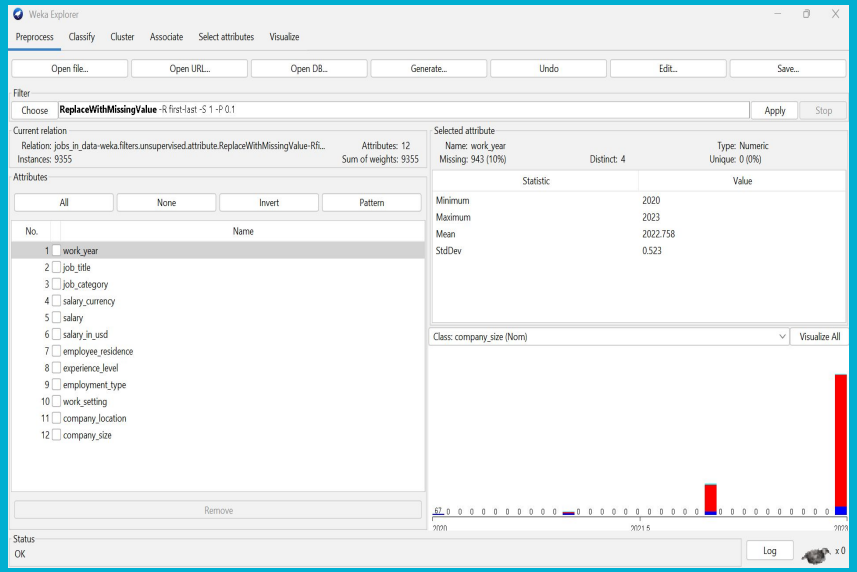
Dealing with Missing Values

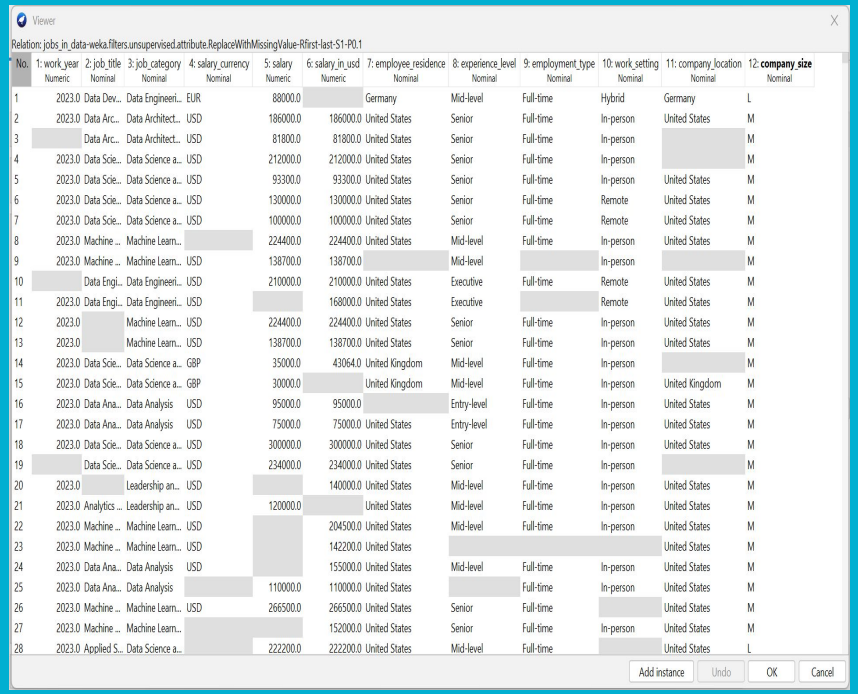
Original dataset without missing values:



Dealing with Missing Values

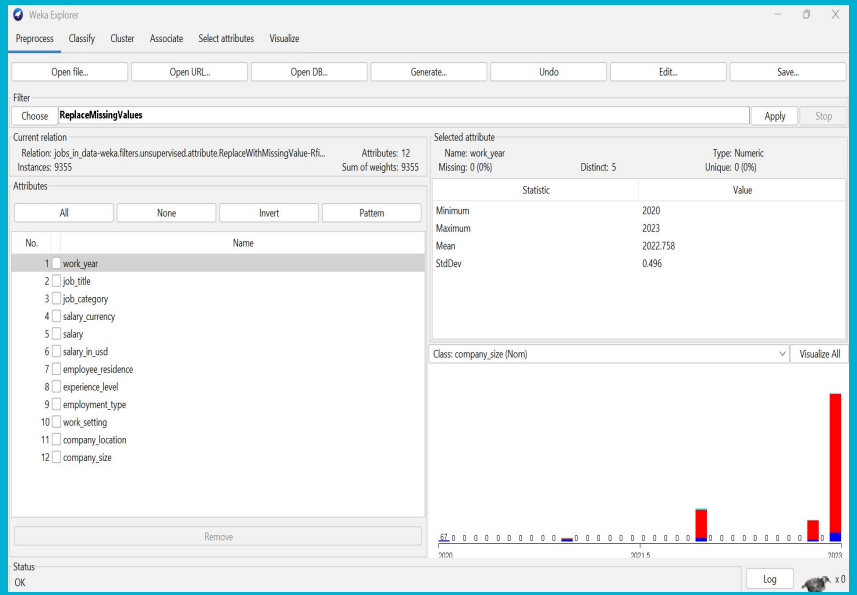
Intentionally created missing values (Replace with missing value):

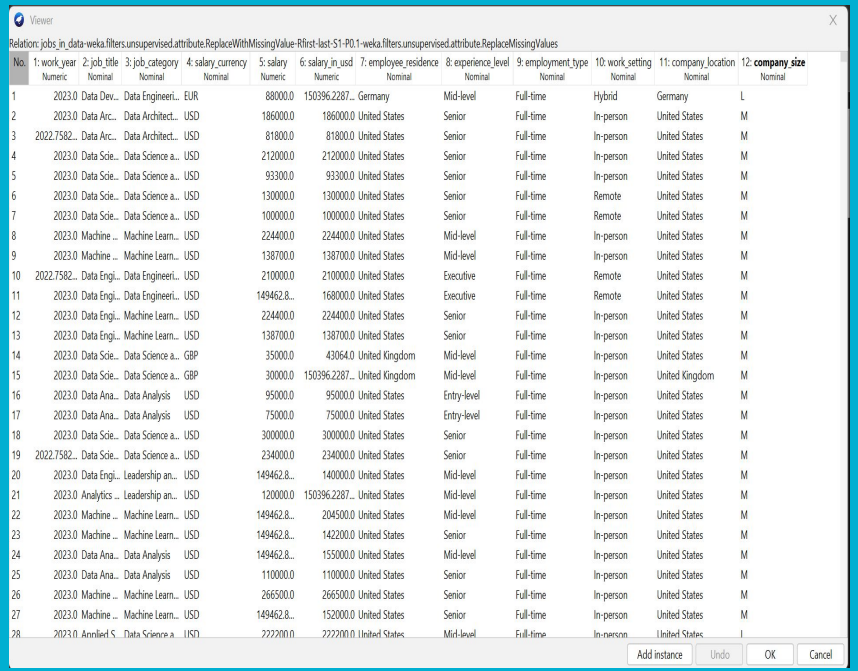




Dealing with Missing Values

Replace missing valves (Imputation with Mean):



