DATA MINING PROJECT

Palestine Technical University – Kadoorie

College of Engineering and Technology

Department of Computer Systems Engineering

Project title:

CUSTOMER SEGMENTATION CLASSIFICATION

By:

Ruaa Qashoo – 201911196

Instructor:

Dr. Anas Melhem

1 January 2023

INTRODUCTION

❖PROBLEM DEFINITAION:

An automobile company has plans to enter new markets with their existing products (P1, P2, P3, P4, and P5). After intensive market research, they've deduced that the behavior of the new market is similar to their existing market.

In their existing market, the sales team has classified all customers into 4 segments (A, B, C, D). Then, they performed segmented outreach and communication for a different segment of customers. This strategy has worked exceptionally well for them. They plan to use the same strategy for the new markets and have identified 2627 new potential customers.

The goal of this project is to study and predict the right group of new customers for an automotive company, so the company can adopt the specific proven marketing strategy to each of them and be more successful in the business.

I decided to use Weka to analyze the problem and then solve it.

dataset description:

There is 8068 instances (customers) for training, and 2627 instance (new potential customers) for testing.

This dataset has 11 Variables (10 attributes & 1 class/target), divided as follows:

Link of The Dataset: Customer Segmentation Classification | Kaggle

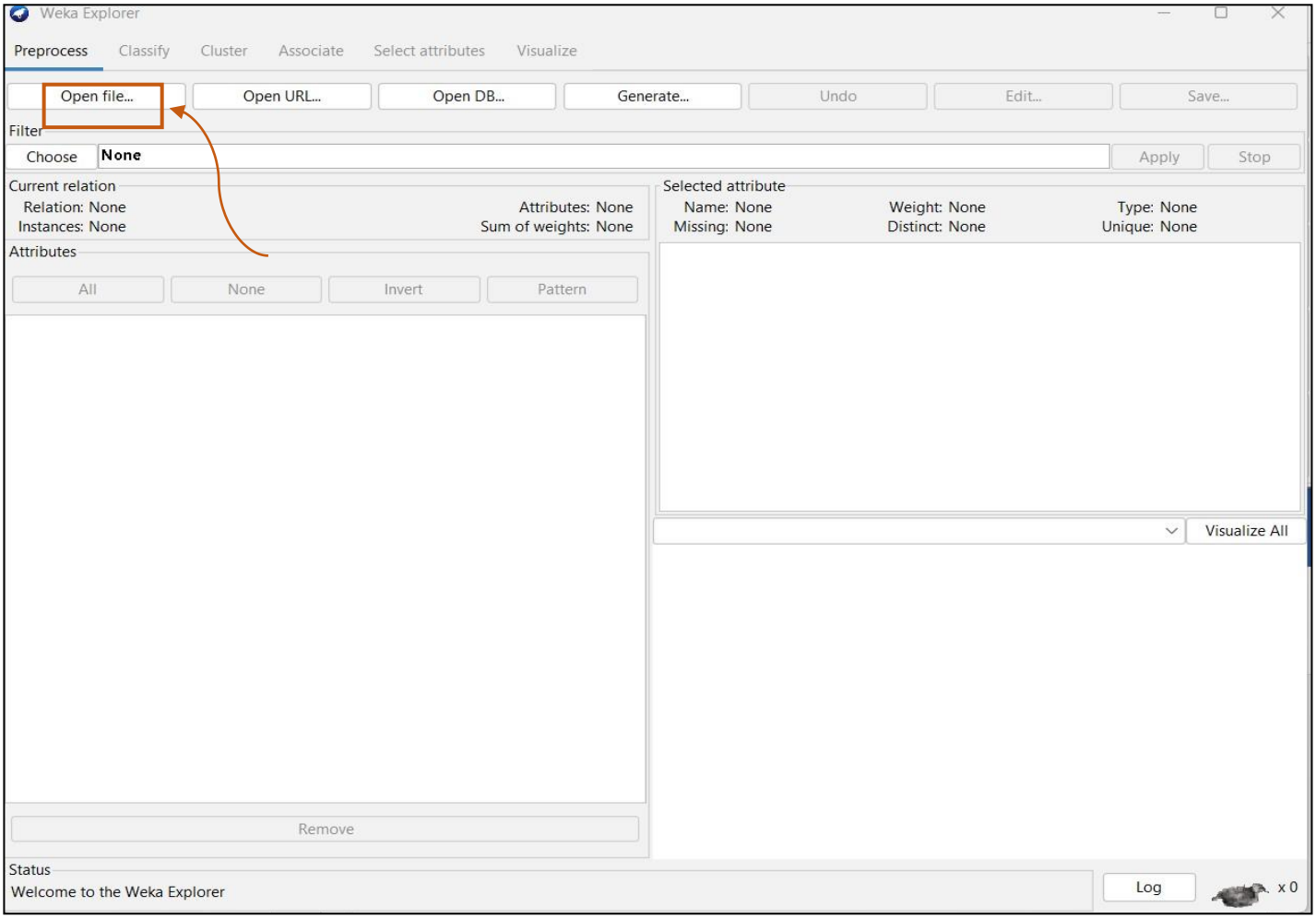
VARIABLES	TYPE	DEFINITION
ID	numeric	Unique ID
GENDER	binary	Gender of the customer
EVER_MARRID	binary	Marital status of the customer
AGE	numeric	Age of the customer
GRADUATED	binary	Is the customer a graduate?
PROFESSION	nominal	Profession of the customer
WORK_EXPERIANCE	numeric	Work Experience in years
SPENDING	Ordinal	Spending score of the customer
FAMILY SIZE	numeric	Number of family members for the customer (including the customer)
VAR_1	nominal	Anonymized Category for the customer
SEGMANT	nominal	(target) Customer Segment of the customer

Problems with this data & how I solve it

<< Clean/Prepare the data >>

Missing Values:

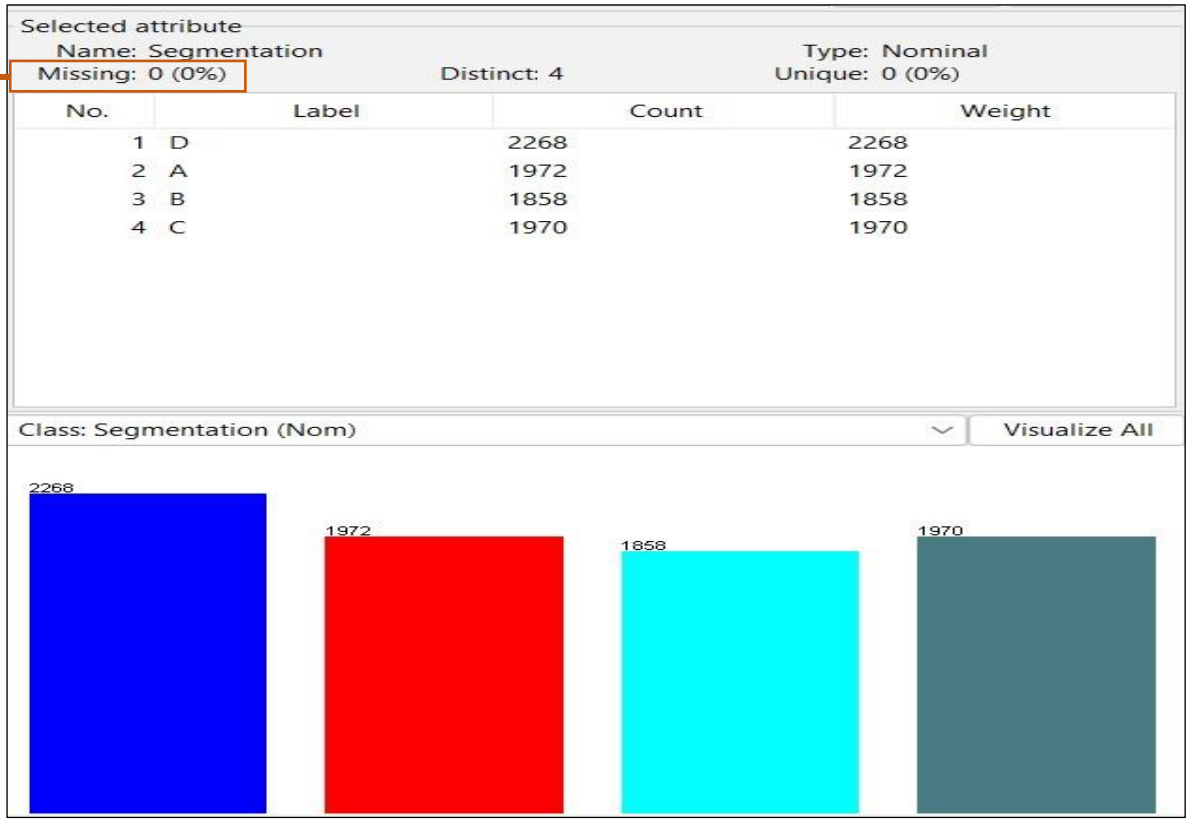
after uploading the data set:



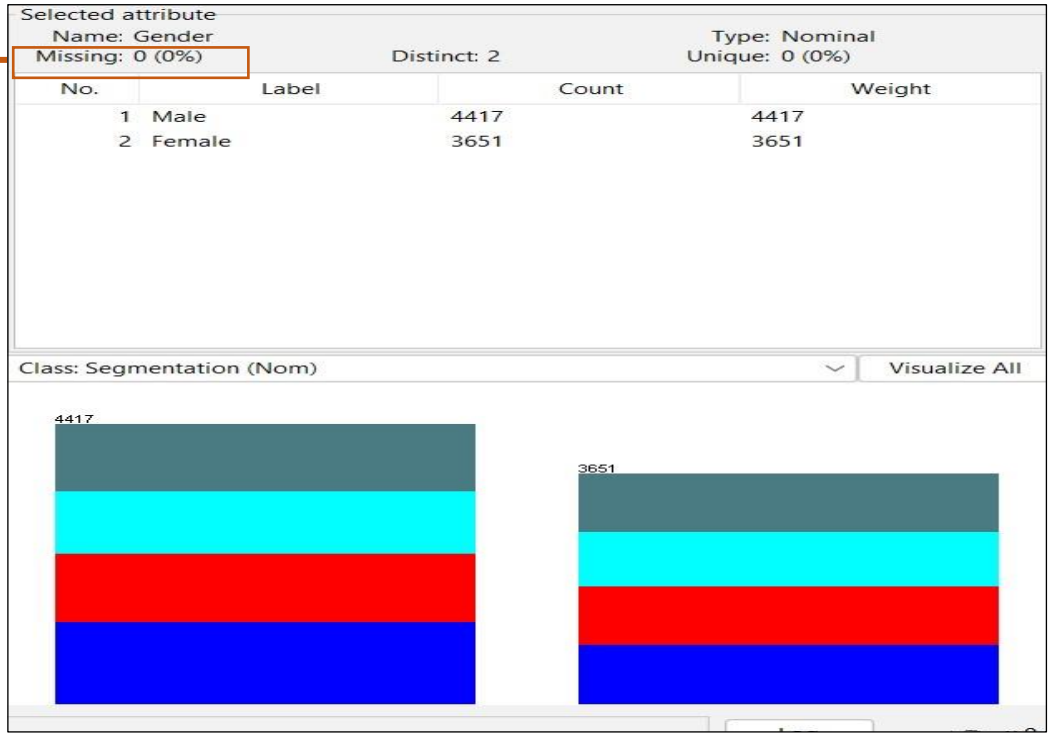
click open file → chose our training dataset → this information about each attribute will appear:

NOTE: as we can see there is 4 colors in each chart

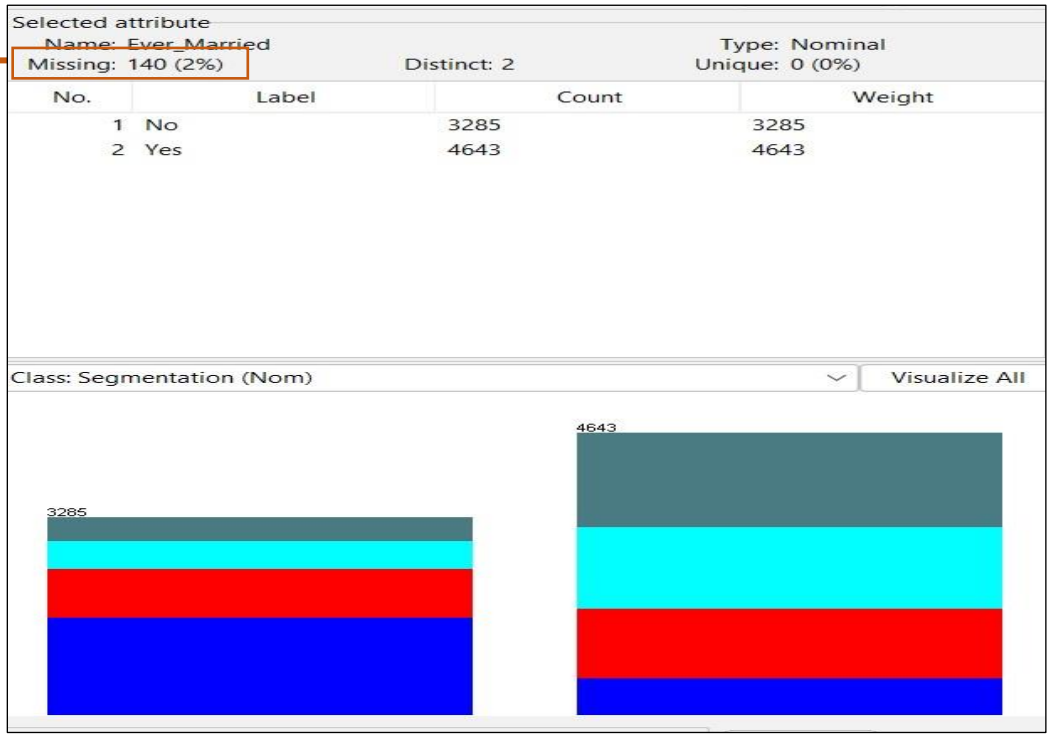
- represent data from class A
- represent data from class B
- represent data from class C
- represent data from class D



