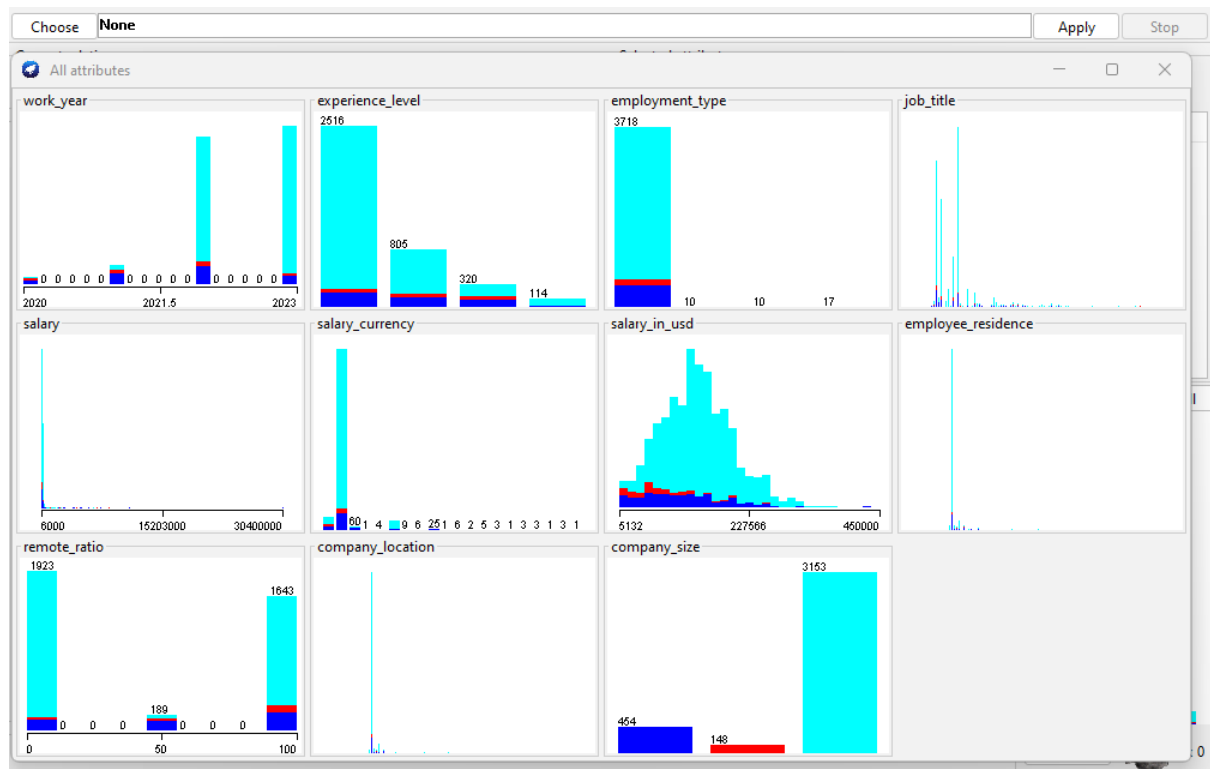


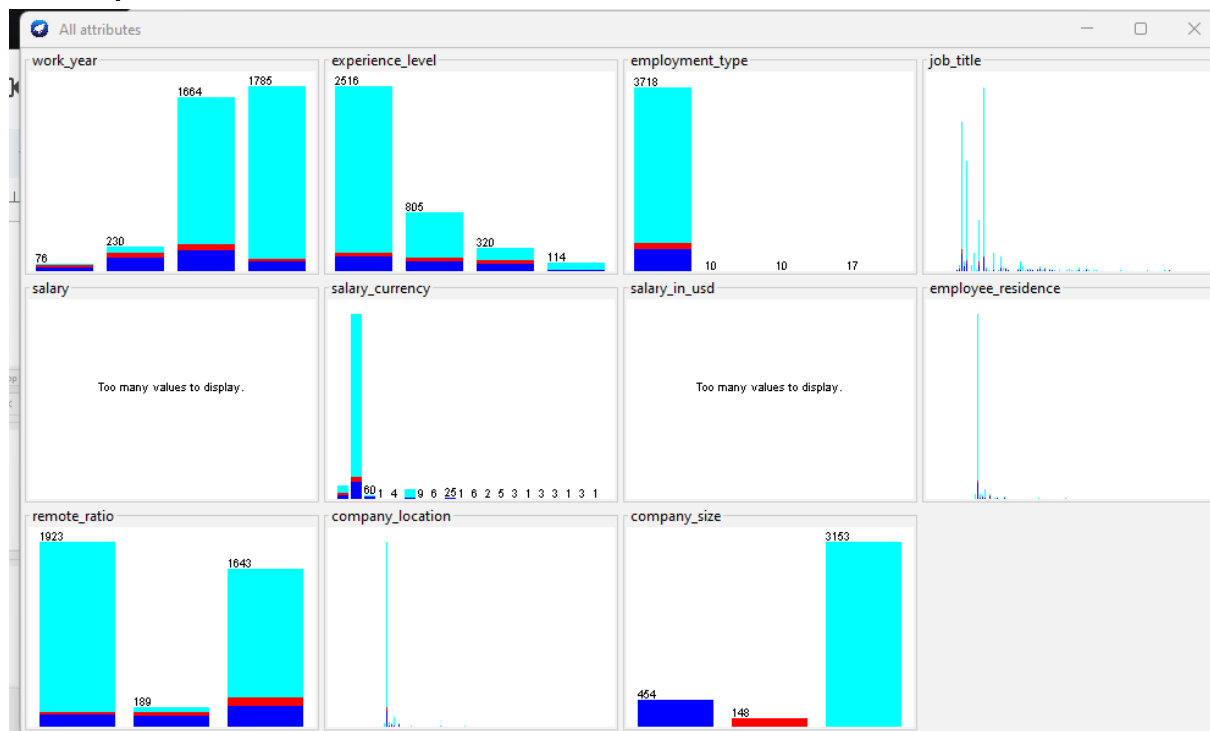
Ejercicio 2