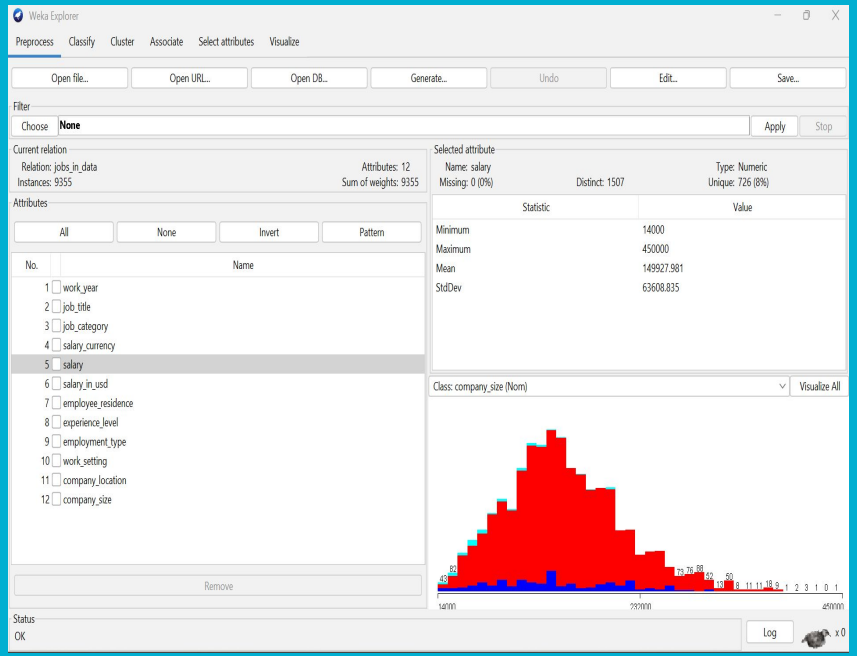


Dealing with Missing Values

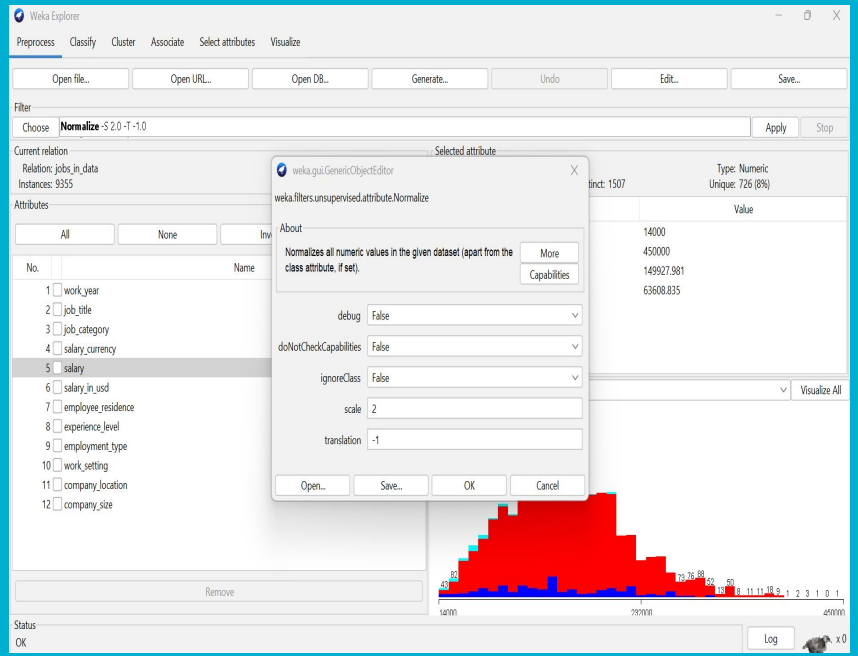
- The intentional creation of missing values allowed us to simulate scenarios where data may be incomplete.
- By deliberately introducing missing values in critical attributes such as salary and experience level, we test the dataset's resilience and its ability to handle gaps in information.
- Missing values were replaced using imputation with the mean of the respective attributes.
- This process ensures robustness testing and prepares the dataset for subsequent preprocessing steps.

Normalization (min-max)

Original data

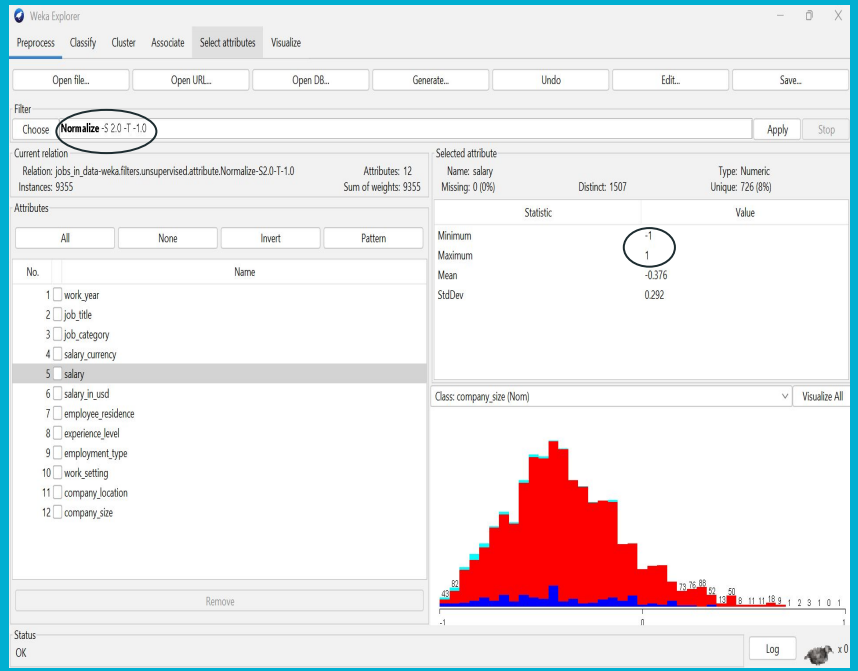


Applying ‘Normalize’ to data



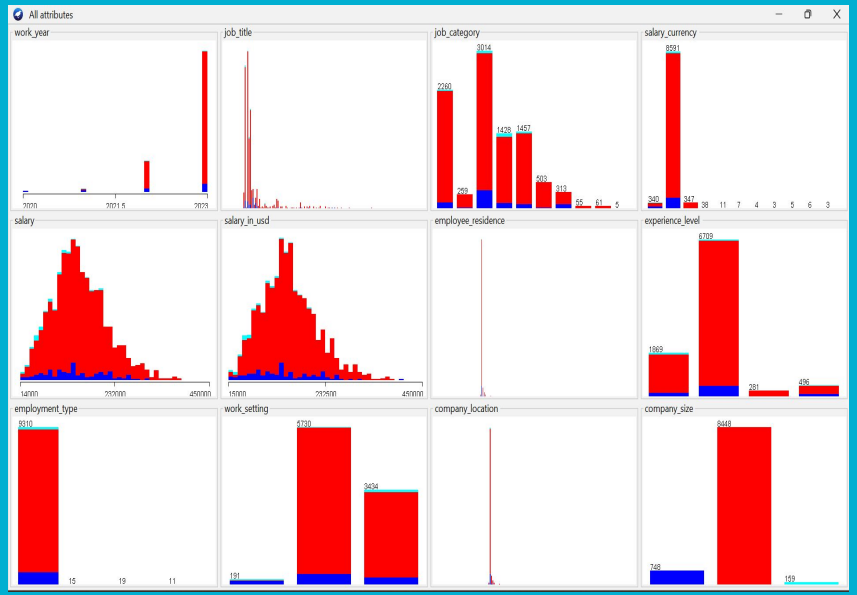
Normalization (min-max)

- Application of min-max scaling to standardize numerical attributes between -1 to 2.

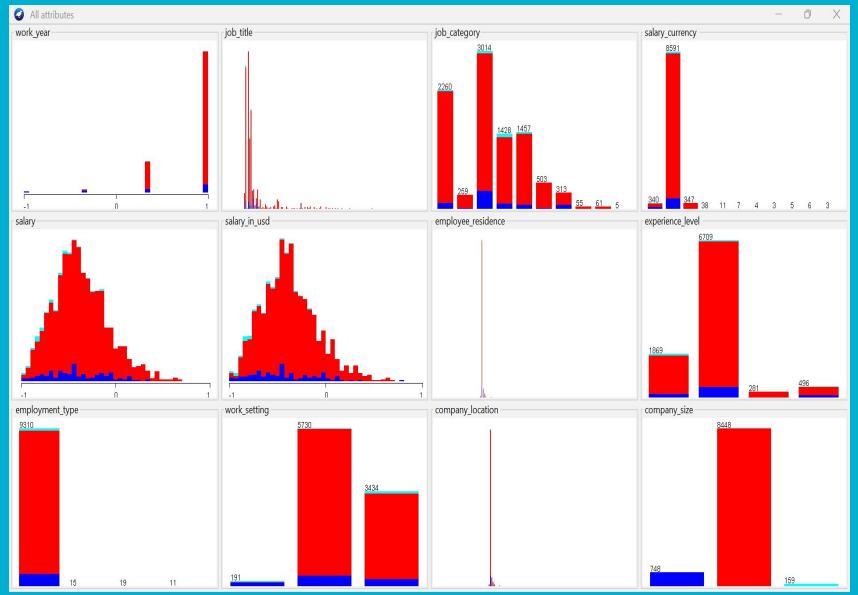


Normalization (min-max)