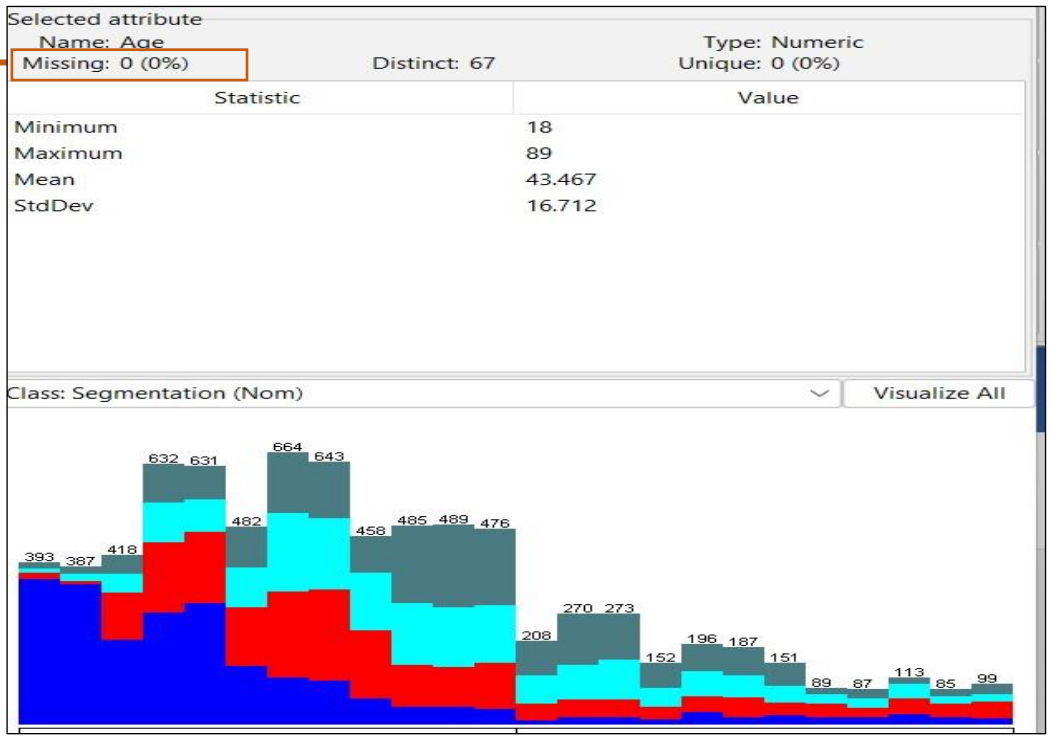
- There is no missing value in Segmentation attribute (class)



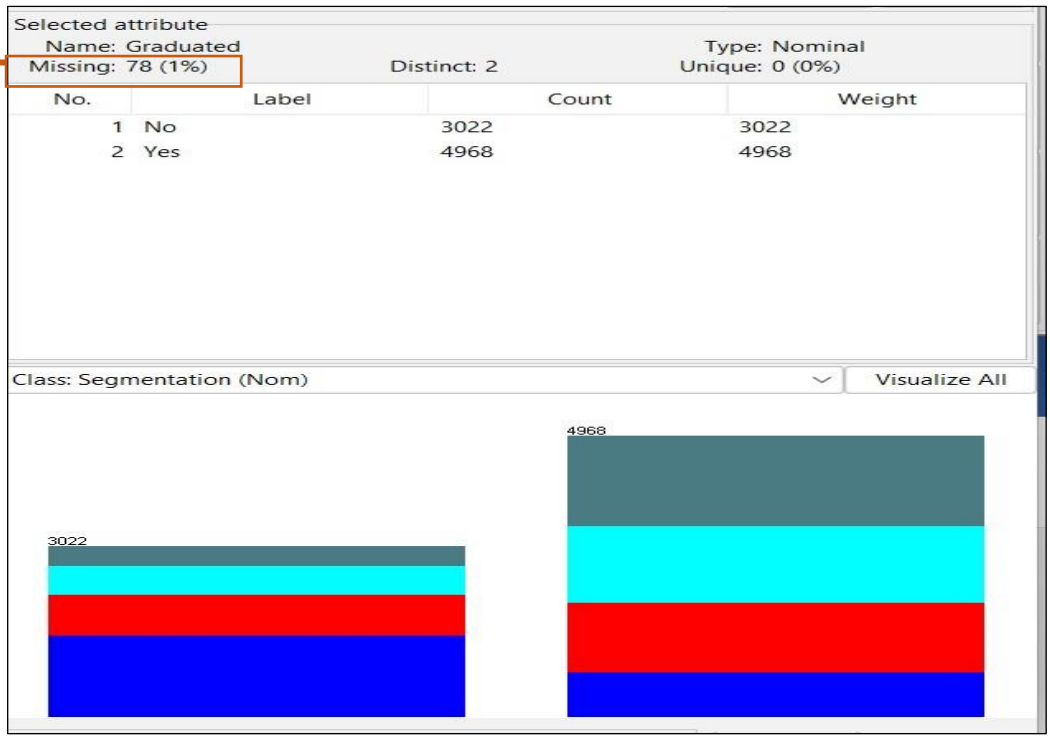
- There is no missing value in Gender attribute



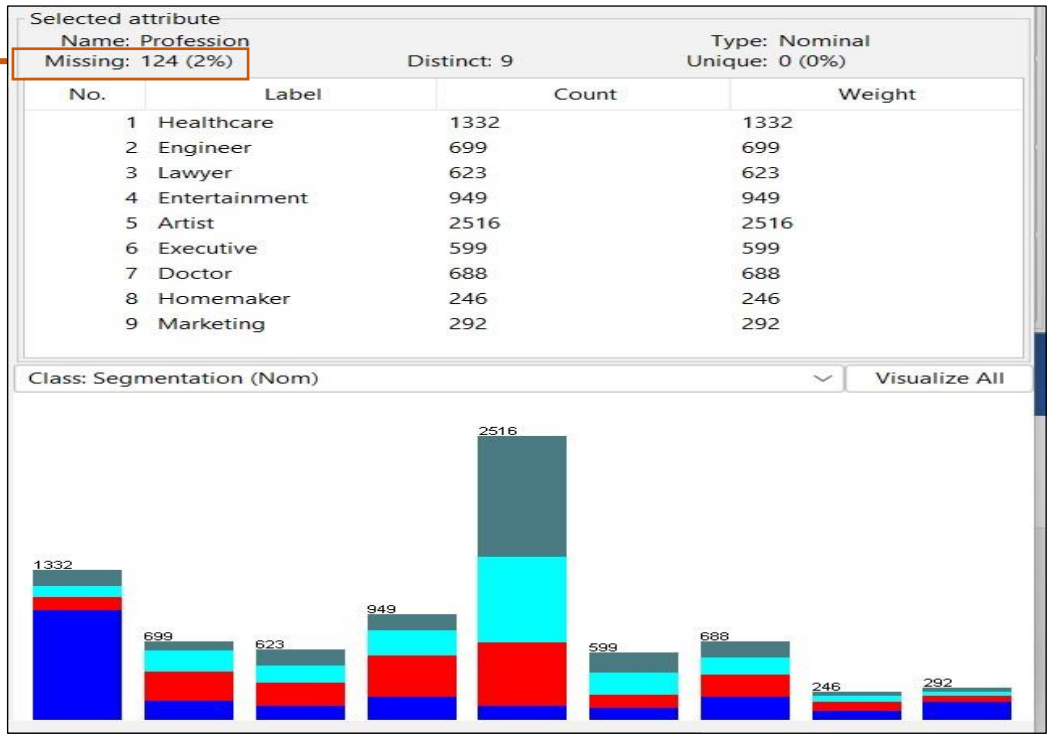
- There is 144 missing value in Ever_Married attribute



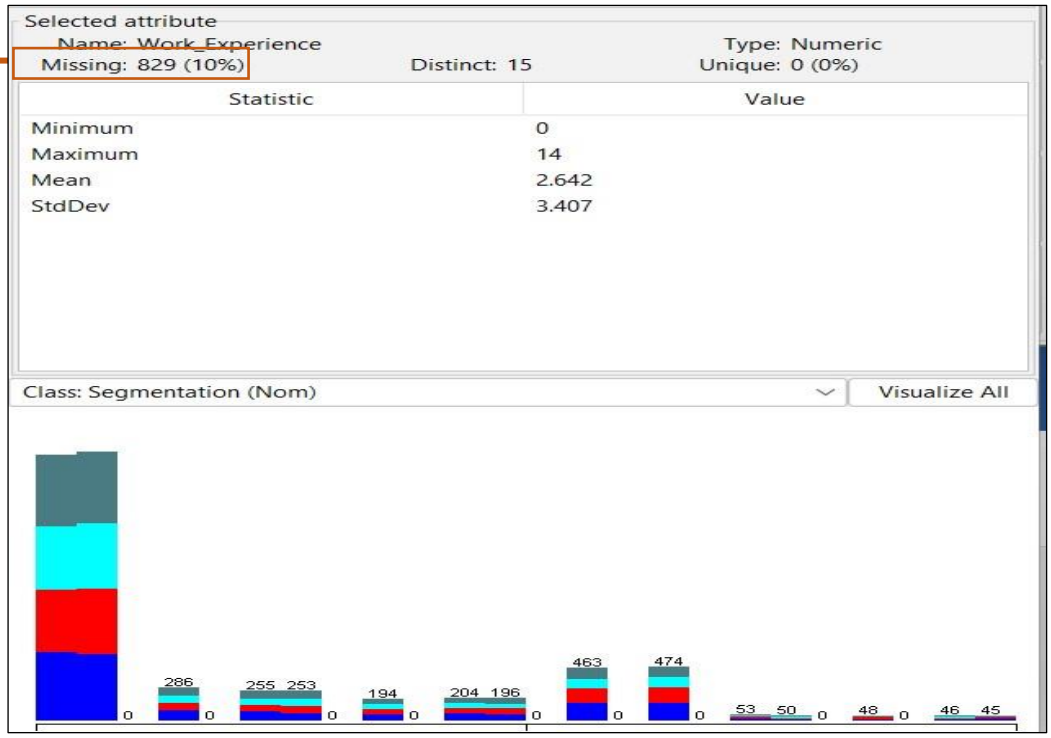
- There is no missing value in Age attribute



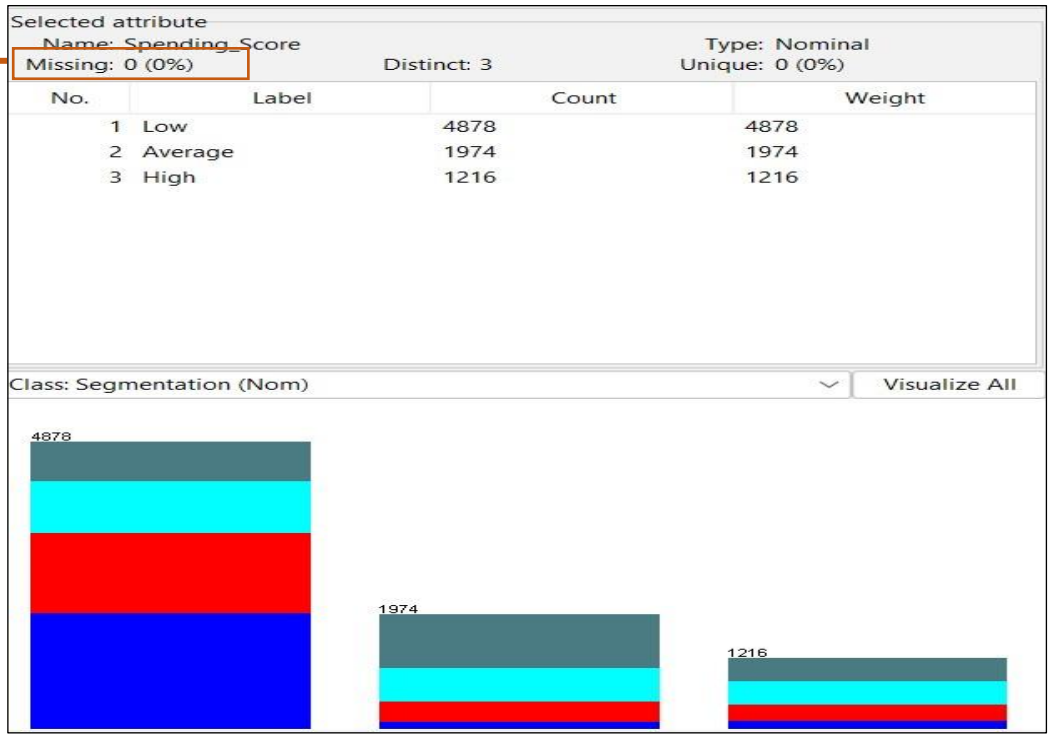
- There is 78 missing value in Graduated attribute



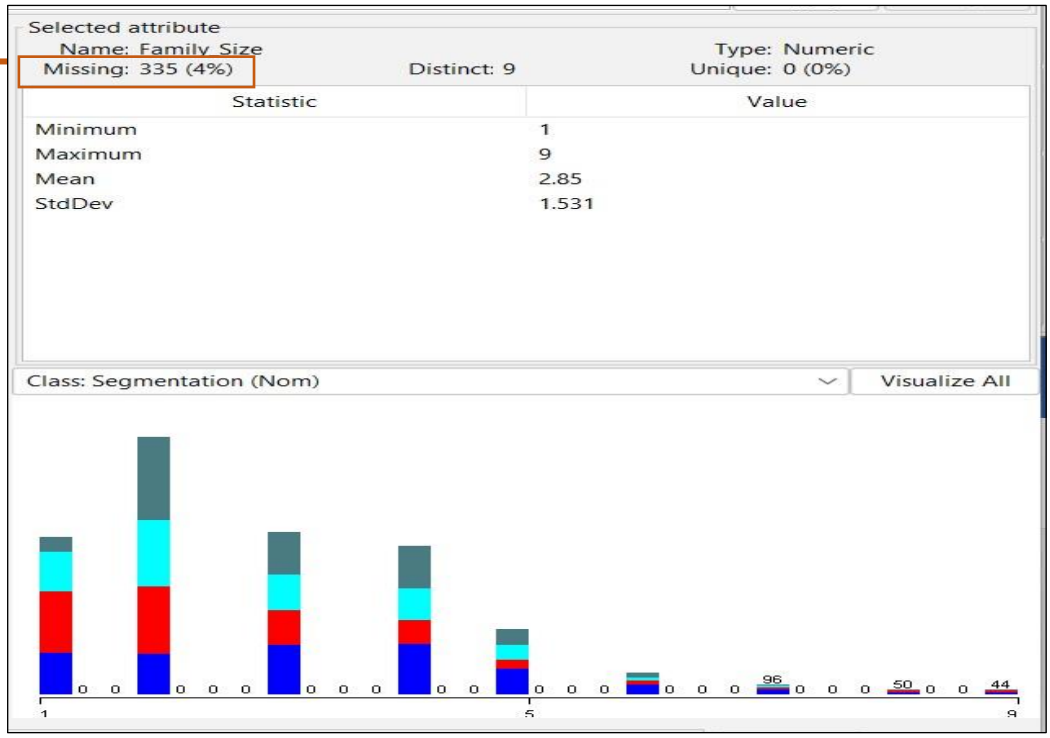
- There is 124 missing value in Profession attribute