Leidy Laura Patzi Tancara

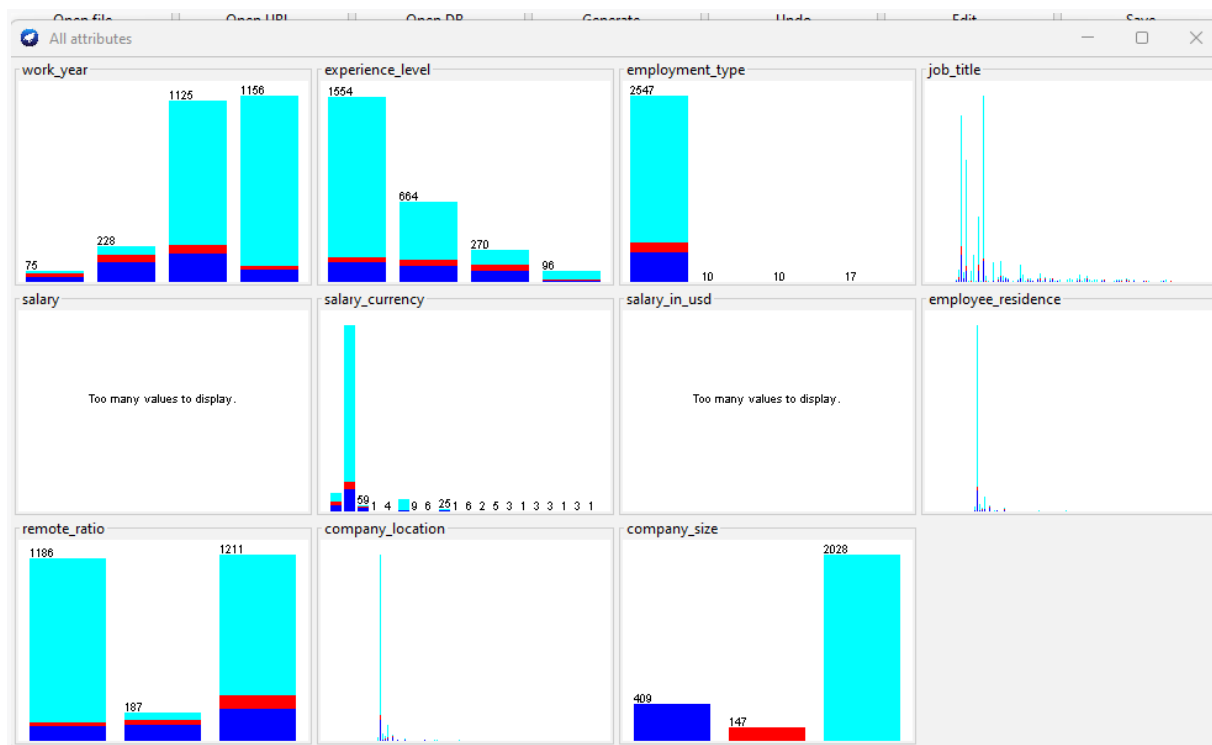
El dataset antes sin preprocesamiento alguno