Orignal

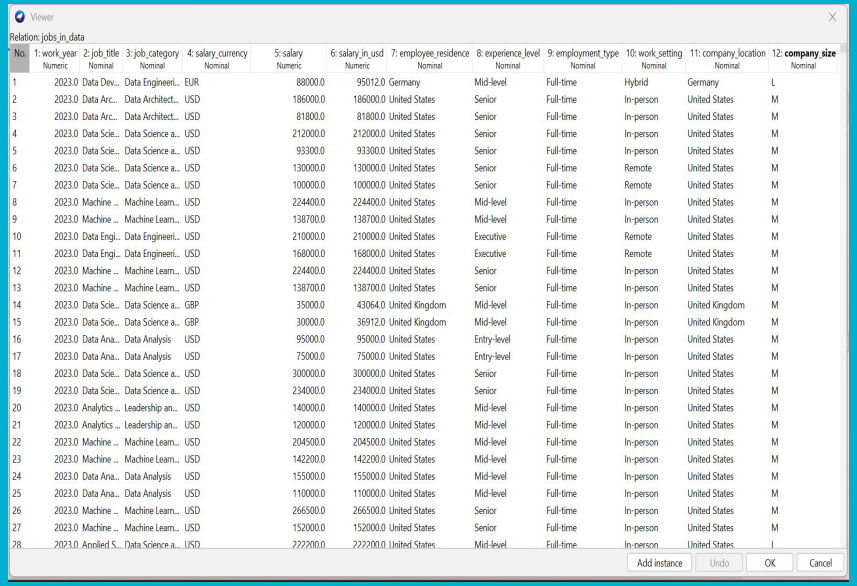


After normalization



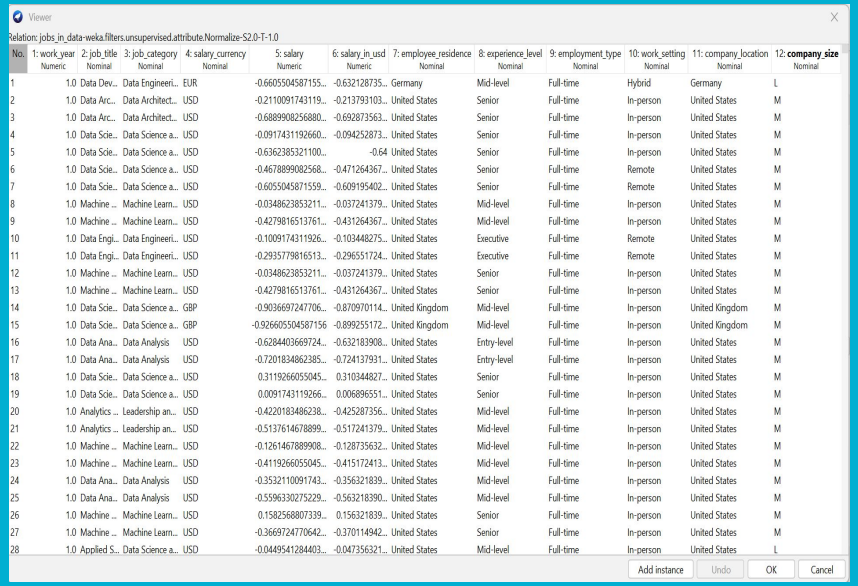
Normalization (min-max)

Orignal



No.	1: work_year Numeric	2: job_title Nominal	3: job_category Nominal	4: salary_currency Nominal	5: salary Numeric	6: salary_in_usd Numeric	7: employee_residence Nominal	8: experience_level Nominal	9: employment_type Nominal	10: work_setting Nominal	11: company_location Nominal	12: company_size Nominal
1	2023.0	Data Dev...	Data Engineeri...	EUR	88000.0	95012.0	Germany	Mid-level	Full-time	Hybrid	Germany	L
2	2023.0	Data Arc...	Data Architect...	USD	186000.0	186000.0	United States	Senior	Full-time	In-person	United States	M
3	2023.0	Data Arc...	Data Architect...	USD	81800.0	81800.0	United States	Senior	Full-time	In-person	United States	M
4	2023.0	Data Scie...	Data Science a...	USD	212000.0	212000.0	United States	Senior	Full-time	In-person	United States	M
5	2023.0	Data Scie...	Data Science a...	USD	93300.0	93300.0	United States	Senior	Full-time	In-person	United States	M
6	2023.0	Data Scie...	Data Science a...	USD	130000.0	130000.0	United States	Senior	Full-time	Remote	United States	M
7	2023.0	Data Scie...	Data Science a...	USD	100000.0	100000.0	United States	Senior	Full-time	Remote	United States	M
8	2023.0	Machine ...	Machine Learn...	USD	224400.0	224400.0	United States	Mid-level	Full-time	In-person	United States	M
9	2023.0	Machine ...	Machine Learn...	USD	138700.0	138700.0	United States	Mid-level	Full-time	In-person	United States	M
10	2023.0	Data Engi...	Data Engineeri...	USD	210000.0	210000.0	United States	Executive	Full-time	Remote	United States	M
11	2023.0	Data Engi...	Data Engineeri...	USD	168000.0	168000.0	United States	Executive	Full-time	Remote	United States	M
12	2023.0	Machine ...	Machine Learn...	USD	224400.0	224400.0	United States	Senior	Full-time	In-person	United States	M
13	2023.0	Machine ...	Machine Learn...	USD	138700.0	138700.0	United States	Senior	Full-time	In-person	United States	M
14	2023.0	Data Scie...	Data Science a...	GBP	35000.0	43064.0	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
15	2023.0	Data Scie...	Data Science a...	GBP	30000.0	36912.0	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
16	2023.0	Data Ana...	Data Analysis	USD	95000.0	95000.0	United States	Entry-level	Full-time	In-person	United States	M
17	2023.0	Data Ana...	Data Analysis	USD	75000.0	75000.0	United States	Entry-level	Full-time	In-person	United States	M
18	2023.0	Data Scie...	Data Science a...	USD	300000.0	300000.0	United States	Senior	Full-time	In-person	United States	M
19	2023.0	Data Scie...	Data Science a...	USD	234000.0	234000.0	United States	Senior	Full-time	In-person	United States	M
20	2023.0	Analytics ...	Leadership an...	USD	140000.0	140000.0	United States	Mid-level	Full-time	In-person	United States	M
21	2023.0	Analytics ...	Leadership an...	USD	120000.0	120000.0	United States	Mid-level	Full-time	In-person	United States	M
22	2023.0	Machine ...	Machine Learn...	USD	204500.0	204500.0	United States	Mid-level	Full-time	In-person	United States	M
23	2023.0	Machine ...	Machine Learn...	USD	142200.0	142200.0	United States	Mid-level	Full-time	In-person	United States	M
24	2023.0	Data Ana...	Data Analysis	USD	155000.0	155000.0	United States	Mid-level	Full-time	In-person	United States	M
25	2023.0	Data Ana...	Data Analysis	USD	110000.0	110000.0	United States	Mid-level	Full-time	In-person	United States	M
26	2023.0	Machine ...	Machine Learn...	USD	266500.0	266500.0	United States	Senior	Full-time	In-person	United States	M
27	2023.0	Machine ...	Machine Learn...	USD	152000.0	152000.0	United States	Senior	Full-time	In-person	United States	M
28	2023.0	Applied S...	Data Science a...	USD	222200.0	222200.0	United States	Mid-level	Full-time	In-person	United States	L