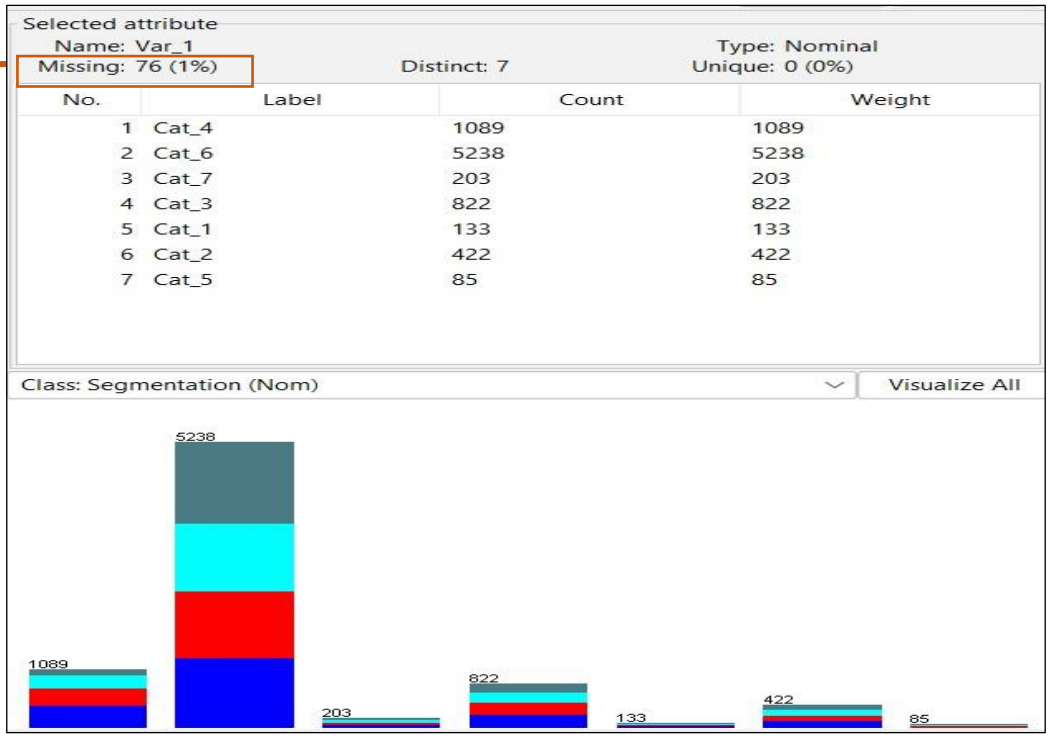
- There is 829 missing value in Work_Experience attribute



- There is no missing value in Spending attribute



- There is 335 missing value in Family_Size attribute

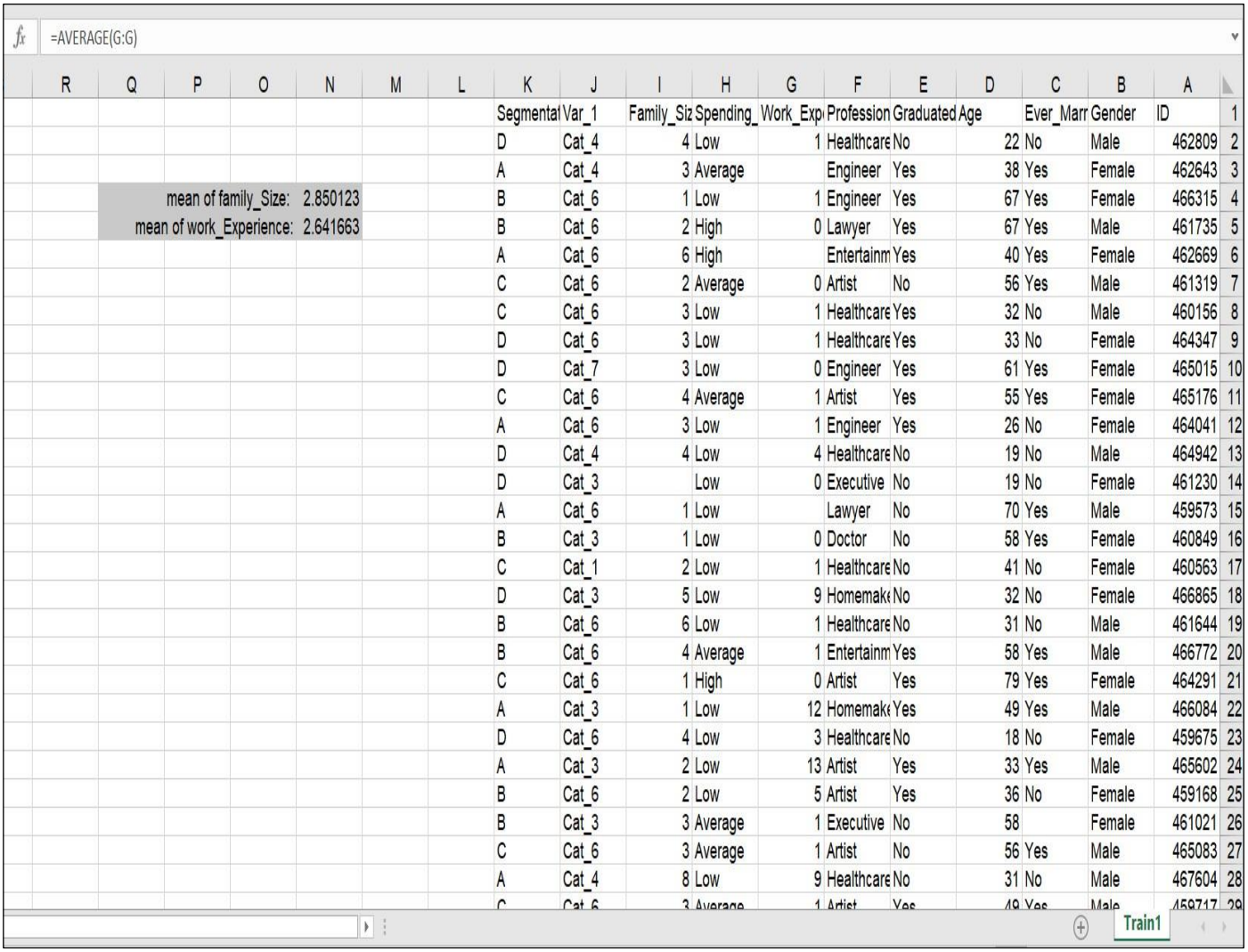


- There is 76 missing value in Var_1 attribute

#Missing Value processing

As above, we see that data contains some missing values. there are total of 8068 entries, but some columns have less than 8068 entries, which means they have missing values:

Columns " Family Size "and " Work Experience " have huge number of null values. Both are numeric attribute, so can be replaced with their mean values. **But** because each of these features represents an integer value (1,2, 3……) and the mean of each respectively is (2.85, 2.64) which is not an integer.



=AVERAGE(G:G)

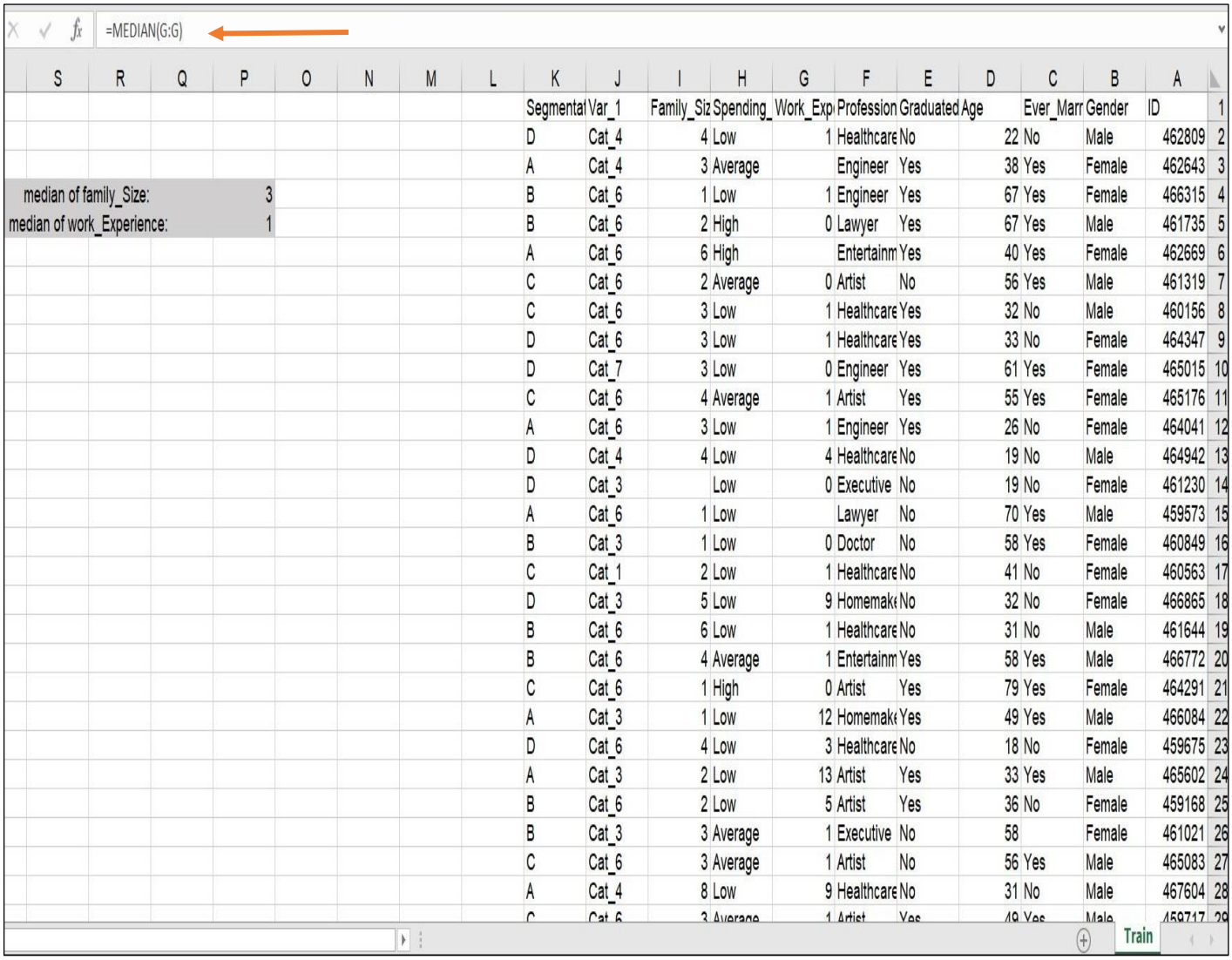
R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	
							Segmental	Var_1	Family_Siz	Spending_	Work_Exp	Profession	Graduated	Age	Ever_Marr	Gender	ID	1
							D	Cat_4	4	Low	1	Healthcare	No	22	No	Male	462809	2
							A	Cat_4	3	Average		Engineer	Yes	38	Yes	Female	462643	3
		mean of family_Size:		2.850123			B	Cat_6	1	Low	1	Engineer	Yes	67	Yes	Female	466315	4
		mean of work_Experience:		2.641663			B	Cat_6	2	High	0	Lawyer	Yes	67	Yes	Male	461735	5
							A	Cat_6	6	High		Entertainm	Yes	40	Yes	Female	462669	6
							C	Cat_6	2	Average	0	Artist	No	56	Yes	Male	461319	7
							C	Cat_6	3	Low	1	Healthcare	Yes	32	No	Male	460156	8
							D	Cat_6	3	Low	1	Healthcare	Yes	33	No	Female	464347	9
							D	Cat_7	3	Low	0	Engineer	Yes	61	Yes	Female	465015	10
							C	Cat_6	4	Average	1	Artist	Yes	55	Yes	Female	465176	11
							A	Cat_6	3	Low	1	Engineer	Yes	26	No	Female	464041	12
							D	Cat_4	4	Low	4	Healthcare	No	19	No	Male	464942	13
							D	Cat_3		Low	0	Executive	No	19	No	Female	461230	14
							A	Cat_6	1	Low		Lawyer	No	70	Yes	Male	459573	15
							B	Cat_3	1	Low	0	Doctor	No	58	Yes	Female	460849	16
							C	Cat_1	2	Low	1	Healthcare	No	41	No	Female	460563	17
							D	Cat_3	5	Low	9	Homemake	No	32	No	Female	466865	18
							B	Cat_6	6	Low	1	Healthcare	No	31	No	Male	461644	19
							B	Cat_6	4	Average	1	Entertainm	Yes	58	Yes	Male	466772	20
							C	Cat_6	1	High	0	Artist	Yes	79	Yes	Female	464291	21
							A	Cat_3	1	Low	12	Homemake	Yes	49	Yes	Male	466084	22
							D	Cat_6	4	Low	3	Healthcare	No	18	No	Female	459675	23
							A	Cat_3	2	Low	13	Artist	Yes	33	Yes	Male	465602	24
							B	Cat_6	2	Low	5	Artist	Yes	36	No	Female	459168	25
							B	Cat_3	3	Average	1	Executive	No	58		Female	461021	26
							C	Cat_6	3	Average	1	Artist	No	56	Yes	Male	465083	27
							A	Cat_4	8	Low	9	Healthcare	No	31	No	Male	467604	28
							C	Cat_6	3	Average	1	Artist	Yes	49	Yes	Male	459717	29

Train1

Also replacing missing with the mean value may affect outliers.