Ahora aplicamos el filtro NumericToNominal

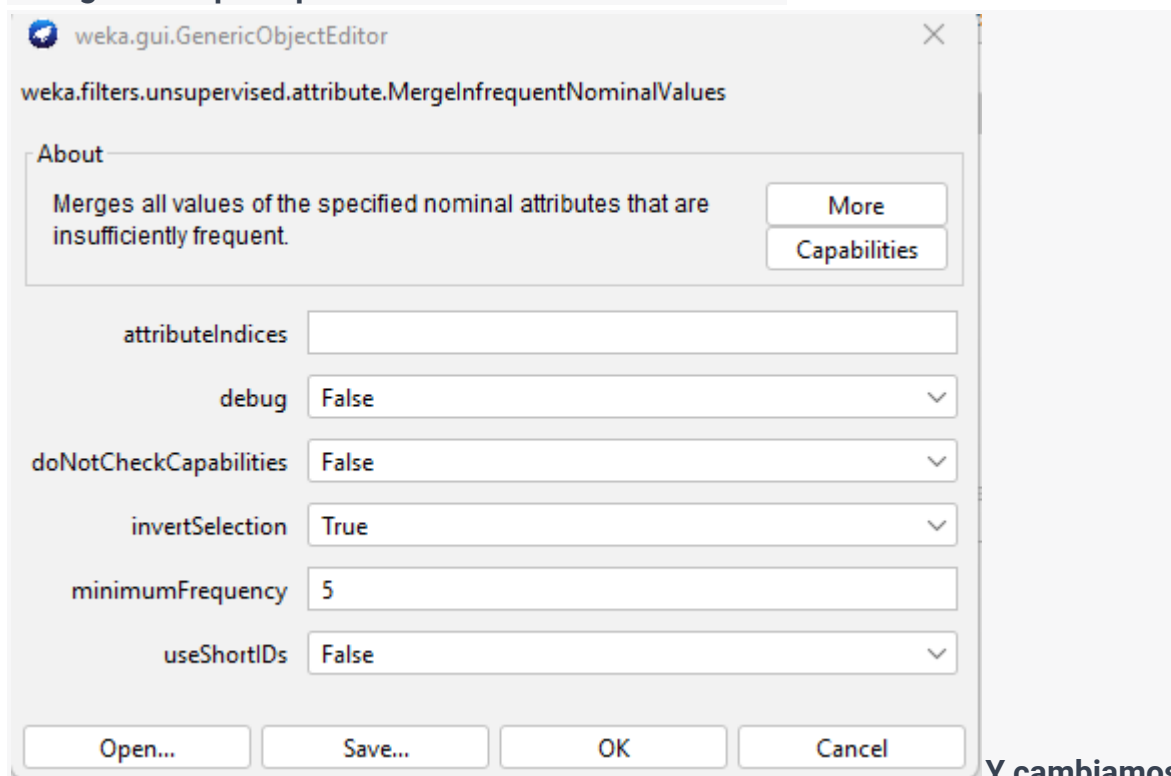


RemoveDuplicates Remover datos duplicados:



MergeInfrequentNominalValues:

Nos sirve para unir variables que tienen baja frecuencia en uno solo, en este caso lo configuramos para que el mínimo de frecuencia sea 5:



el invertSelecccion a true y estos son los resultados:

Sin el filtro:

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **MergeInfrequentNominalValues -N 5 -R** Apply Stop

Current relation: Relation: ds_salaries Instances: 3755 Attributes: 11 Sum of weights: 3755

Attributes: All None Invert Pattern

No. Name

- ☐ work_year
- ☐ experience_level
- ☐ employment_type
- ☒ job_title
- ☐ salary
- ☐ salary_currency
- ☐ salary_in_usd
- ☐ employee_residence
- ☐ remote_ratio
- ☐ company_location
- ☐ company_size

Remove

Selected attribute: Name: job_title Missing: 0 (0%) Distinct: 93 Type: Nominal Unique: 18 (0%)

No.	Label	Count	Weight
4	Applied Scientist	58	58
5	Data Analyst	612	612
6	Data Modeler	2	2
7	Research Engineer	37	37
8	Analytics Engineer	103	103
9	Business Intelligence E...	4	4
10	Machine Learning Engi...	289	289
11	Data Strategist	2	2
12	Data Engineer	1040	1040
13	Computer Vision Engin...	18	18

Class: company_size (Nom) Visualize All

Status: OK Log x 0

Aplicando el filtro:

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **MergeInfrequentNominalValues -N 5 -R -V** Apply Stop

Current relation: Relation: ds_salaries-weka.filters.unsupervised.attribute.Merg... Instances: 3755 Attributes: 11 Sum of weights: 3755

Attributes: All None Invert Pattern

No. Name

- ☐ work_year
- ☐ experience_level
- ☐ employment_type
- ☒ job_title_merged_infrequent_values
- ☐ salary
- ☐ salary_currency_merged_infrequent_values
- ☐ salary_in_usd
- ☐ employee_residence_merged_infrequent_values
- ☐ remote_ratio
- ☐ company_location_merged_infrequent_values
- ☐ company_size

Remove

Selected attribute: Name: job_title_merged_infrequent_values Missing: 0 (0%) Distinct: 50 Type: Nominal Unique: 0 (0%)

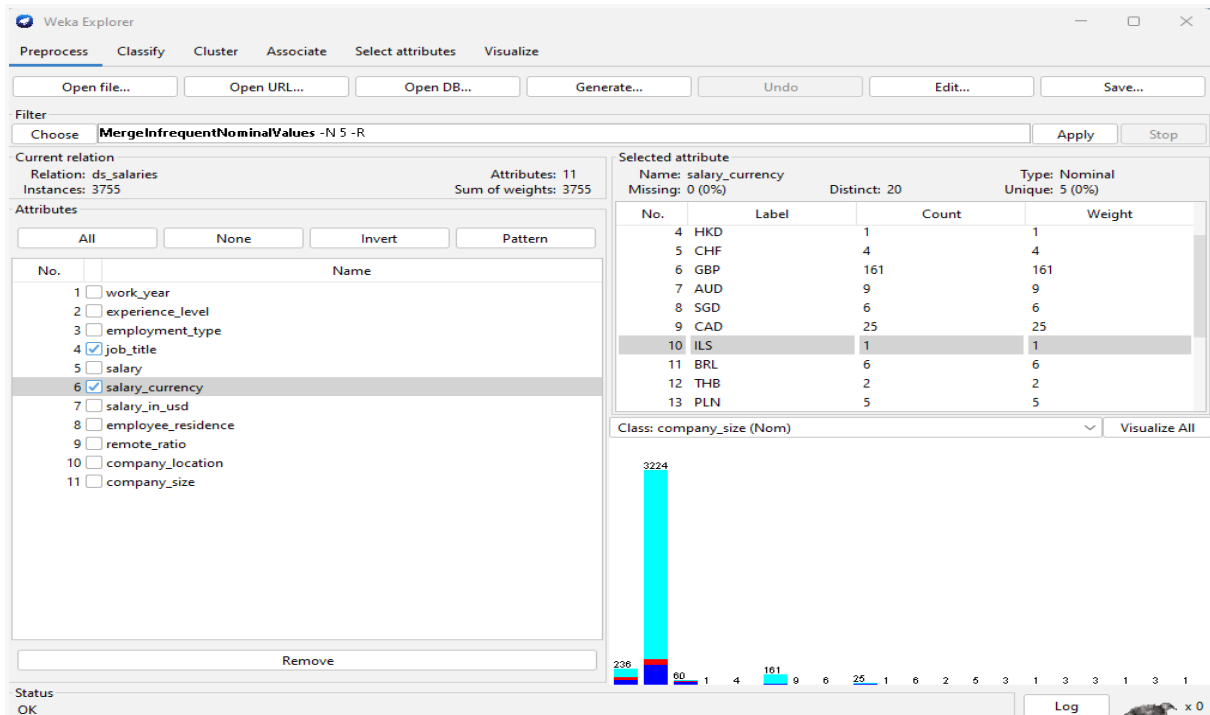
No.	Label	Count	Weight
1	Data Modeler_or_Busin...	84	84
2	Principal Data Scientist	8	8
3	ML Engineer	34	34
4	Data Scientist	840	840
5	Applied Scientist	58	58
6	Data Analyst	612	612
7	Research Engineer	37	37
8	Analytics Engineer	103	103
9	Machine Learning Engi...	289	289
10	Data Engineer	1040	1040