After normalization



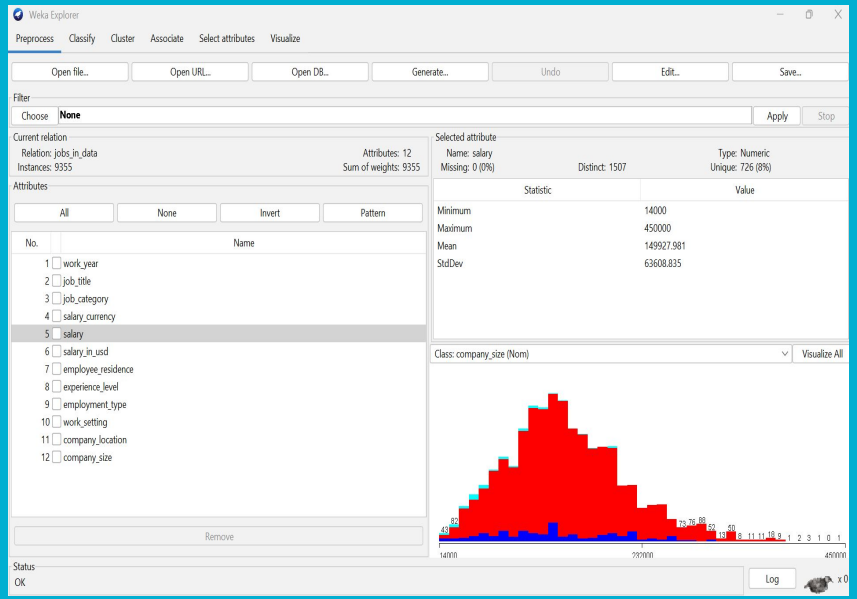
No.	1: work_year Numeric	2: job_title Nominal	3: job_category Nominal	4: salary_currency Nominal	5: salary Numeric	6: salary_in_usd Numeric	7: employee_residence Nominal	8: experience_level Nominal	9: employment_type Nominal	10: work_setting Nominal	11: company_location Nominal	12: company_size Nominal
1	1.0	Data Dev...	Data Engineeri...	EUR	-0.6605504587155...	-0.632128735...	Germany	Mid-level	Full-time	Hybrid	Germany	L
2	1.0	Data Arc...	Data Architect...	USD	-0.2110091743119...	-0.213793103...	United States	Senior	Full-time	In-person	United States	M
3	1.0	Data Arc...	Data Architect...	USD	-0.6889908256880...	-0.692873563...	United States	Senior	Full-time	In-person	United States	M
4	1.0	Data Scie...	Data Science a...	USD	-0.0917431192660...	-0.094252873...	United States	Senior	Full-time	In-person	United States	M
5	1.0	Data Scie...	Data Science a...	USD	-0.6362385321100...	-0.64	United States	Senior	Full-time	In-person	United States	M
6	1.0	Data Scie...	Data Science a...	USD	-0.4678899082568...	-0.471264367...	United States	Senior	Full-time	Remote	United States	M
7	1.0	Data Scie...	Data Science a...	USD	-0.6055045871559...	-0.609195402...	United States	Senior	Full-time	Remote	United States	M
8	1.0	Machine ...	Machine Learn...	USD	-0.0348623853211...	-0.037241379...	United States	Mid-level	Full-time	In-person	United States	M
9	1.0	Machine ...	Machine Learn...	USD	-0.4279816513761...	-0.431264367...	United States	Mid-level	Full-time	In-person	United States	M
10	1.0	Data Engi...	Data Engineeri...	USD	-0.1009174311926...	-0.103448275...	United States	Executive	Full-time	Remote	United States	M
11	1.0	Data Engi...	Data Engineeri...	USD	-0.2935779816513...	-0.296551724...	United States	Executive	Full-time	Remote	United States	M
12	1.0	Machine ...	Machine Learn...	USD	-0.0348623853211...	-0.037241379...	United States	Senior	Full-time	In-person	United States	M
13	1.0	Machine ...	Machine Learn...	USD	-0.4279816513761...	-0.431264367...	United States	Senior	Full-time	In-person	United States	M
14	1.0	Data Scie...	Data Science a...	GBP	-0.9036697247706...	-0.870970114...	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
15	1.0	Data Scie...	Data Science a...	GBP	-0.926605504587156	-0.899255172...	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
16	1.0	Data Ana...	Data Analysis	USD	-0.6284403669724...	-0.632183908...	United States	Entry-level	Full-time	In-person	United States	M
17	1.0	Data Ana...	Data Analysis	USD	-0.7201834862385...	-0.724137931...	United States	Entry-level	Full-time	In-person	United States	M
18	1.0	Data Scie...	Data Science a...	USD	0.3119266055045...	0.310344827...	United States	Senior	Full-time	In-person	United States	M
19	1.0	Data Scie...	Data Science a...	USD	0.0091743119266...	0.006896551...	United States	Senior	Full-time	In-person	United States	M
20	1.0	Analytics ...	Leadership an...	USD	-0.4220183486238...	-0.425287356...	United States	Mid-level	Full-time	In-person	United States	M
21	1.0	Analytics ...	Leadership an...	USD	-0.5137614678899...	-0.517241379...	United States	Mid-level	Full-time	In-person	United States	M
22	1.0	Machine ...	Machine Learn...	USD	-0.1261467889908...	-0.128735632...	United States	Mid-level	Full-time	In-person	United States	M
23	1.0	Machine ...	Machine Learn...	USD	-0.4119266055045...	-0.415172413...	United States	Mid-level	Full-time	In-person	United States	M
24	1.0	Data Ana...	Data Analysis	USD	-0.3532110091743...	-0.356321839...	United States	Mid-level	Full-time	In-person	United States	M
25	1.0	Data Ana...	Data Analysis	USD	-0.5596330275229...	-0.563218390...	United States	Mid-level	Full-time	In-person	United States	M
26	1.0	Machine ...	Machine Learn...	USD	0.1582568807339...	0.156321839...	United States	Senior	Full-time	In-person	United States	M
27	1.0	Machine ...	Machine Learn...	USD	-0.3669724770642...	-0.370114942...	United States	Senior	Full-time	In-person	United States	M
28	1.0	Applied S...	Data Science a...	USD	-0.0449541284403...	-0.047356321...	United States	Mid-level	Full-time	In-person	United States	L

Normalization (min-max)

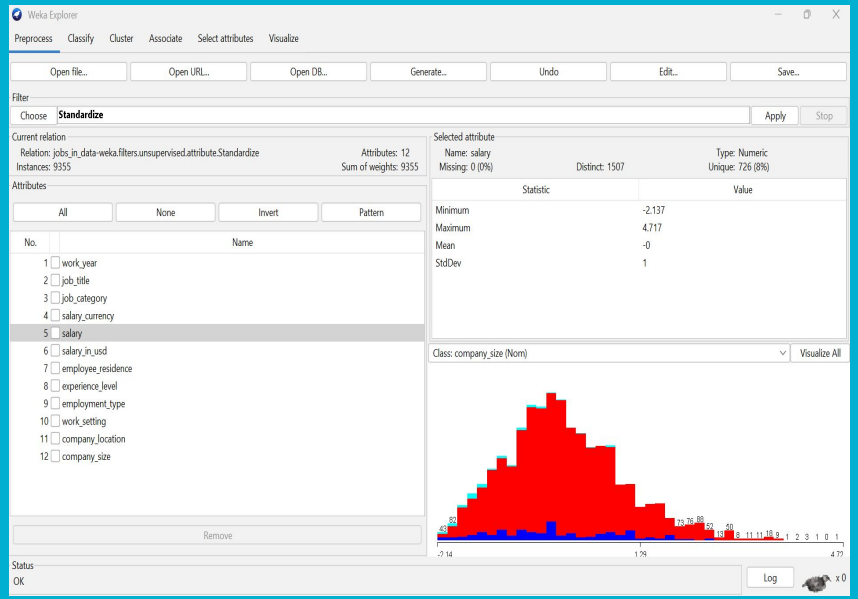
- The original dataset contains numerical features with varying scales, which may impact the performance of certain algorithms.
- Min-Max normalization is applied to ensure that numerical features are standardized between -1 and 2. This helps prevent attributes with larger scales from dominating the analysis, making the dataset more suitable for a wide range of algorithms.