So, I decided to replace the missing values in these columns with the median value of each one:

- Median of family_size = 3
- Median of work_Experience = 1



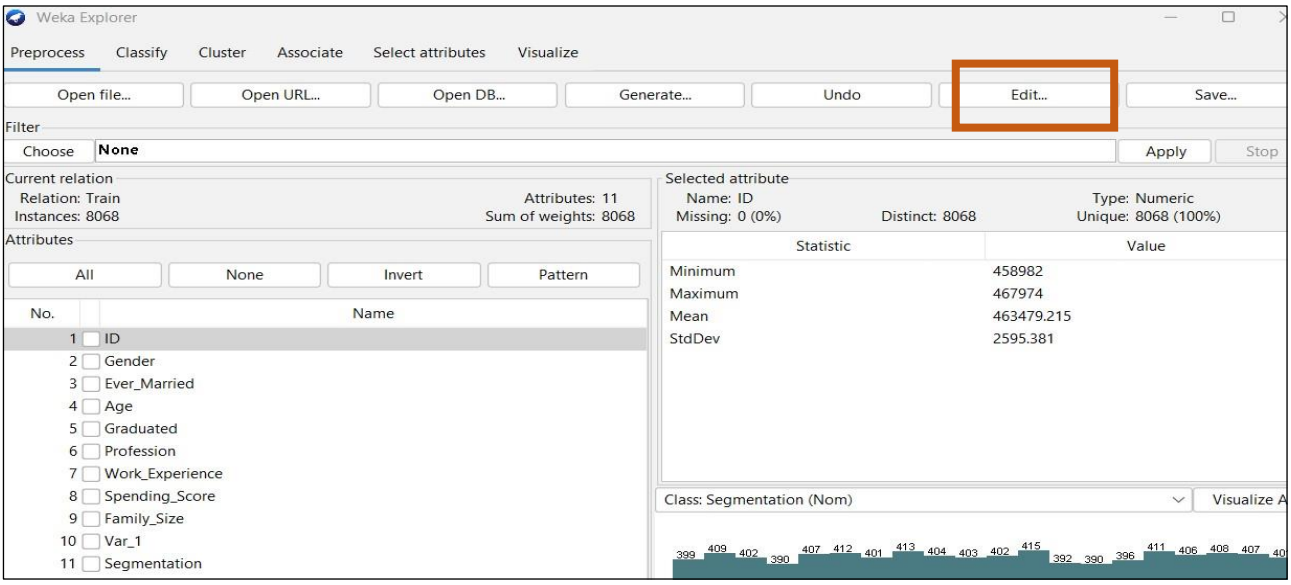
=MEDIAN(G:G)

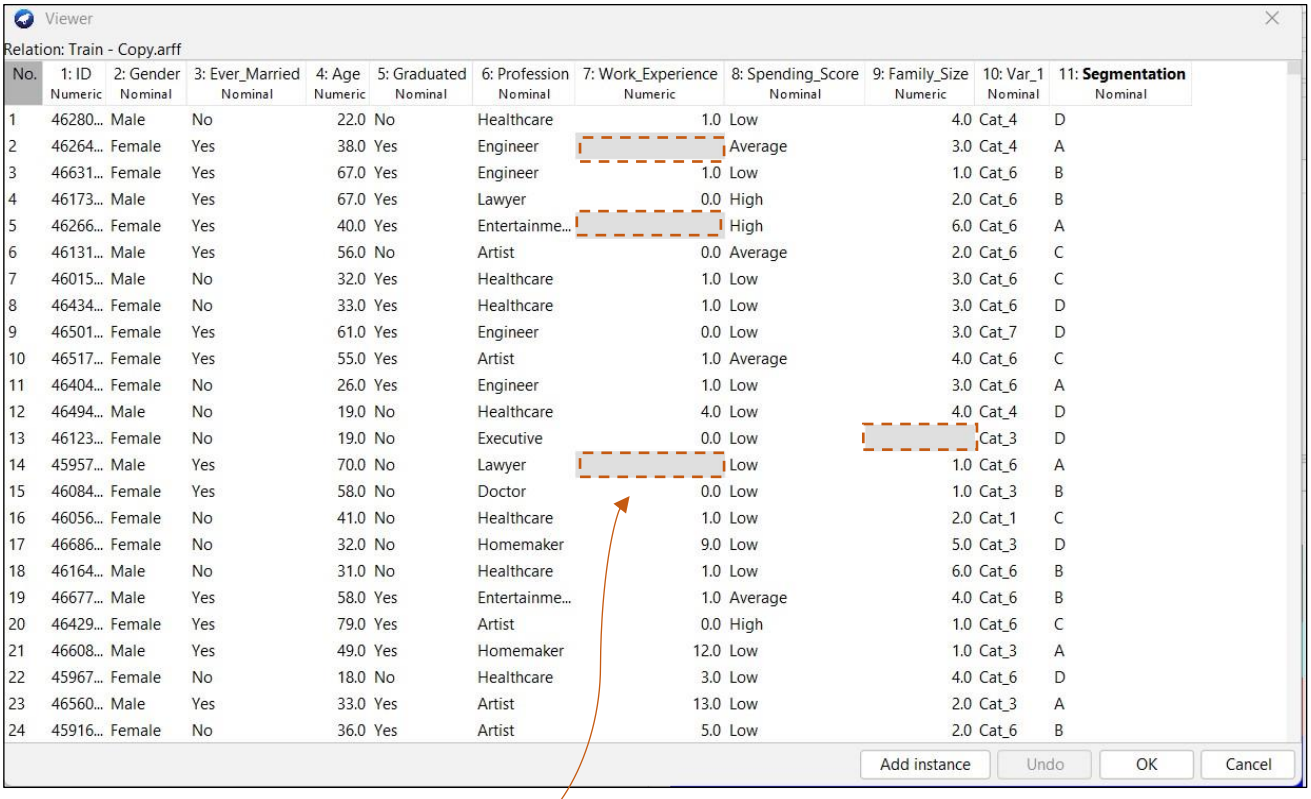
	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	
									Segmental	Var_1	Family_Siz	Spending_	Work_Exp	Profession	Graduated	Age	Ever_Marr	Gender	ID	1
									D	Cat_4	4	Low	1	Healthcare	No	22	No	Male	462809	2
									A	Cat_4	3	Average		Engineer	Yes	38	Yes	Female	462643	3
			median of family_Size:	3					B	Cat_6	1	Low	1	Engineer	Yes	67	Yes	Female	466315	4
			median of work_Experience:	1					B	Cat_6	2	High	0	Lawyer	Yes	67	Yes	Male	461735	5
									A	Cat_6	6	High		Entertainm	Yes	40	Yes	Female	462669	6
									C	Cat_6	2	Average	0	Artist	No	56	Yes	Male	461319	7
									C	Cat_6	3	Low	1	Healthcare	Yes	32	No	Male	460156	8
									D	Cat_6	3	Low	1	Healthcare	Yes	33	No	Female	464347	9
									D	Cat_7	3	Low	0	Engineer	Yes	61	Yes	Female	465015	10
									C	Cat_6	4	Average	1	Artist	Yes	55	Yes	Female	465176	11
									A	Cat_6	3	Low	1	Engineer	Yes	26	No	Female	464041	12
									D	Cat_4	4	Low	4	Healthcare	No	19	No	Male	464942	13
									D	Cat_3		Low	0	Executive	No	19	No	Female	461230	14
									A	Cat_6	1	Low		Lawyer	No	70	Yes	Male	459573	15
									B	Cat_3	1	Low	0	Doctor	No	58	Yes	Female	460849	16
									C	Cat_1	2	Low	1	Healthcare	No	41	No	Female	460563	17
									D	Cat_3	5	Low	9	Homemak	No	32	No	Female	466865	18
									B	Cat_6	6	Low	1	Healthcare	No	31	No	Male	461644	19
									B	Cat_6	4	Average	1	Entertainm	Yes	58	Yes	Male	466772	20
									C	Cat_6	1	High	0	Artist	Yes	79	Yes	Female	464291	21
									A	Cat_3	1	Low	12	Homemak	Yes	49	Yes	Male	466084	22
									D	Cat_6	4	Low	3	Healthcare	No	18	No	Female	459675	23
									A	Cat_3	2	Low	13	Artist	Yes	33	Yes	Male	465602	24
									B	Cat_6	2	Low	5	Artist	Yes	36	No	Female	459168	25
									B	Cat_3	3	Average	1	Executive	No	58		Female	461021	26
									C	Cat_6	3	Average	1	Artist	No	56	Yes	Male	465083	27
									A	Cat_4	8	Low	9	Healthcare	No	31	No	Male	467604	28
									C	Cat_6	3	Average	1	Artist	Yes	49	Yes	Male	459717	29

Train

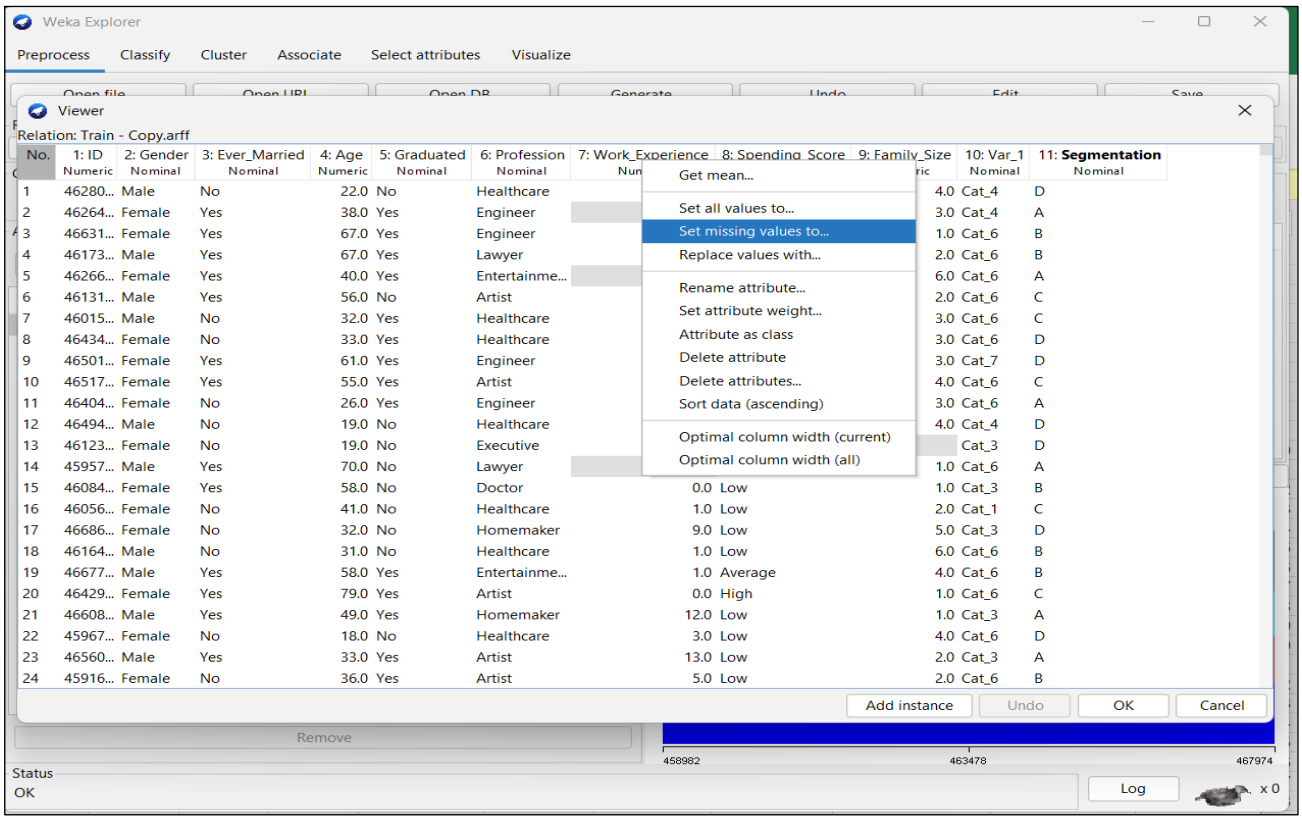
Replacing them in weka:

open the training dataset → click Edit → write click on attribute which needed to edit → then choose set messing value to and enter the median value (that we calculated above):

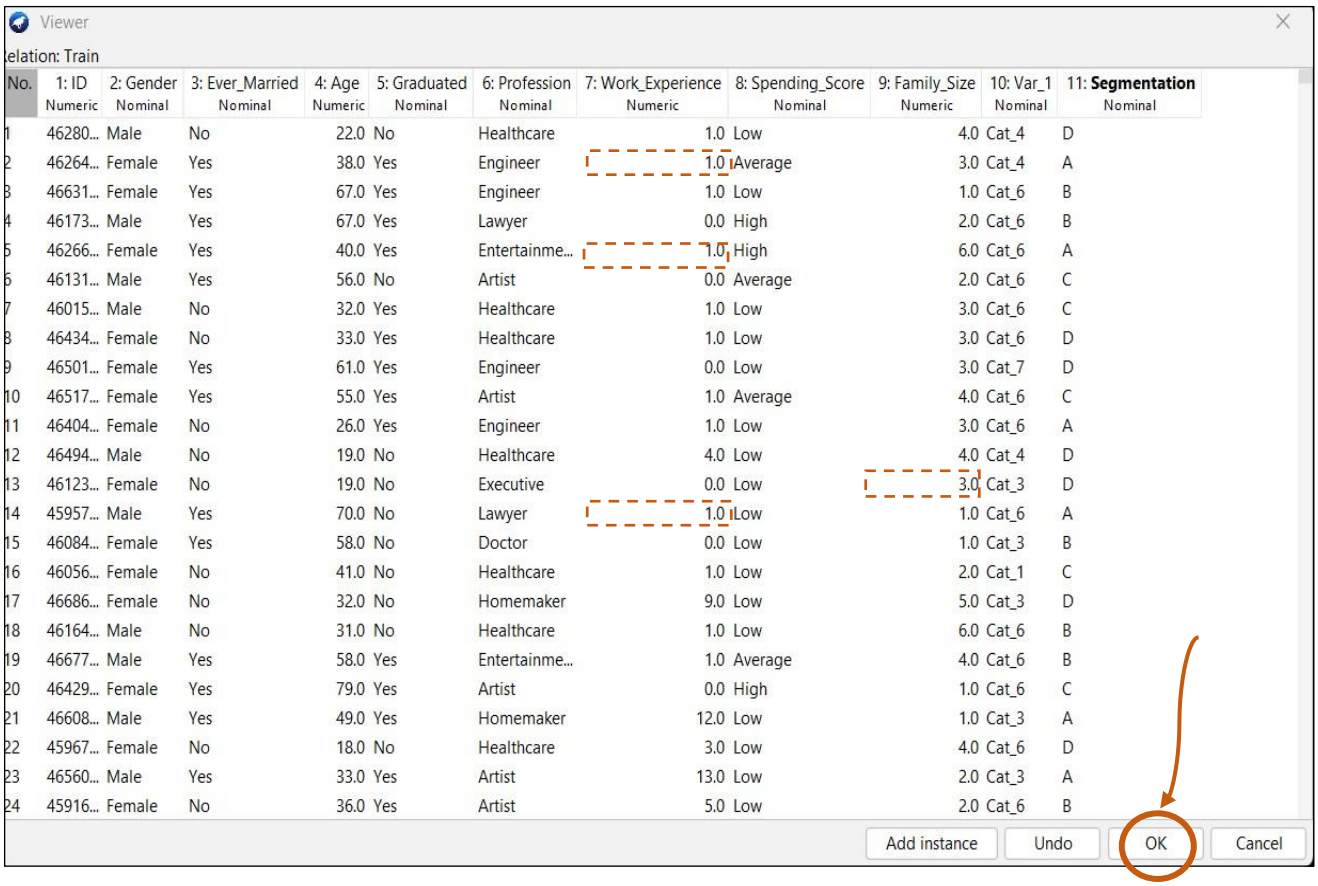




We can see some of messing value



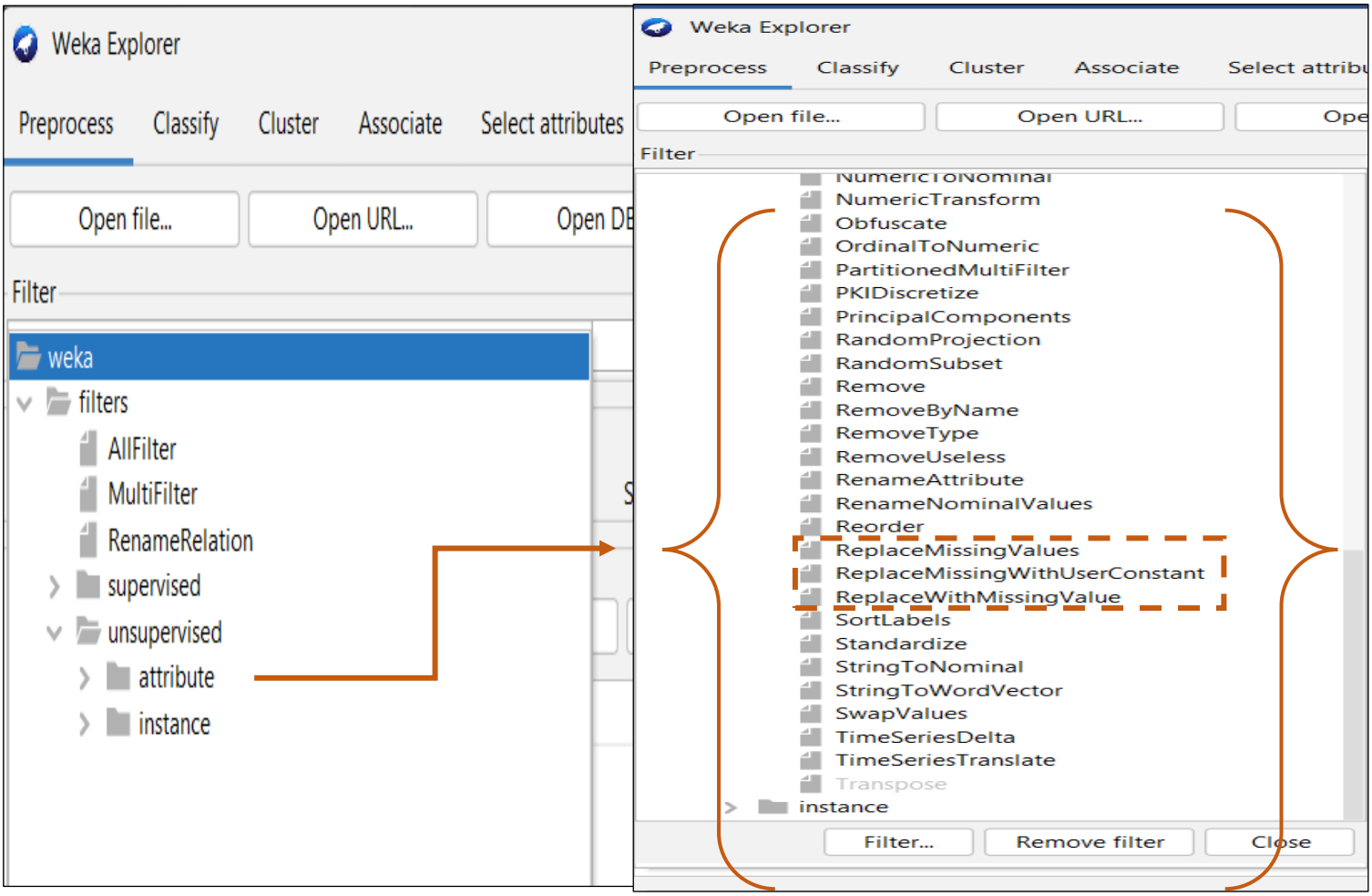
After replacing them we click " **OK** " to save the changes we made

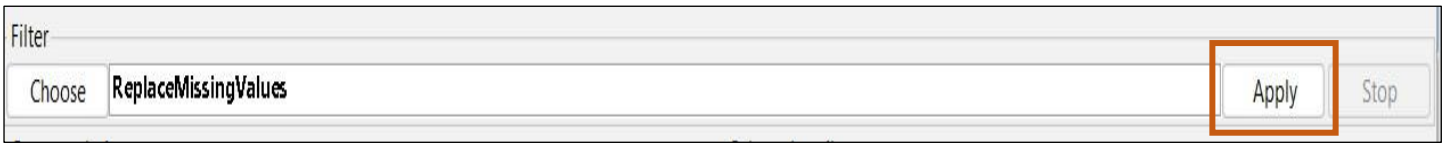


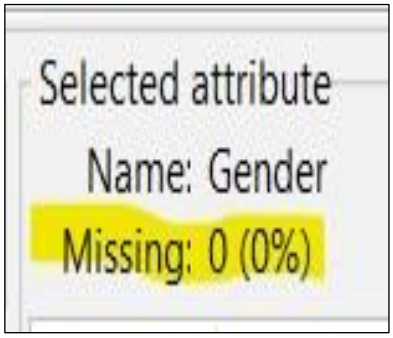
Columns " Ever_Married " , " Graduated " , " Profession " , " Var_1 " have null values.

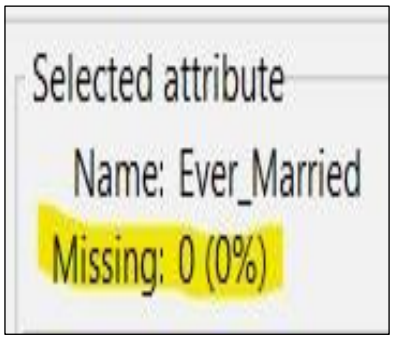
All of them either nominal or binary, so can be replaced with their **Mode** values.

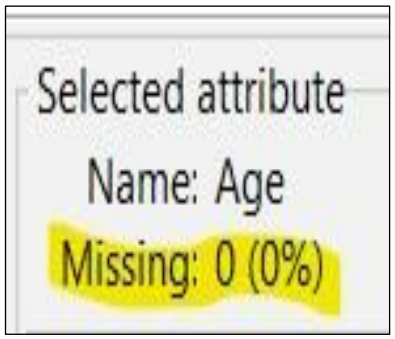
Note: in weka there is 3 replacing filter I chose " ReplaceMissingValues " because this filter allow replacing the missing with Mode value.

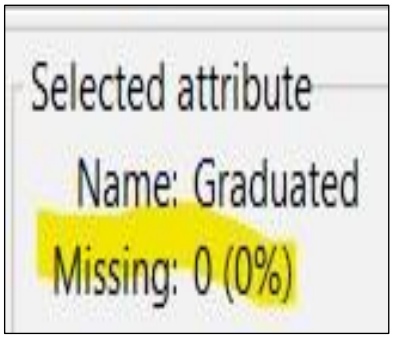


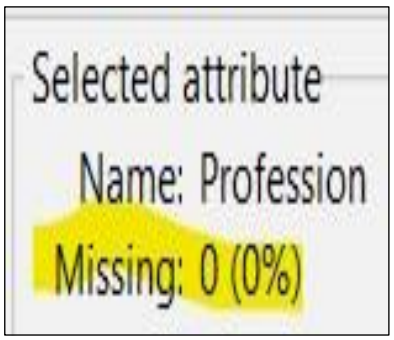


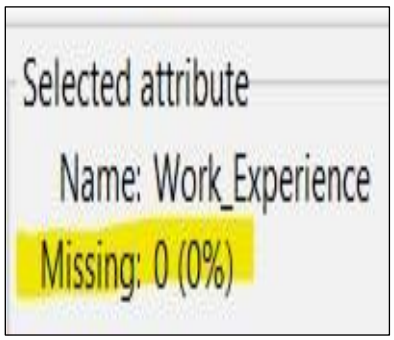


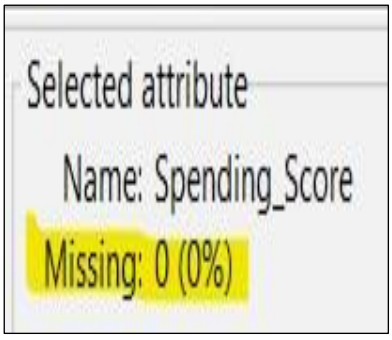


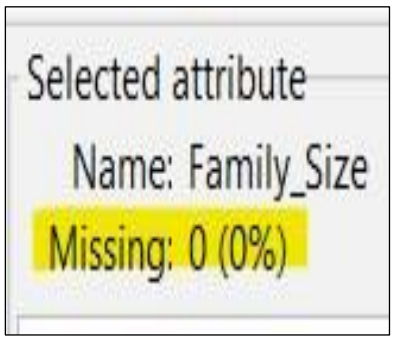


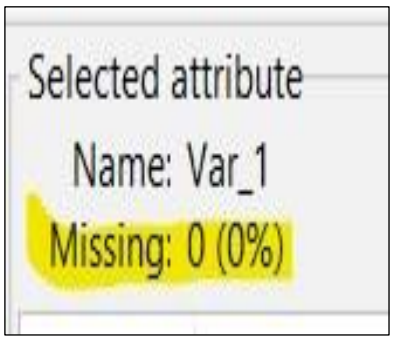












Now as we can see we replace all missing values

(**No missing values now**)

Outlier/ extreme values:

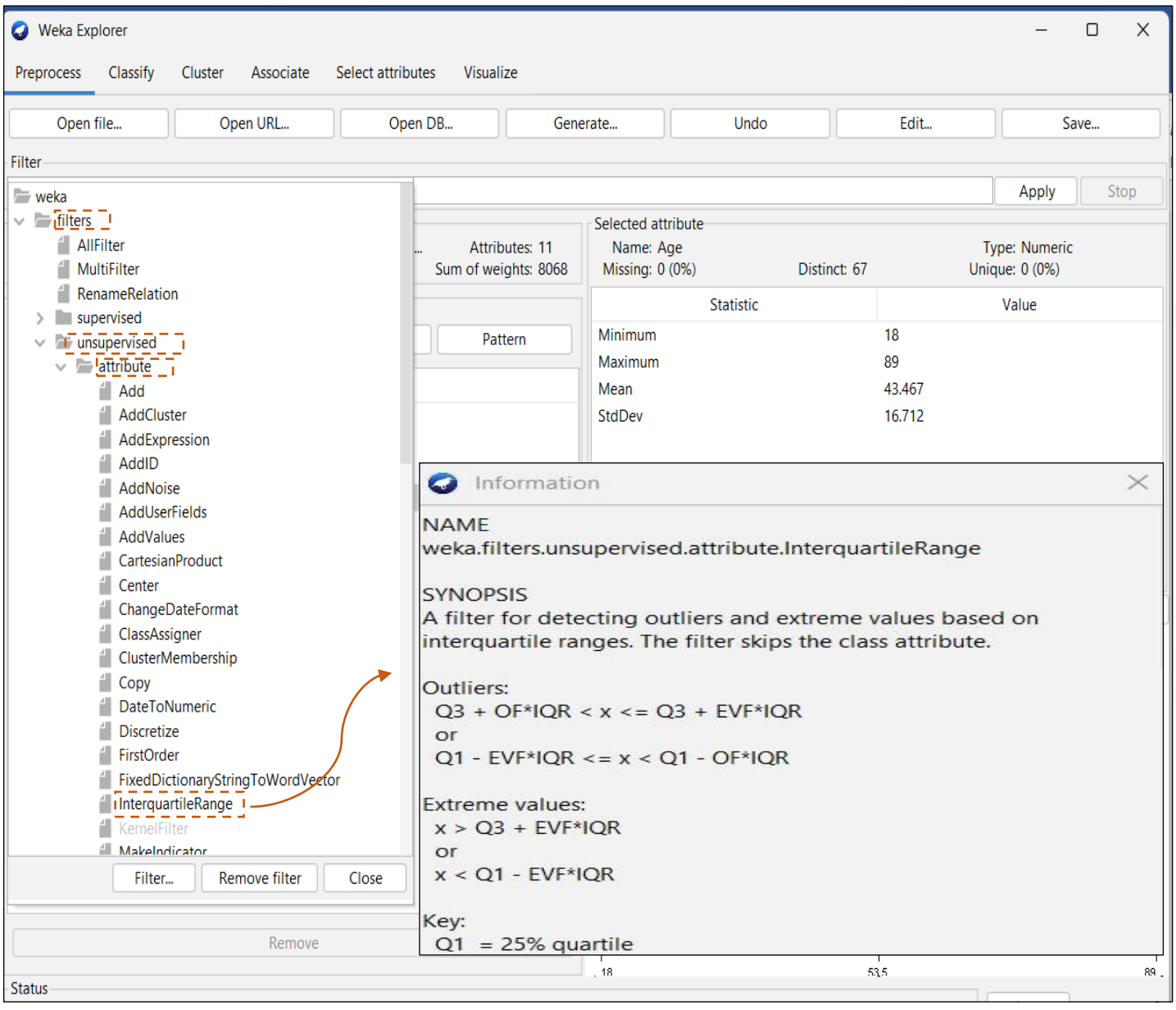
Now we need to find if there are any outlier/ extreme values in our data set:

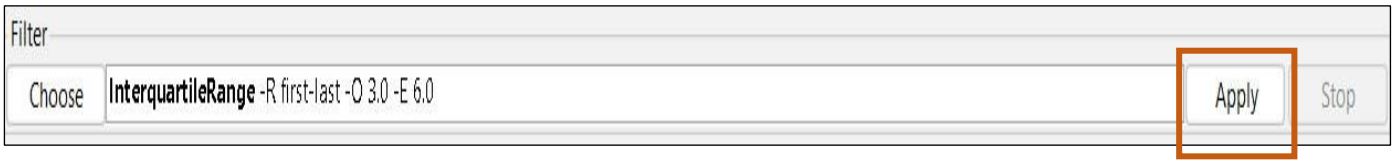
A common rule for identifying suspected outliers is to single out values falling at least 1.5×IQR above the third quartile or below the first quartile.

Finding them in weka:

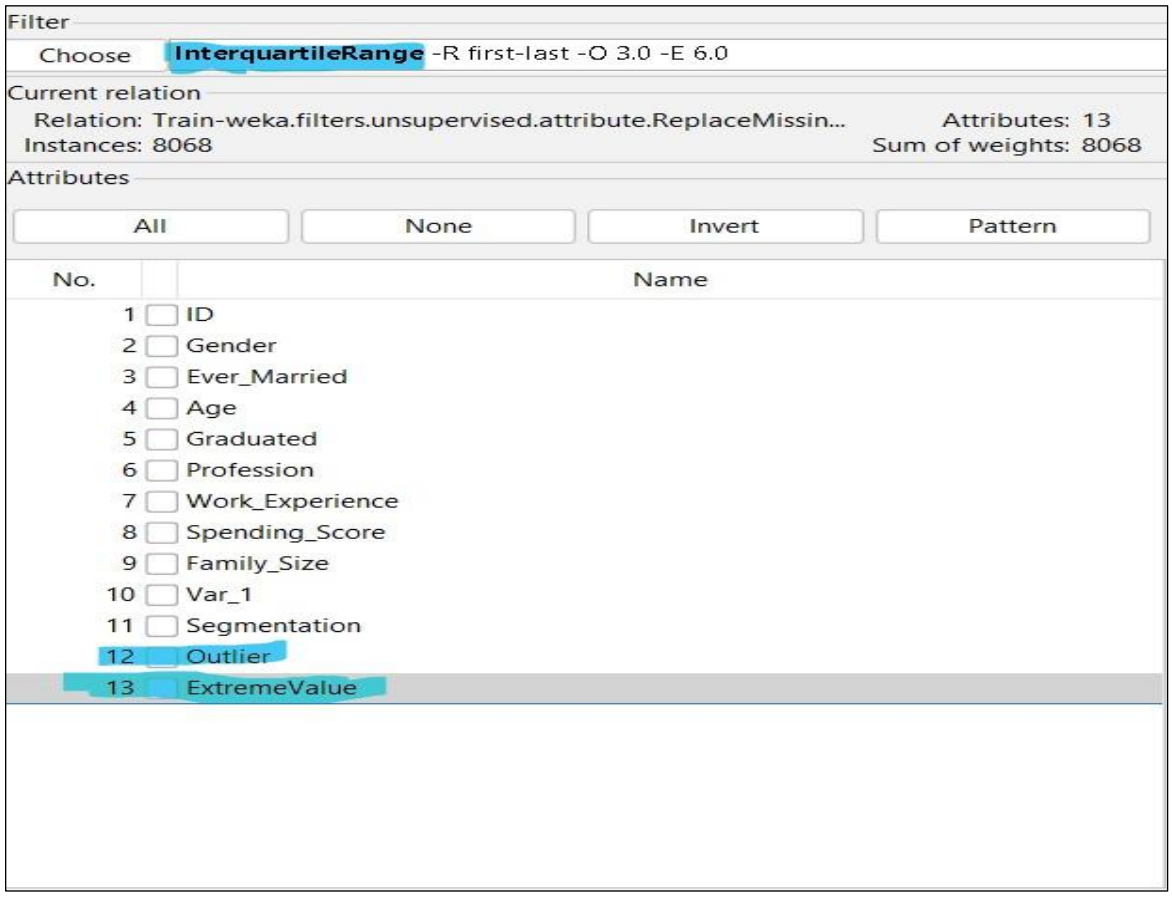
***Note**: **First you must process the missing values as above**

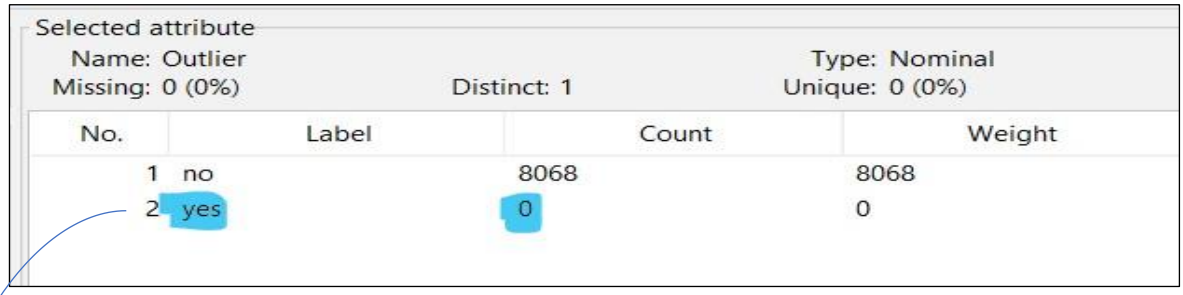
Filter → unsupervised → attribute → InterquartileRange (IQR)

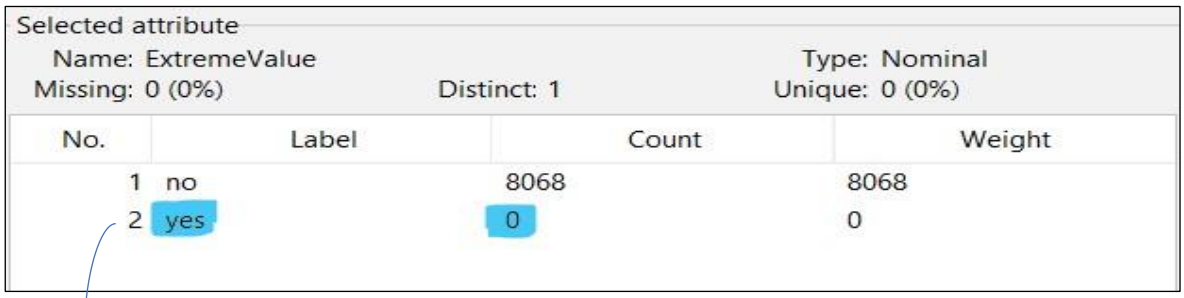




After we apply this filter 2 new column added to dataset (Outlier , ExtremeValue):







In our dataset there is no Outlier/Extreme

Note:

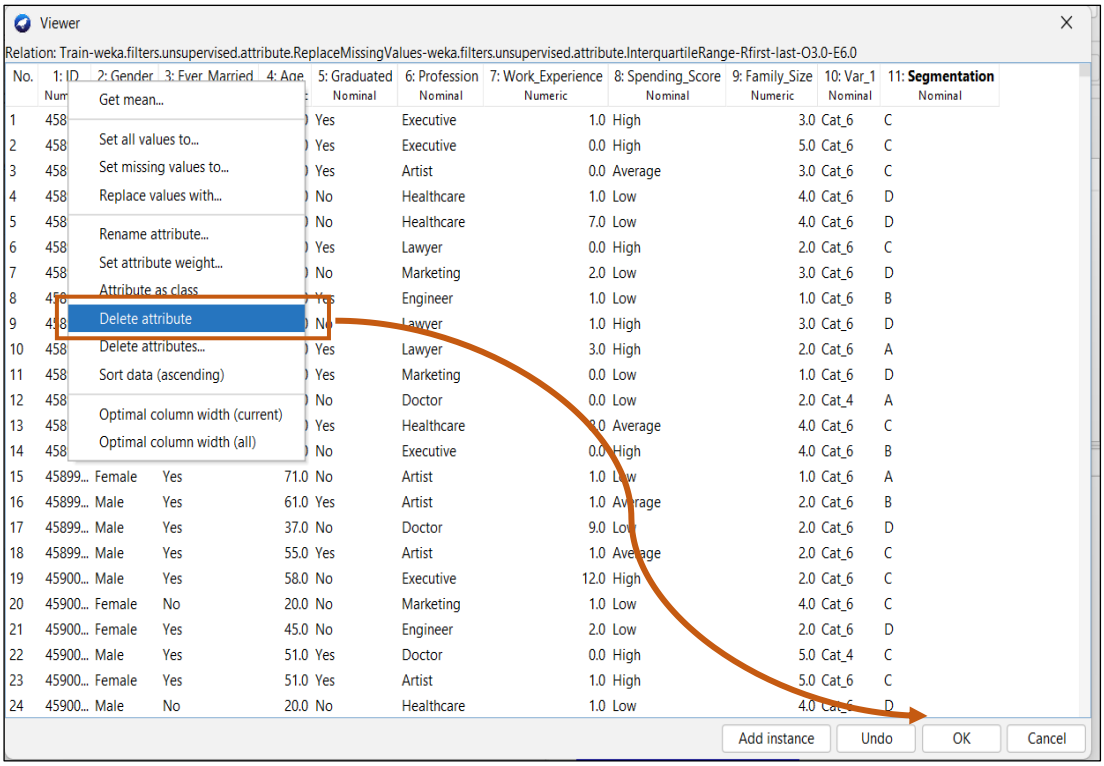
If there were outliers, we could solve them by using this filter which remove all instances that contain the outliers:

Filter →unsupervised→ instance→ **RemoveWithValue.**

After completion of the process, the variables (Extreme Values, outliers) deleted.

Note:

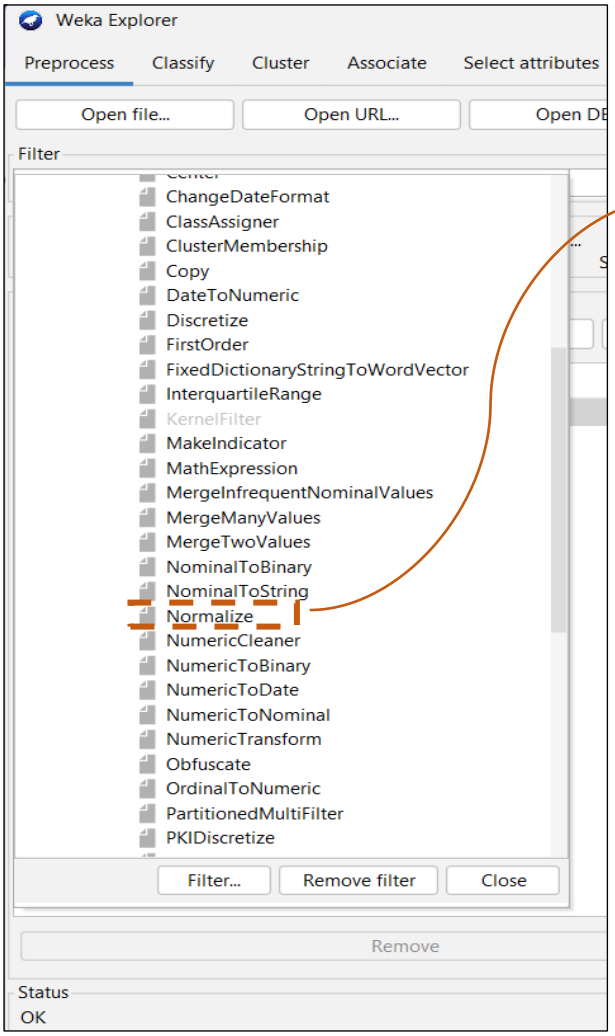
we can Remove column ID as it´s not important to the model



Note:

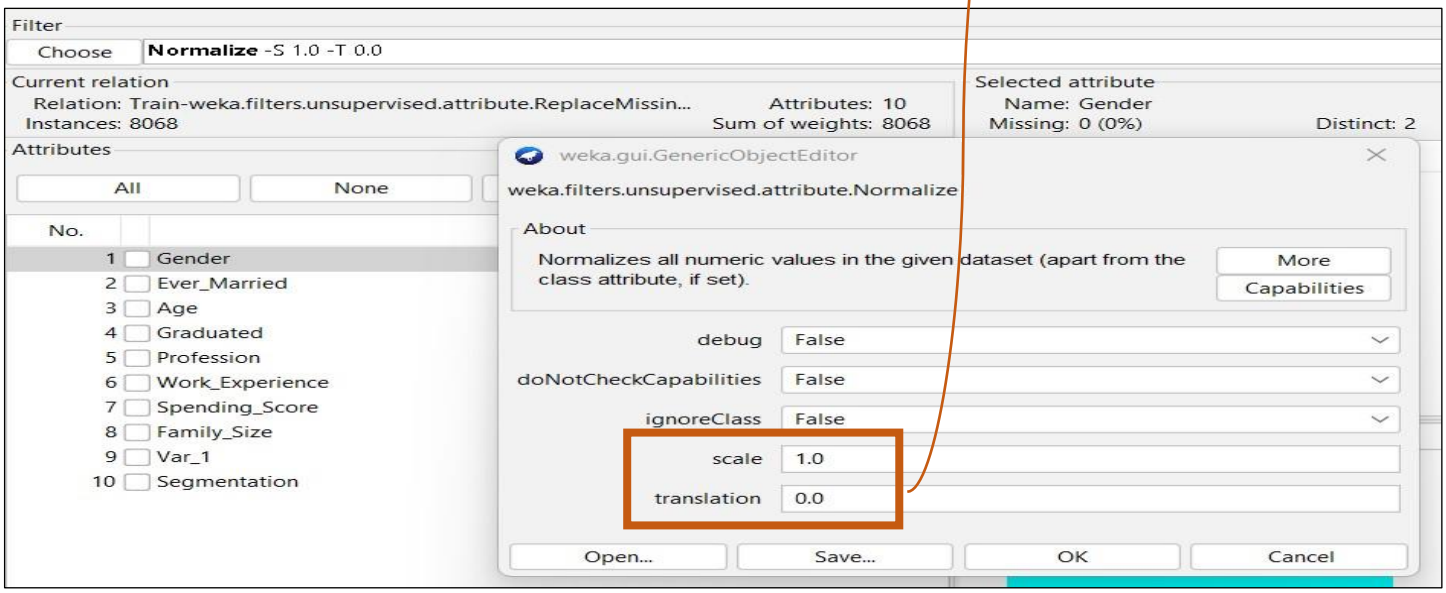
Because we need to make clustering for data, All Numeric attributes should be transformed to a similar scale to be effective:

Filter →unsupervised→ instance→ **Normalize.**



Normalizes all numeric values in the given dataset.

The resulting values are in [0,1] for the data used to compute the normalization.

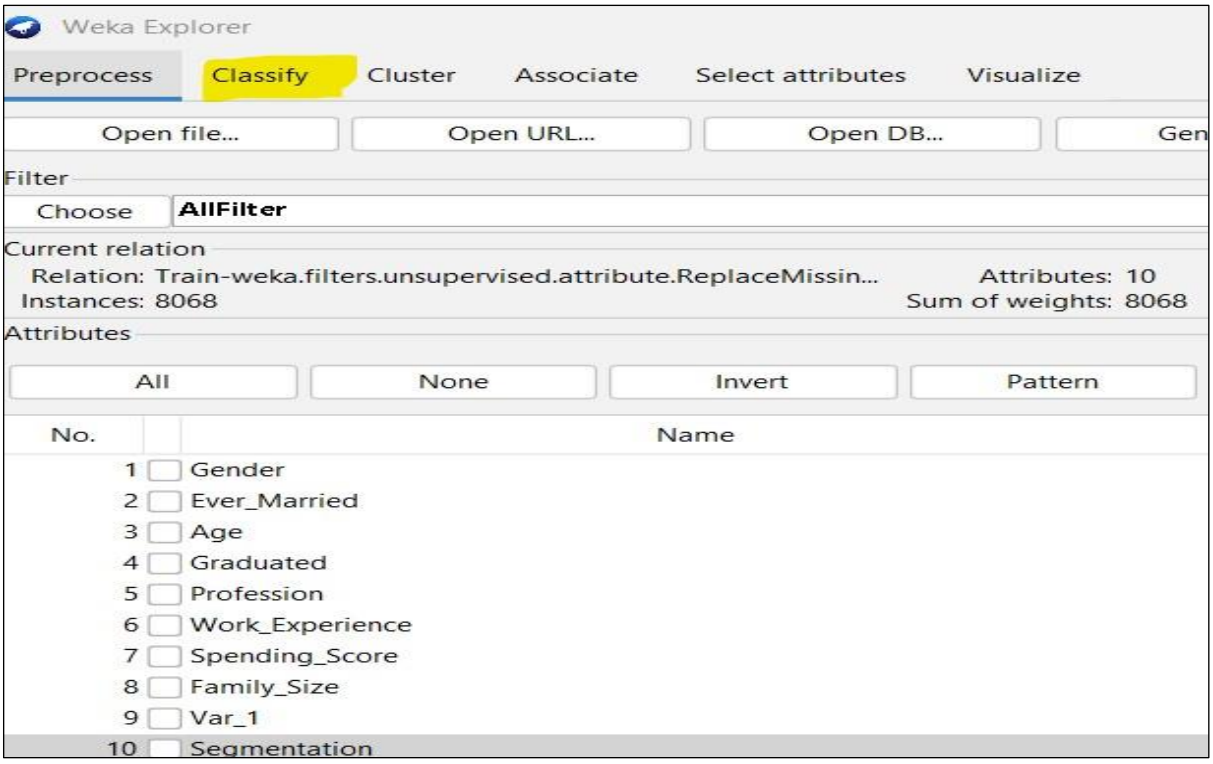


We finish data preparation (fill in missing values and get rid of the extreme values and outliers (if there is)).

Now our dataset ready for applying data mining model

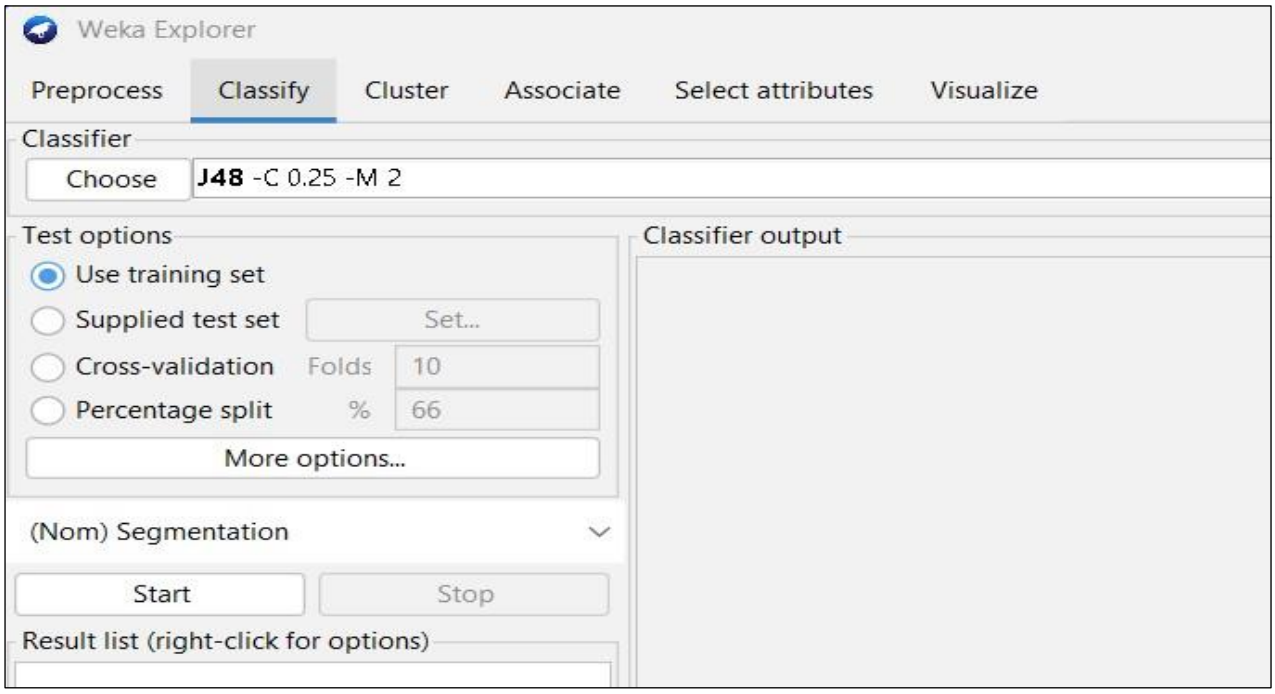
Data Mining Modeling

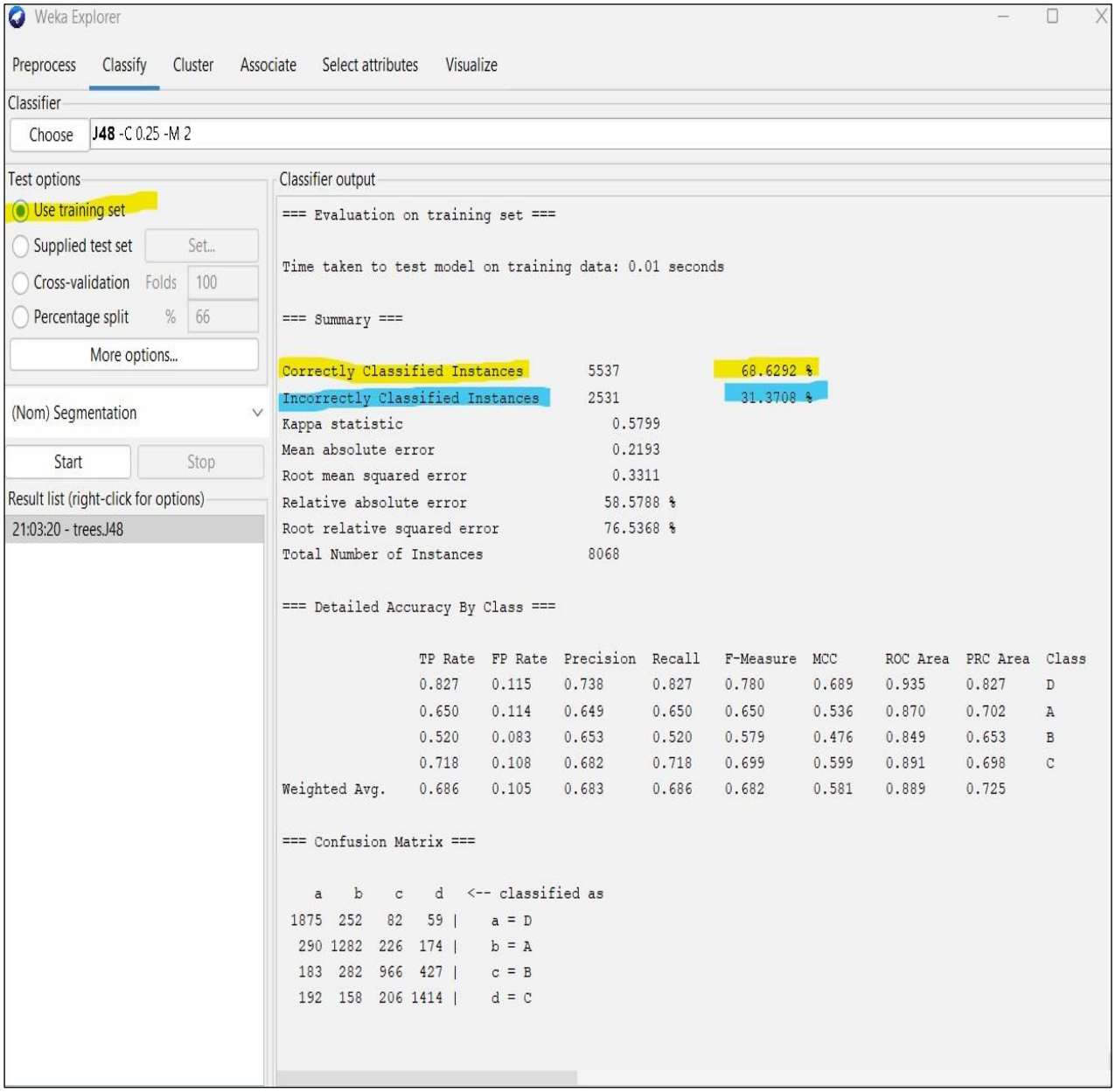
Classification:



Classification using Decision Tree:

Classify → choose → tree →<u>j48</u>



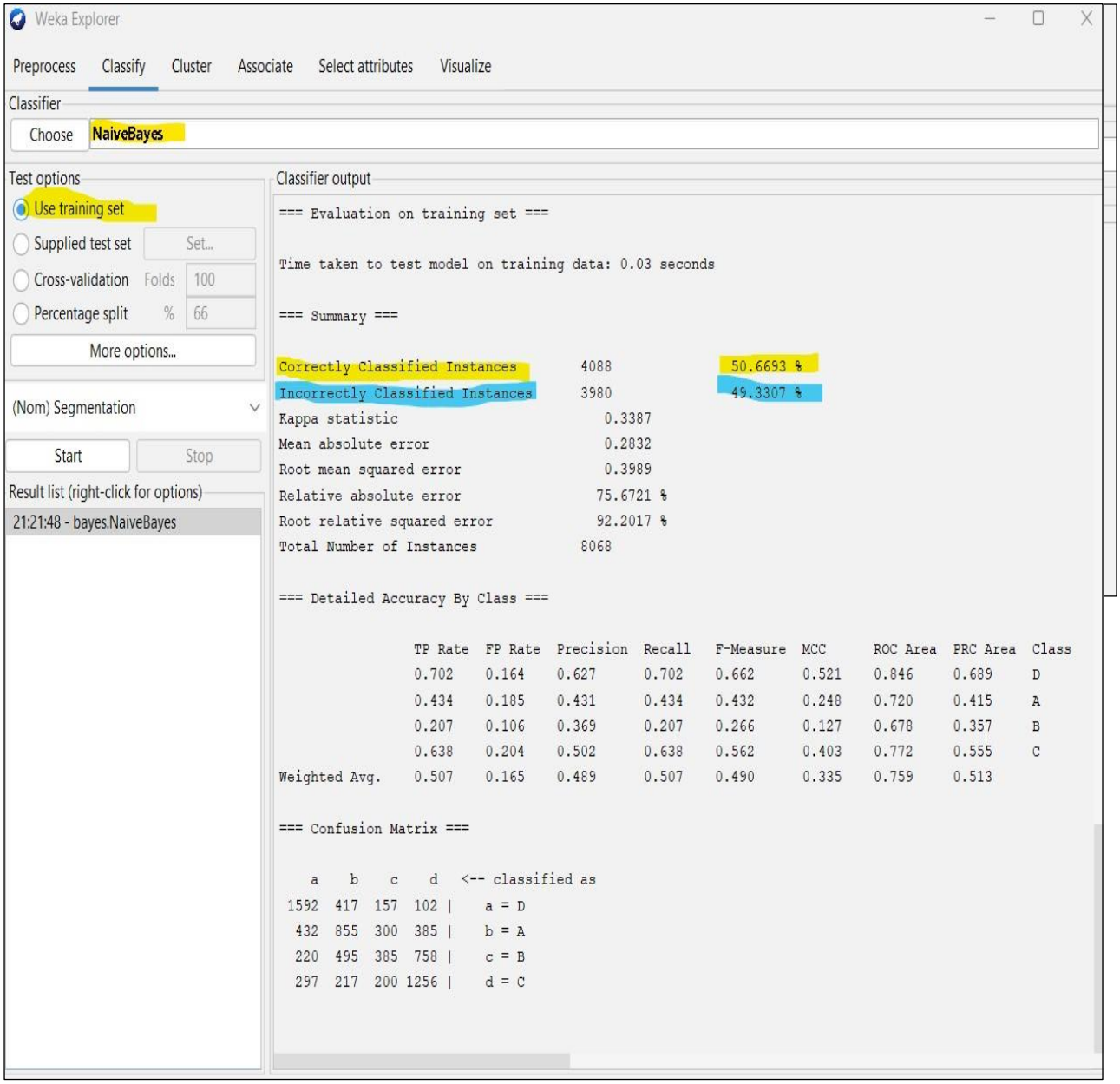


We can see that the **correctly** classified instance is: 68.6292%

We can see that the **incorrectly** classified instance is: 31.3708%

Classification using Naïve Bayes:

Classify → choose → bayes→**NaiveBayes**



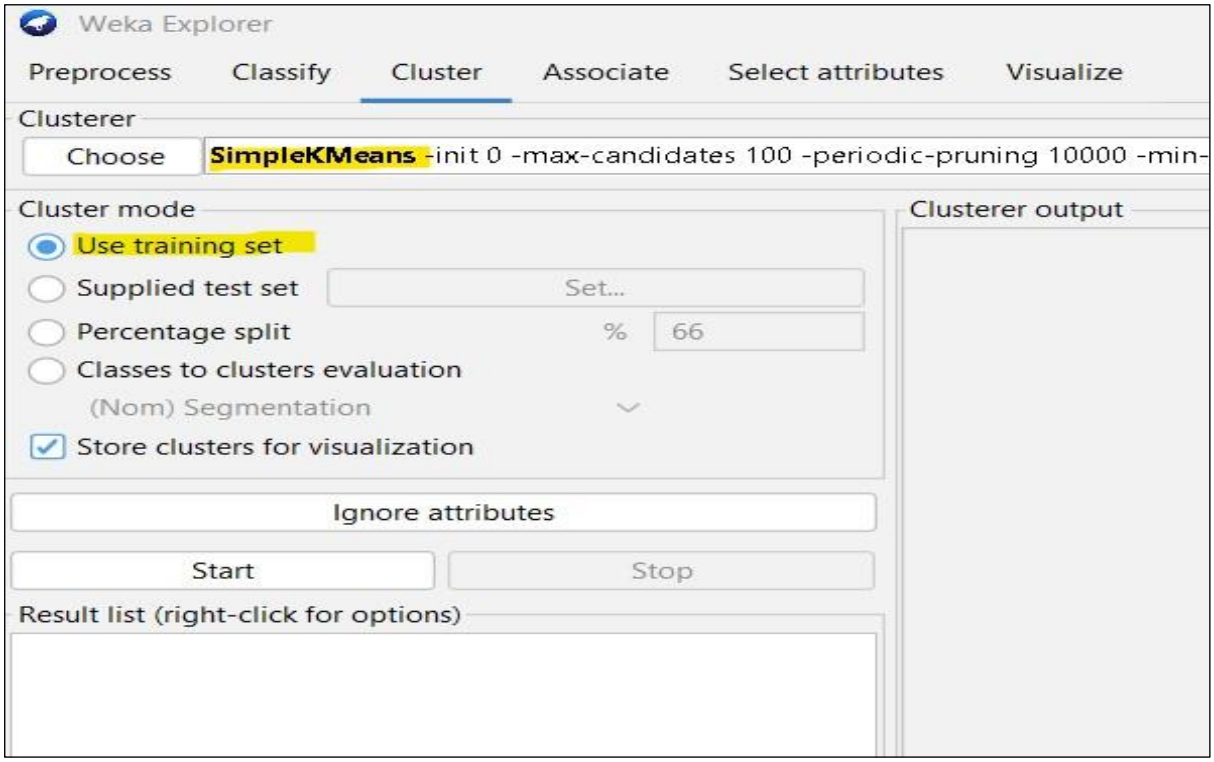
We can see that the **correctly** classified instance is: 50.6693%
We can see that the **incorrectly** classified instance is: 49.3307%

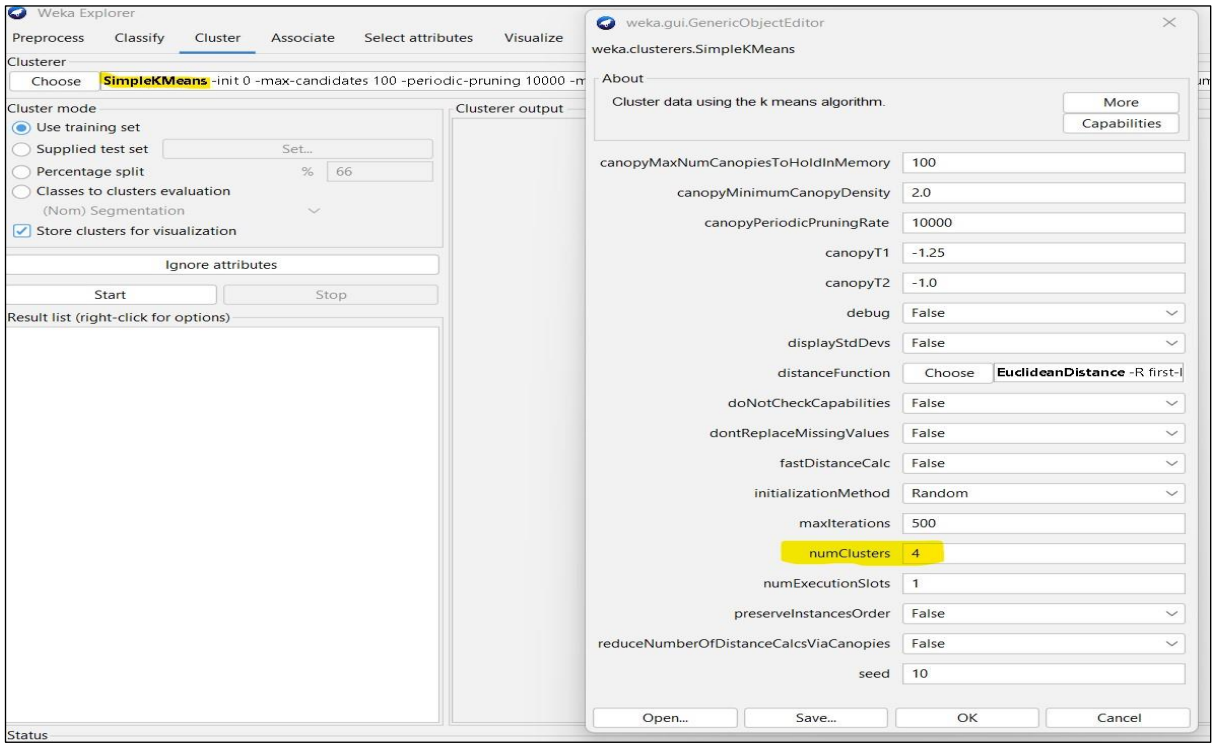
We can see that classification by using the first model (j84 Decision Tree) more accurate than classification by using the second model (naïve Bayes).

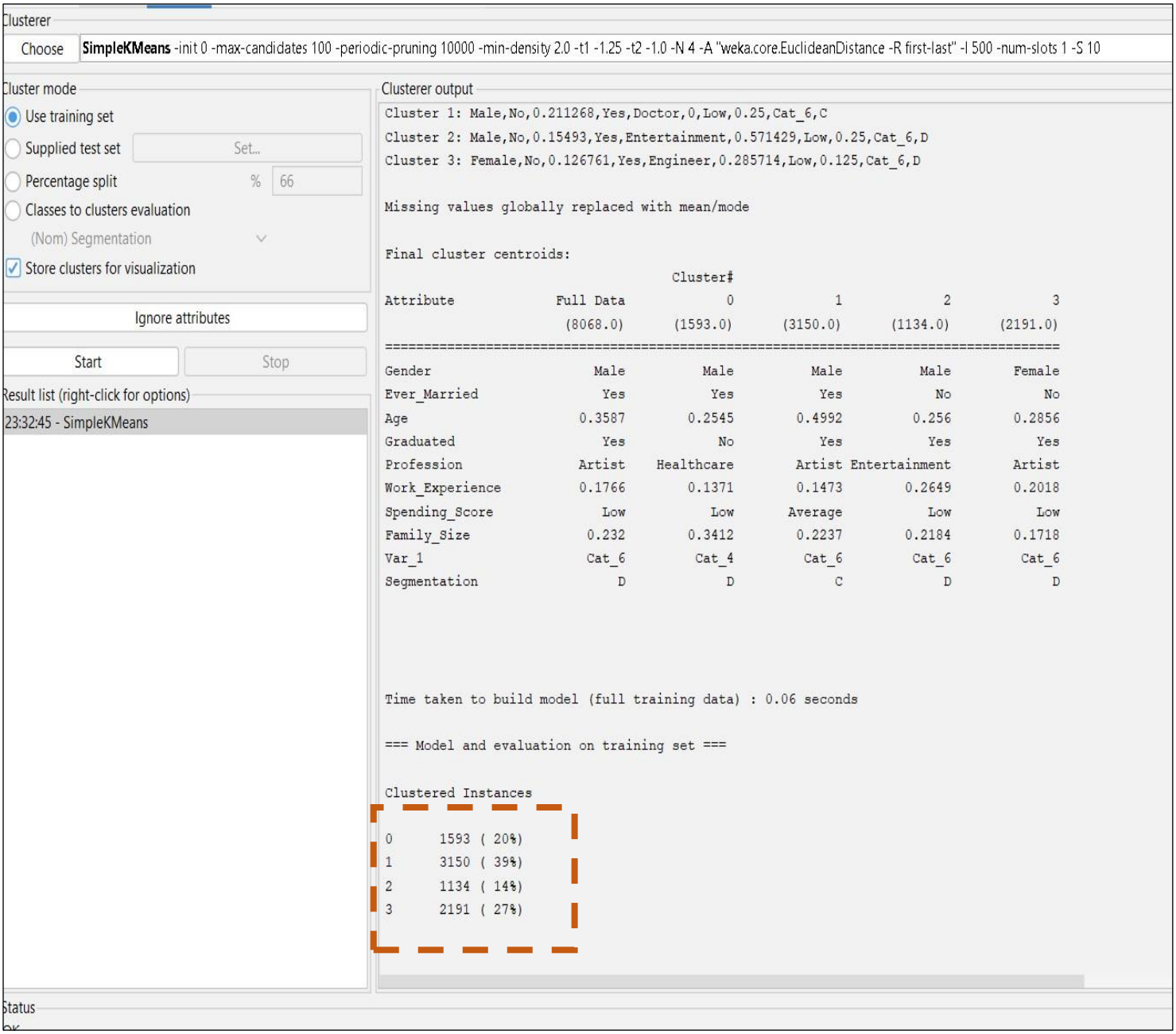
So, we chose the j48 Decision Tree for Classification

Clustring:

Simple K-Means






Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) Segmentation
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
23:32:45 - SimpleKMeans
Clusterer output
Cluster 1: Male,No,0.211268,Yes,Doctor,0,Low,0.25,Cat_6,C
Cluster 2: Male,No,0.15493,Yes,Entertainment,0.571429,Low,0.25,Cat_6,D
Cluster 3: Female,No,0.126761,Yes,Engineer,0.285714,Low,0.125,Cat_6,D
Missing values globally replaced with mean/mode
Final cluster centroids:
Cluster#
Attribute Full Data 0 1 2 3
(8068.0) (1593.0) (3150.0) (1134.0) (2191.0)
Gender Male Male Male Male Female
Ever_Married Yes Yes Yes No No
Age 0.3587 0.2545 0.4992 0.256 0.2856
Graduated Yes No Yes Yes Yes
Profession Artist Healthcare Artist Entertainment Artist
Work_Experience 0.1766 0.1371 0.1473 0.2649 0.2018
Spending_Score Low Low Average Low Low
Family_Size 0.232 0.3412 0.2237 0.2184 0.1718
Var_1 Cat_6 Cat_4 Cat_6 Cat_6 Cat_6
Segmentation D D C D D
Time taken to build model (full training data) : 0.06 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 1593 (20%)
1 3150 (39%)
2 1134 (14%)
3 2191 (27%)
Status

Decision

In this project we went through all the process from defining the business objective, knowing that dataset, clean and preparing the data exploring features and distributions, data modelling and presenting different algorithms to select the best to predict the Customer Segmentation, what will help the business adopt the best marketing strategies to each of them and bring more market share and revenue to the company.

The chosen model was j84 model since it´s the most accurate, although it doesn't present a high accuracy. We could reach a more accurate model having more data about customers, it´s something to explore and go deeper in the organization with the business team and the data engineer in order to explore if more relevant features are available.

The End