Class: company_size (Nom) Visualize All

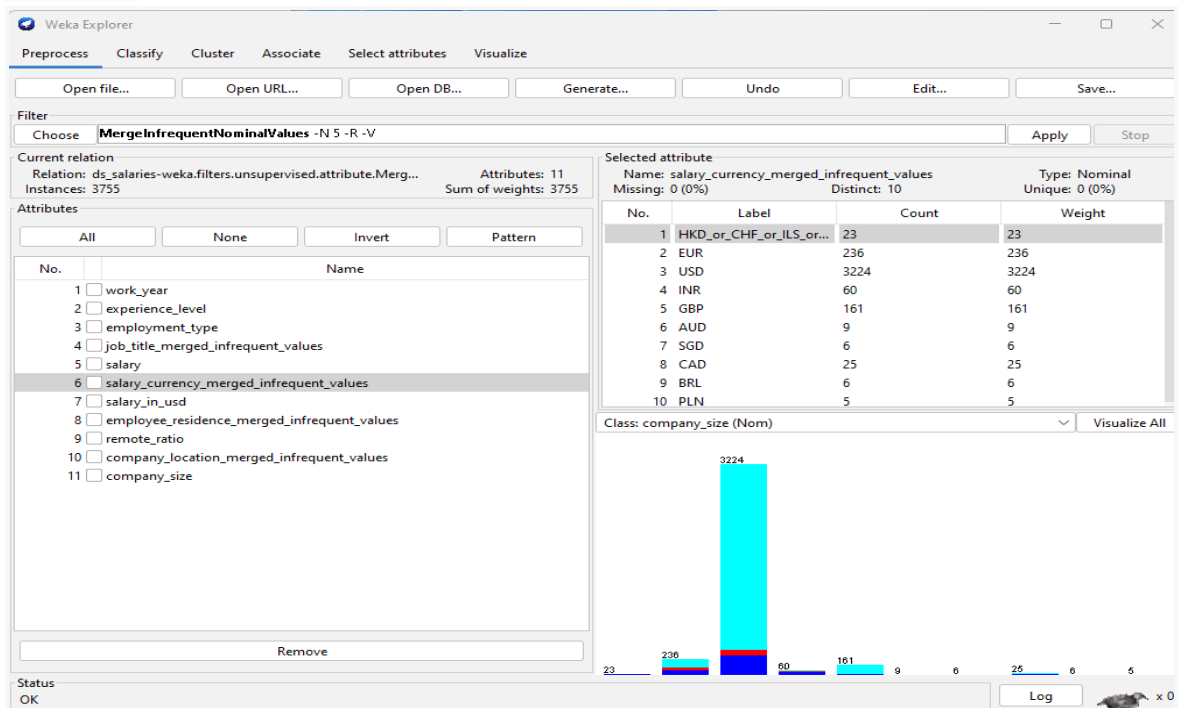
Status: OK Log x 0

Podemos observar que el nro 6 data Modeler que tenia una frecuencia de 2 ya no esta pero hay una nueva columna 1 donde muestra todas las variables con poca frecuencia y juntos se hicieron un total de 84 datos frecuentes.