Normalization (z-score)

Original

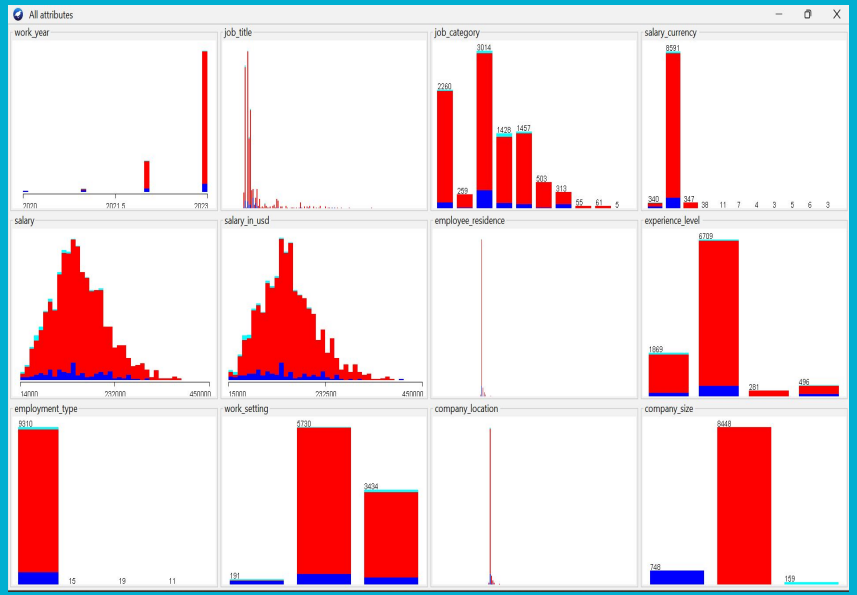


After applying ‘Standardize’

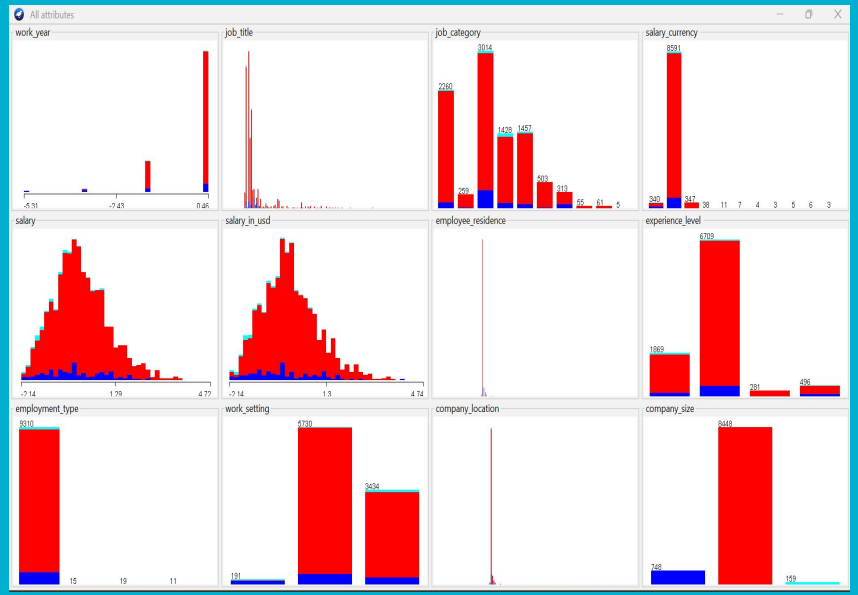


Normalization (z-score)

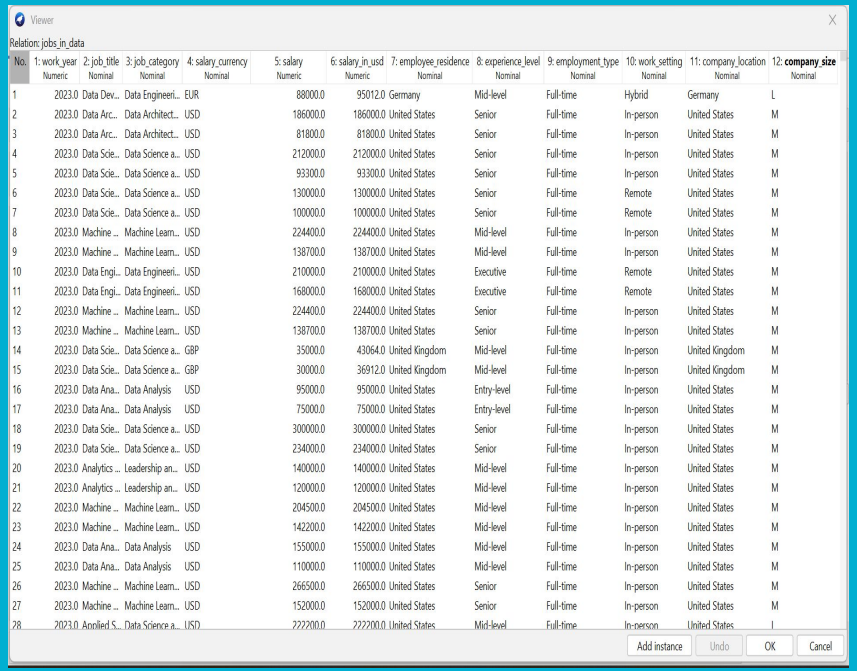
Original



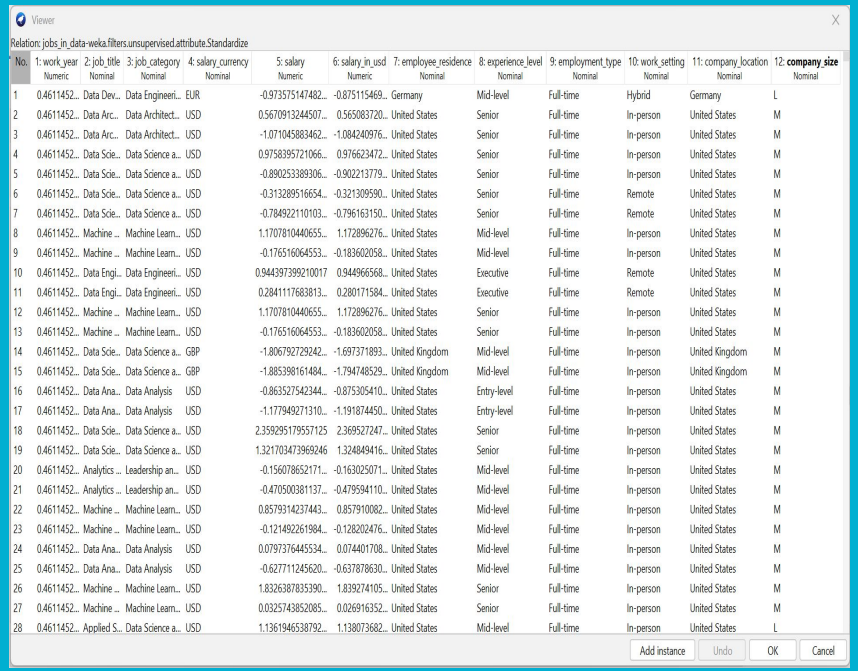
After applying ‘Standardize’



Normalization (z-score)



No.	1: work_year Numeric	2: job_title Nominal	3: job_category Nominal	4: salary_currency Nominal	5: salary Numeric	6: salary_in_usd Numeric	7: employee_residence Nominal	8: experience_level Nominal	9: employment_type Nominal	10: work_setting Nominal	11: company_location Nominal	12: **company_size** Nominal
1	2023.0	Data Dev...	Data Engineeri...	EUR	88000.0	95012.0	Germany	Mid-level	Full-time	Hybrid	Germany	L
2	2023.0	Data Arc...	Data Architect...	USD	186000.0	186000.0	United States	Senior	Full-time	In-person	United States	M
3	2023.0	Data Arc...	Data Architect...	USD	81800.0	81800.0	United States	Senior	Full-time	In-person	United States	M
4	2023.0	Data Scie...	Data Science a...	USD	212000.0	212000.0	United States	Senior	Full-time	In-person	United States	M
5	2023.0	Data Scie...	Data Science a...	USD	93300.0	93300.0	United States	Senior	Full-time	In-person	United States	M
6	2023.0	Data Scie...	Data Science a...	USD	130000.0	130000.0	United States	Senior	Full-time	Remote	United States	M
7	2023.0	Data Scie...	Data Science a...	USD	100000.0	100000.0	United States	Senior	Full-time	Remote	United States	M
8	2023.0	Machine ...	Machine Learn...	USD	224400.0	224400.0	United States	Mid-level	Full-time	In-person	United States	M
9	2023.0	Machine ...	Machine Learn...	USD	138700.0	138700.0	United States	Mid-level	Full-time	In-person	United States	M
10	2023.0	Data Engi...	Data Engineeri...	USD	210000.0	210000.0	United States	Executive	Full-time	Remote	United States	M
11	2023.0	Data Engi...	Data Engineeri...	USD	168000.0	168000.0	United States	Executive	Full-time	Remote	United States	M
12	2023.0	Machine ...	Machine Learn...	USD	224400.0	224400.0	United States	Senior	Full-time	In-person	United States	M
13	2023.0	Machine ...	Machine Learn...	USD	138700.0	138700.0	United States	Senior	Full-time	In-person	United States	M
14	2023.0	Data Scie...	Data Science a...	GBP	35000.0	43064.0	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
15	2023.0	Data Scie...	Data Science a...	GBP	30000.0	36912.0	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
16	2023.0	Data Ana...	Data Analysis	USD	95000.0	95000.0	United States	Entry-level	Full-time	In-person	United States	M
17	2023.0	Data Ana...	Data Analysis	USD	75000.0	75000.0	United States	Entry-level	Full-time	In-person	United States	M
18	2023.0	Data Scie...	Data Science a...	USD	300000.0	300000.0	United States	Senior	Full-time	In-person	United States	M
19	2023.0	Data Scie...	Data Science a...	USD	234000.0	234000.0	United States	Senior	Full-time	In-person	United States	M
20	2023.0	Analytics ...	Leadership an...	USD	140000.0	140000.0	United States	Mid-level	Full-time	In-person	United States	M
21	2023.0	Analytics ...	Leadership an...	USD	120000.0	120000.0	United States	Mid-level	Full-time	In-person	United States	M
22	2023.0	Machine ...	Machine Learn...	USD	204500.0	204500.0	United States	Mid-level	Full-time	In-person	United States	M
23	2023.0	Machine ...	Machine Learn...	USD	142200.0	142200.0	United States	Mid-level	Full-time	In-person	United States	M
24	2023.0	Data Ana...	Data Analysis	USD	155000.0	155000.0	United States	Mid-level	Full-time	In-person	United States	M
25	2023.0	Data Ana...	Data Analysis	USD	110000.0	110000.0	United States	Mid-level	Full-time	In-person	United States	M
26	2023.0	Machine ...	Machine Learn...	USD	266500.0	266500.0	United States	Senior	Full-time	In-person	United States	M
27	2023.0	Machine ...	Machine Learn...	USD	152000.0	152000.0	United States	Senior	Full-time	In-person	United States	M
28	2023.0	Applied S...	Data Science a...	USD	222200.0	222200.0	United States	Mid-level	Full-time	In-person	United States	L



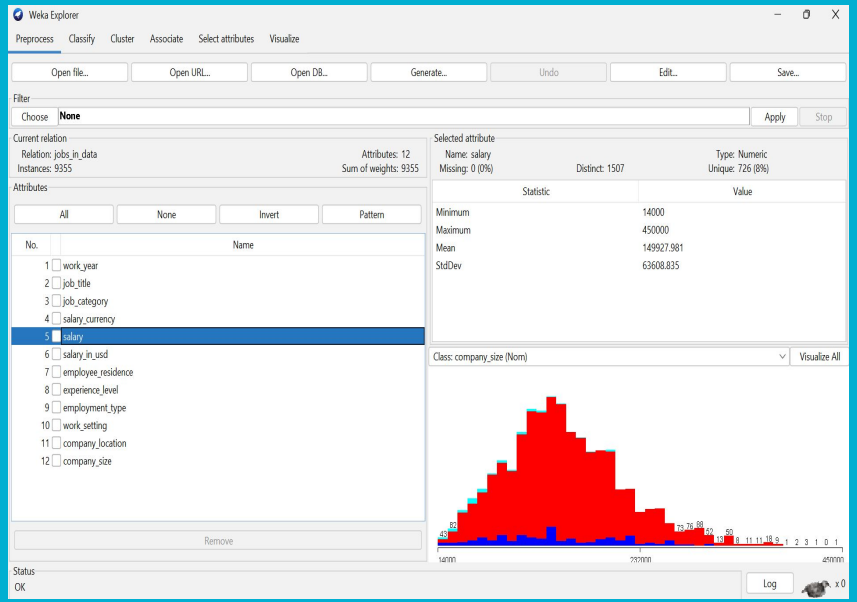
No.	1: work_year Numeric	2: job_title Nominal	3: job_category Nominal	4: salary_currency Nominal	5: salary Numeric	6: salary_in_usd Numeric	7: employee_residence Nominal	8: experience_level Nominal	9: employment_type Nominal	10: work_setting Nominal	11: company_location Nominal	12: **company_size** Nominal
1	0.4611452...	Data Dev...	Data Engineeri...	EUR	-0.973575147482...	-0.875115469...	Germany	Mid-level	Full-time	Hybrid	Germany	L
2	0.4611452...	Data Arc...	Data Architect...	USD	0.5670913244507...	0.565083720...	United States	Senior	Full-time	In-person	United States	M
3	0.4611452...	Data Arc...	Data Architect...	USD	-1.071045883462...	-1.084240976...	United States	Senior	Full-time	In-person	United States	M
4	0.4611452...	Data Scie...	Data Science a...	USD	0.9758395721066...	0.976623472...	United States	Senior	Full-time	In-person	United States	M
5	0.4611452...	Data Scie...	Data Science a...	USD	-0.890253389306...	-0.902213779...	United States	Senior	Full-time	In-person	United States	M
6	0.4611452...	Data Scie...	Data Science a...	USD	-0.313289516654...	-0.321309590...	United States	Senior	Full-time	Remote	United States	M
7	0.4611452...	Data Scie...	Data Science a...	USD	-0.784922110103...	-0.796163150...	United States	Senior	Full-time	Remote	United States	M
8	0.4611452...	Machine ...	Machine Learn...	USD	1.1707810440655...	1.172896276...	United States	Mid-level	Full-time	In-person	United States	M
9	0.4611452...	Machine ...	Machine Learn...	USD	-0.176516064553...	-0.183602058...	United States	Mid-level	Full-time	In-person	United States	M
10	0.4611452...	Data Engi...	Data Engineeri...	USD	0.944397399210017	0.944966568...	United States	Executive	Full-time	Remote	United States	M
11	0.4611452...	Data Engi...	Data Engineeri...	USD	0.2841117683813...	0.280171584...	United States	Executive	Full-time	Remote	United States	M
12	0.4611452...	Machine ...	Machine Learn...	USD	1.1707810440655...	1.172896276...	United States	Senior	Full-time	In-person	United States	M
13	0.4611452...	Machine ...	Machine Learn...	USD	-0.176516064553...	-0.183602058...	United States	Senior	Full-time	In-person	United States	M
14	0.4611452...	Data Scie...	Data Science a...	GBP	-1.806792729242...	-1.697371893...	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
15	0.4611452...	Data Scie...	Data Science a...	GBP	-1.885398161484...	-1.794748529...	United Kingdom	Mid-level	Full-time	In-person	United Kingdom	M
16	0.4611452...	Data Ana...	Data Analysis	USD	-0.863527542344...	-0.875305410...	United States	Entry-level	Full-time	In-person	United States	M
17	0.4611452...	Data Ana...	Data Analysis	USD	-1.177949271310...	-1.191874450...	United States	Entry-level	Full-time	In-person	United States	M
18	0.4611452...	Data Scie...	Data Science a...	USD	2.359295179557125	2.369527247...	United States	Senior	Full-time	In-person	United States	M
19	0.4611452...	Data Scie...	Data Science a...	USD	1.3217034739692 46	1.324849416...	United States	Senior	Full-time	In-person	United States	M
20	0.4611452...	Analytics ...	Leadership an...	USD	-0.156078652171...	-0.163025071...	United States	Mid-level	Full-time	In-person	United States	M
21	0.4611452...	Analytics ...	Leadership an...	USD	-0.470500381137...	-0.479594110...	United States	Mid-level	Full-time	In-person	United States	M
22	0.4611452...	Machine ...	Machine Learn...	USD	0.857931423744 3...	0.857910082...	United States	Mid-level	Full-time	In-person	United States	M
23	0.4611452...	Machine ...	Machine Learn...	USD	-0.121492261984...	-0.128202476...	United States	Mid-level	Full-time	In-person	United States	M
24	0.4611452...	Data Ana...	Data Analysis	USD	0.0797376445534...	0.074401708...	United States	Mid-level	Full-time	In-person	United States	M
25	0.4611452...	Data Ana...	Data Analysis	USD	-0.627711245620...	-0.637878630...	United States	Mid-level	Full-time	In-person	United States	M
26	0.4611452...	Machine ...	Machine Learn...	USD	1.8326387835390...	1.839274105...	United States	Senior	Full-time	In-person	United States	M
27	0.4611452...	Machine ...	Machine Learn...	USD	0.0325743852085...	0.026916352...	United States	Senior	Full-time	In-person	United States	M
28	0.4611452...	Applied S...	Data Science a...	USD	1.1361946538792...	1.138073682...	United States	Mid-level	Full-time	In-person	United States	L

Normalization (z-score)

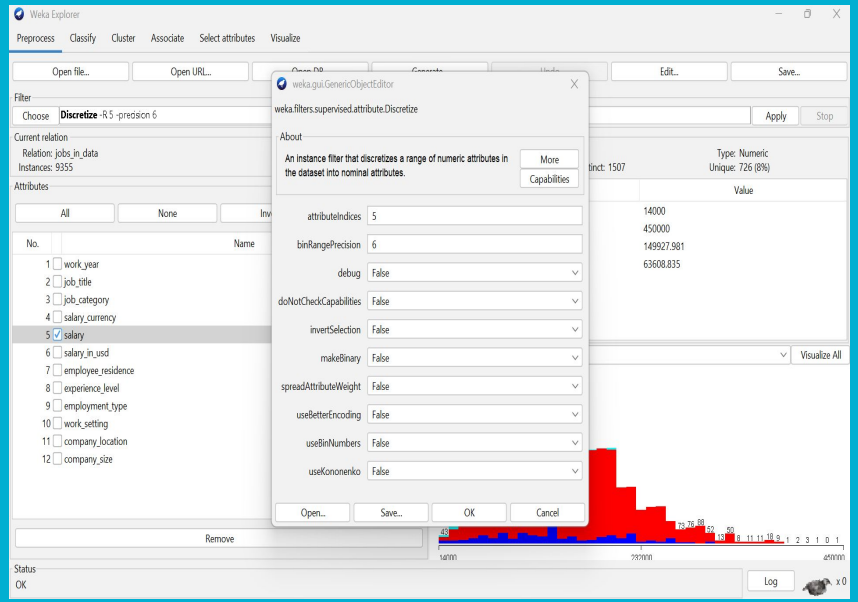
- Z-Score normalization (Standardize) is applied to ensure that numerical features have a mean of 0 and a standard deviation of 1. This helps in centering and scaling the data, making it suitable for algorithms that assume normally distributed features.

Discretization

Original

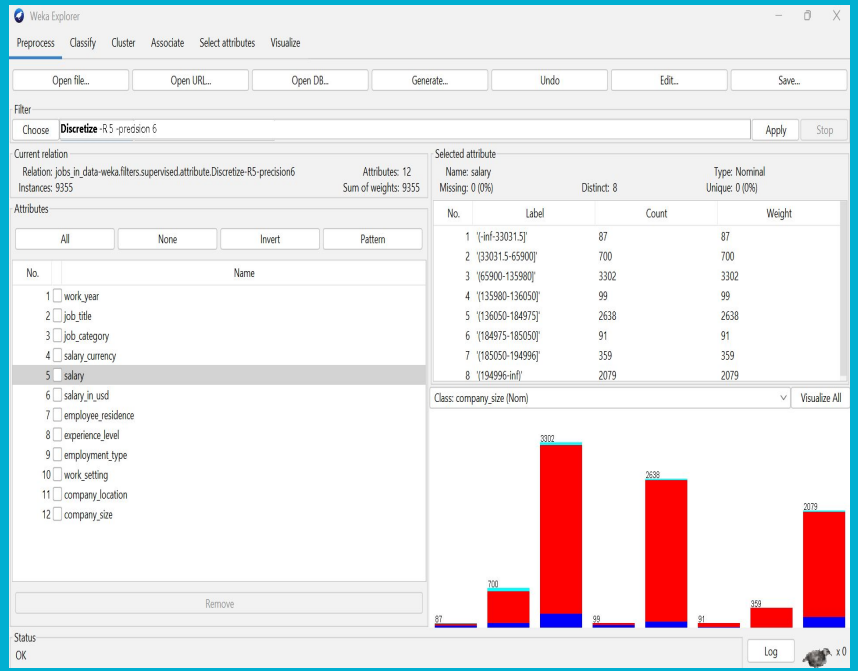


Applying ‘Discretize’



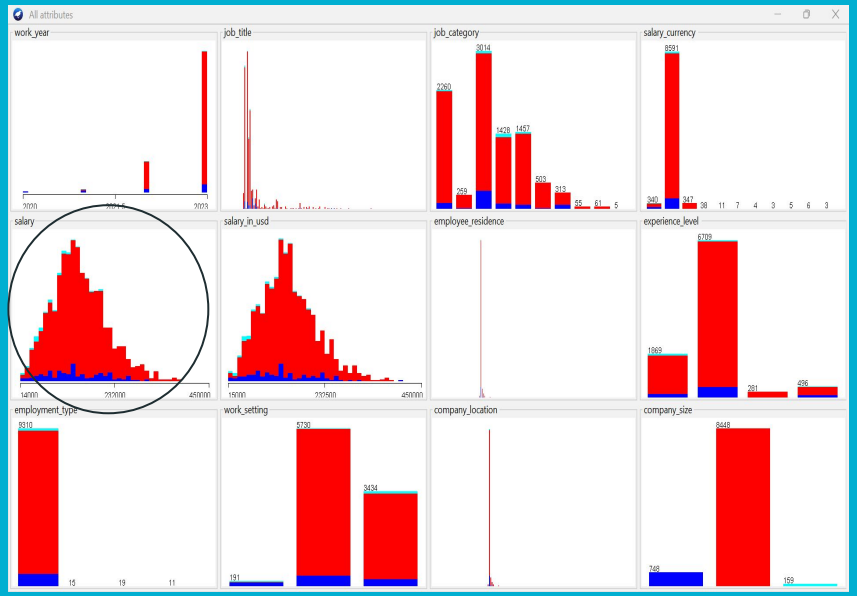
Discretization

- Implementation of supervised discretization techniques to transform continuous variables into discrete bins.
- Conversion of numerical attributes into nominal categories to simplify analysis and enhance interpretability.

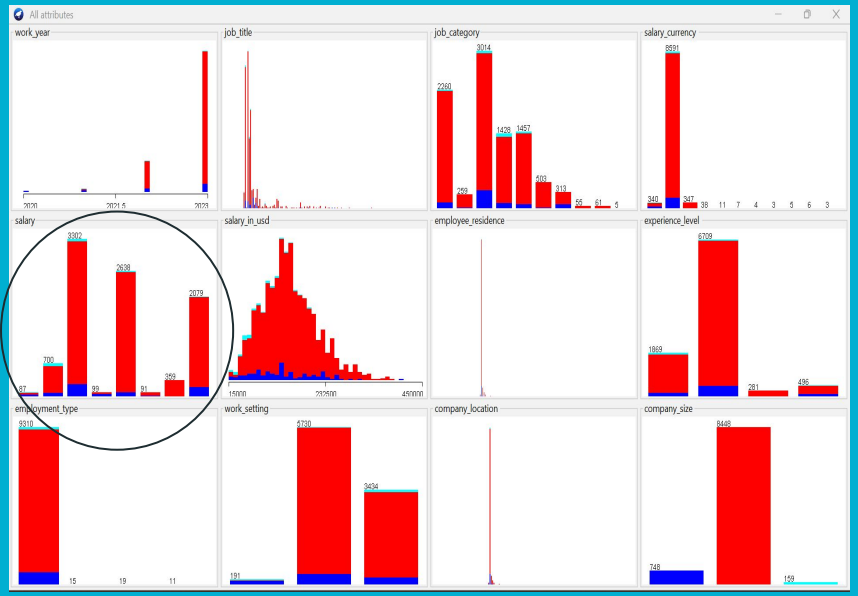


Discretization

Original



After ‘Discretize’

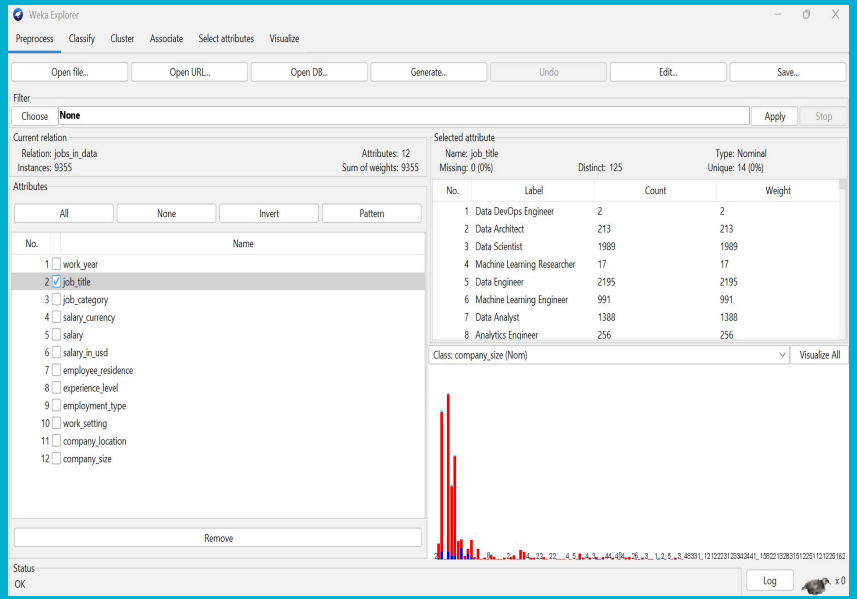


Discretization

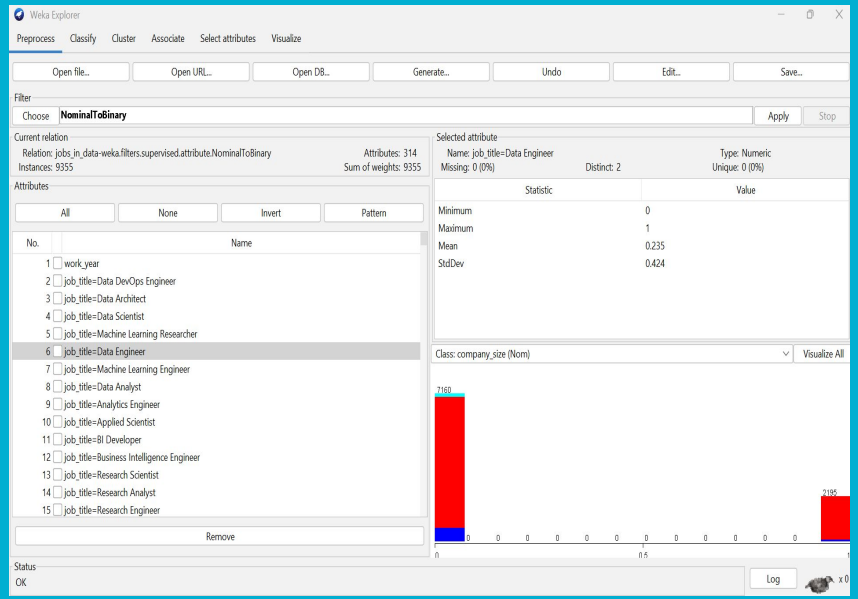
- Continuous variables, such as salary, may benefit from discretization to simplify the dataset and identify patterns.
- Supervised discretization methods allows us to use class information to guide the binning process, leading to more meaningful and interpretable results.
- Transformation of numerical attributes into nominal categories enhances the dataset's interpretability and facilitates analysis, particularly when exploring relationships between different attributes.

Change Nominal to Binary

Original



Applied Nominal to Binary

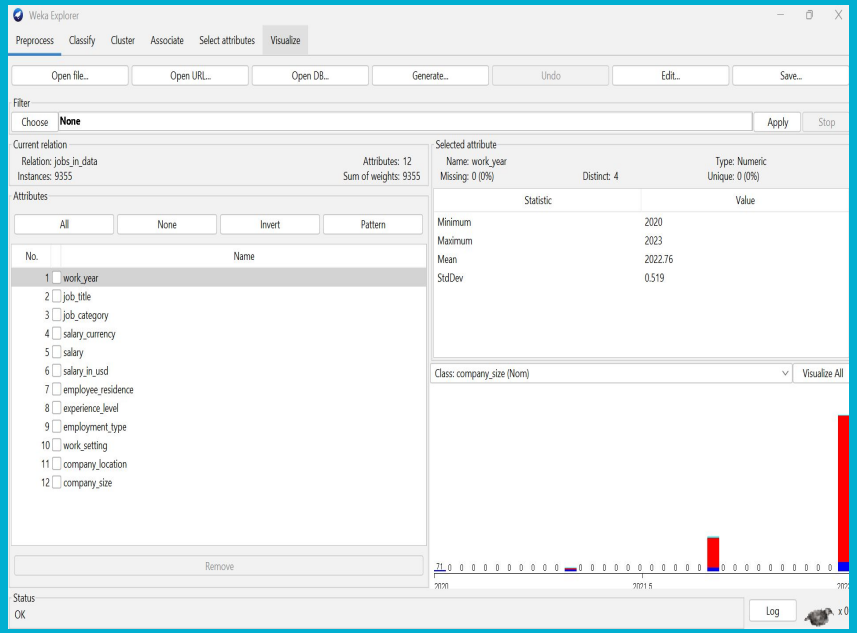


Change Nominal to Binary

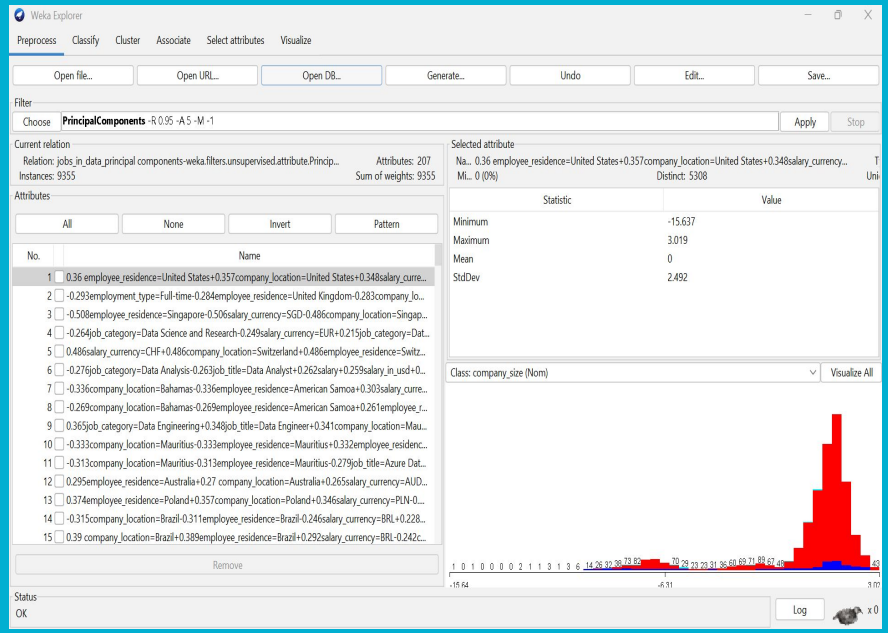
- Nominal attributes, such as job title, job category, often require transformation for better algorithm compatibility.
- The supervised approach ensures that the transformation is guided by class information, preserving the underlying relationships within the data.

Data Reduction-PCA

Original



Application of PCA



Data Reduction-PCA

- High-dimensional datasets can pose challenges for computational efficiency and interpretability.
- Principal Component Analysis (PCA) is applied to identify the principal components, effectively reducing the dataset's dimensionality. This facilities faster computation and aids in visualizing essential patterns and relationships within the data.

Summary of Operations

- The preprocessing operations collectively enhance dataset robustness by addressing missing values, standardize numerical features to prevent scale dominance, and simplify the dataset through supervised discretization. Transformation of nominal attributes into binary format improves algorithm compatibility, while PCA reduced dimensionality for improved efficiency and interpretability.
- Together, these preprocessing steps prepare the dataset for analysis, ensuring it is more robust, standardized, and suitable for a wide range of analytical techniques.

Thank You