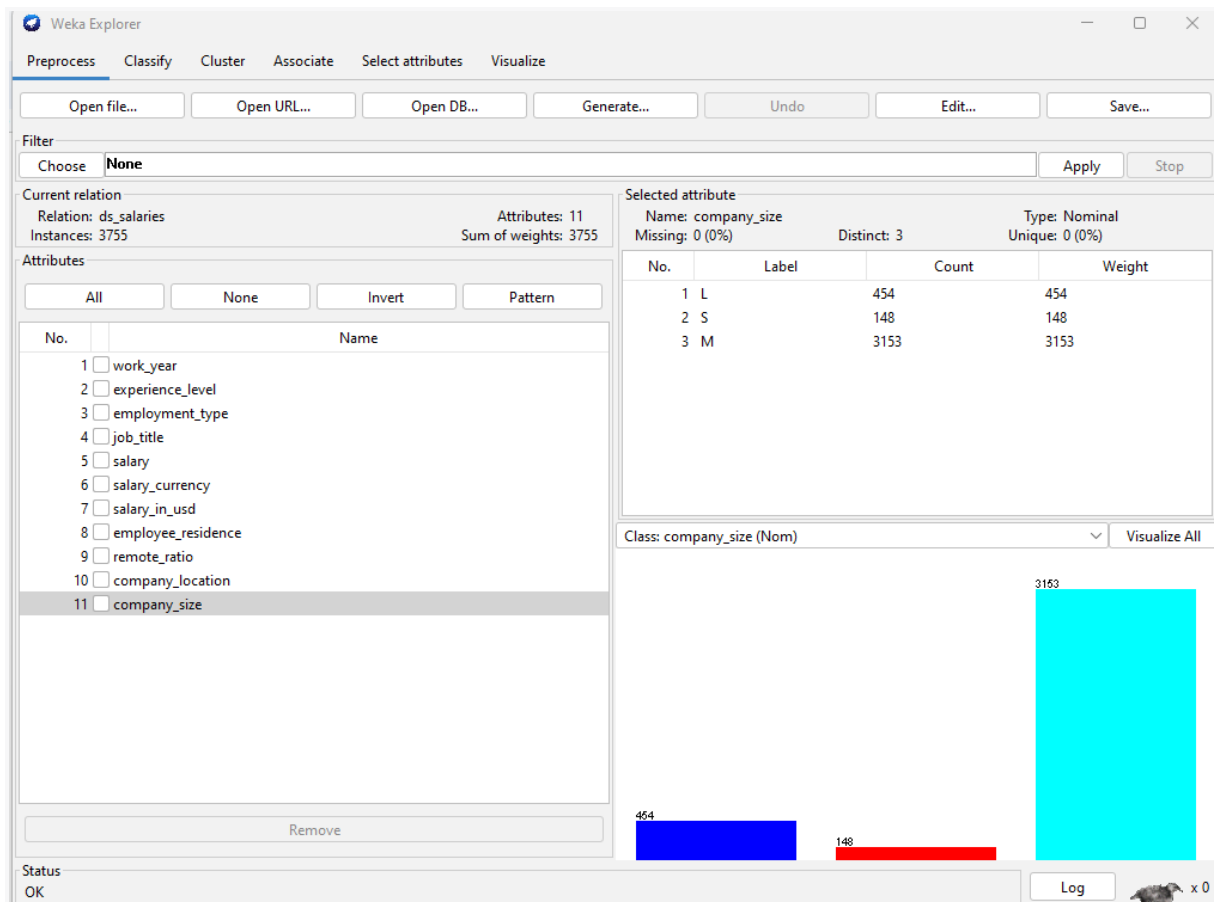
Aplicando el filtro a otra columna:
sin filtro:



Con filtro



MakeIndicator: Convierte valores categoricos a binarios:
sin filtro:



Con filtro:

