Data Mining Project

Faculty of Prince Al-Hussein Bin Abdallah II
for Information Technology

The Hashemite University

Presented to:
Dr. Subhieh Moh'd Faraj S. Elsalhi

Prepared by:

Mahmoud Ibrahim Mahmoud Aqel 1936150

May 2023

part one

1.1 Initial data exploration and preparation

1.1.1 A1 Identify the type of each attribute.

Attribute Name	Attribute Type
Age	Ratio
Work class	Nominal
Education	Nominal
Education-Num	Ratio
Marital-Status	Nominal
Occupation	Nominal
Relationship	Nominal
Race	Nominal
Gender	Nominal
Capital-Gain	Ratio
Capital-Loss	Ratio
Hours-Per-Week	Ratio
Native-Country	Nominal
Fnlwgt	Nominal

Table 1

1.1.2 A2 Identify the values of the summarizing properties for each attribute.

1. Location

1.1. Open weka

1.2. From right side click (Explorer)

1.3. From preprocess tab click (Open file)

1.4. Open your custom dataset file

1.5. From the attributes panel you can get attributes' name and its location.

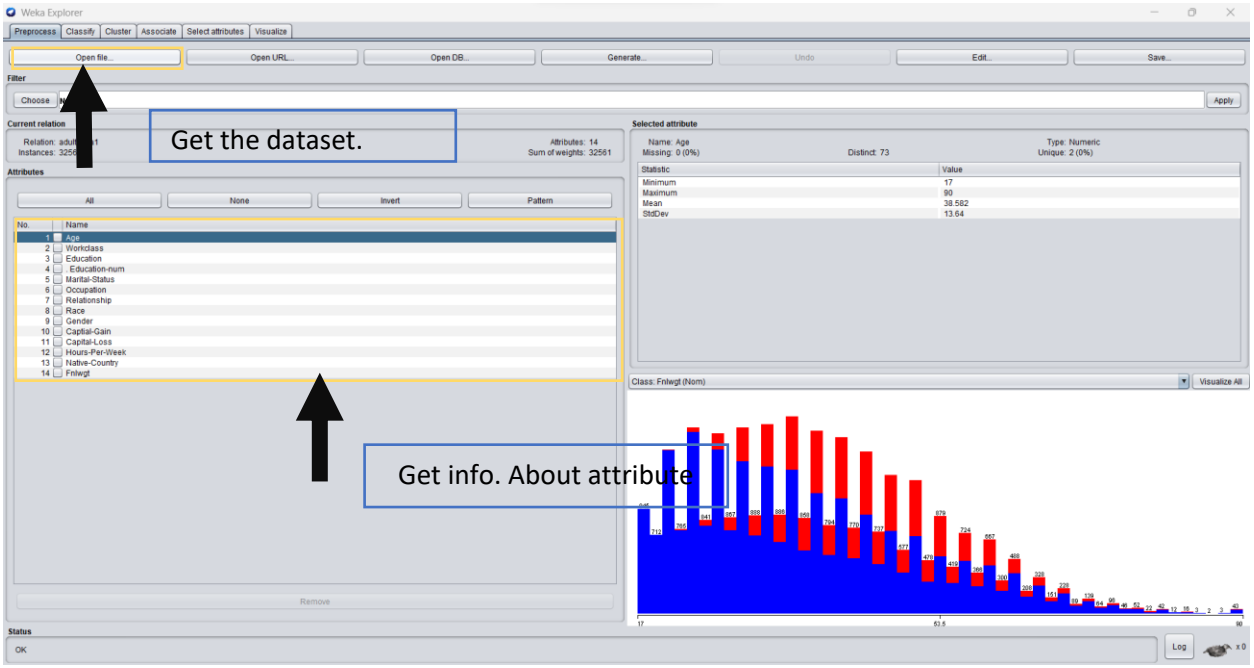

Figure 1

2. Max

2.1. Select the attribute that you want to get maximum instance value of it

2.2. From selected attribute panel you can get maximum instance value.

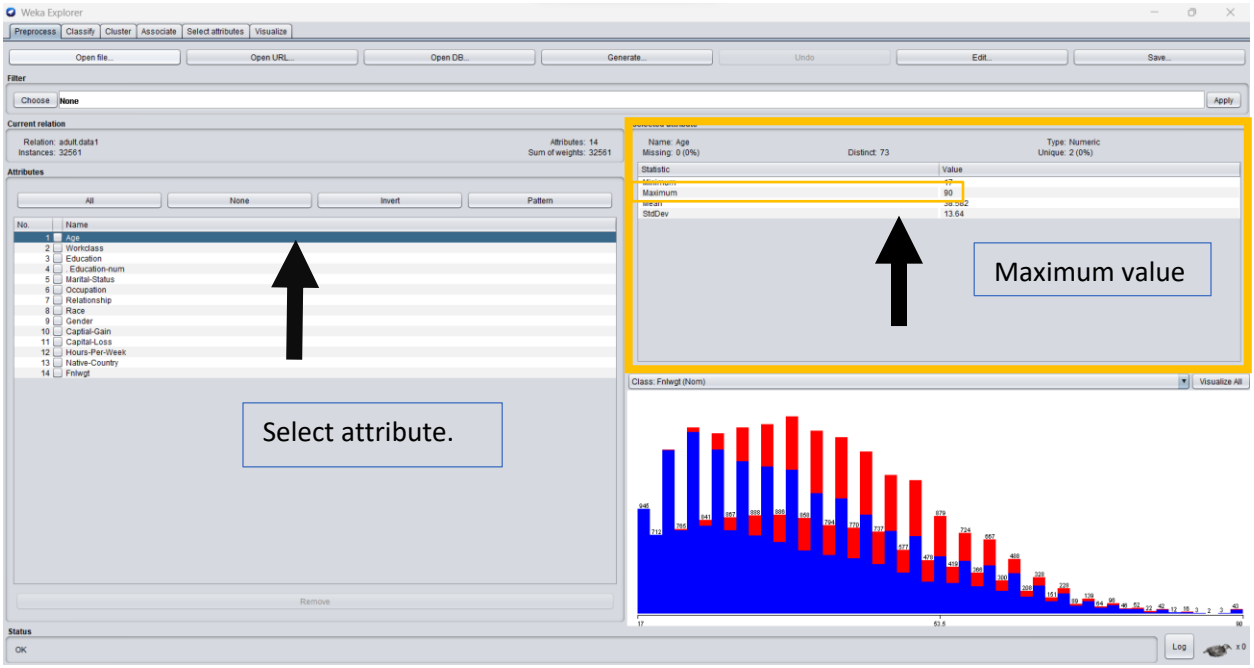

Figure 2

3. Min

3.1. Select the attribute that you want to get minimum instance value of it

3.2. From selected attribute panel you can get minimum instance value.

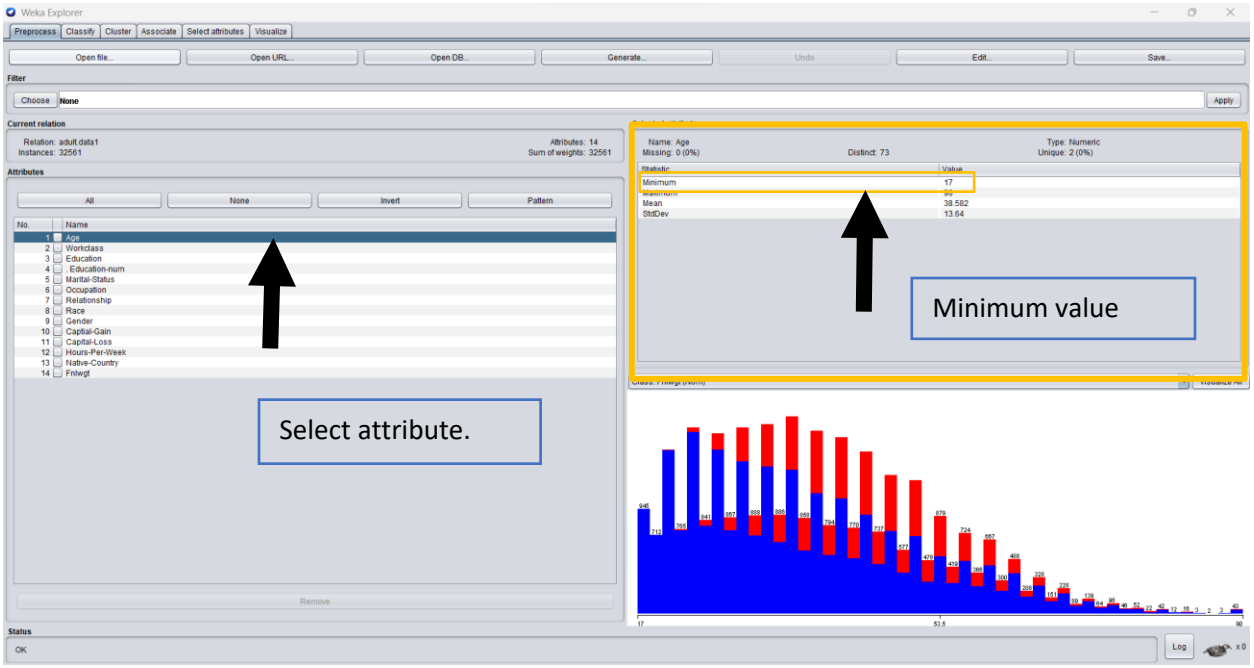

Figure 3

4. Range (non nominal)

4.1. range = max-min

5. Range(nominal)

5.1. Select attribute that you want to get range fot it from attribute panel

5.2. From selected attribute panel you can get range that have name (distinct).

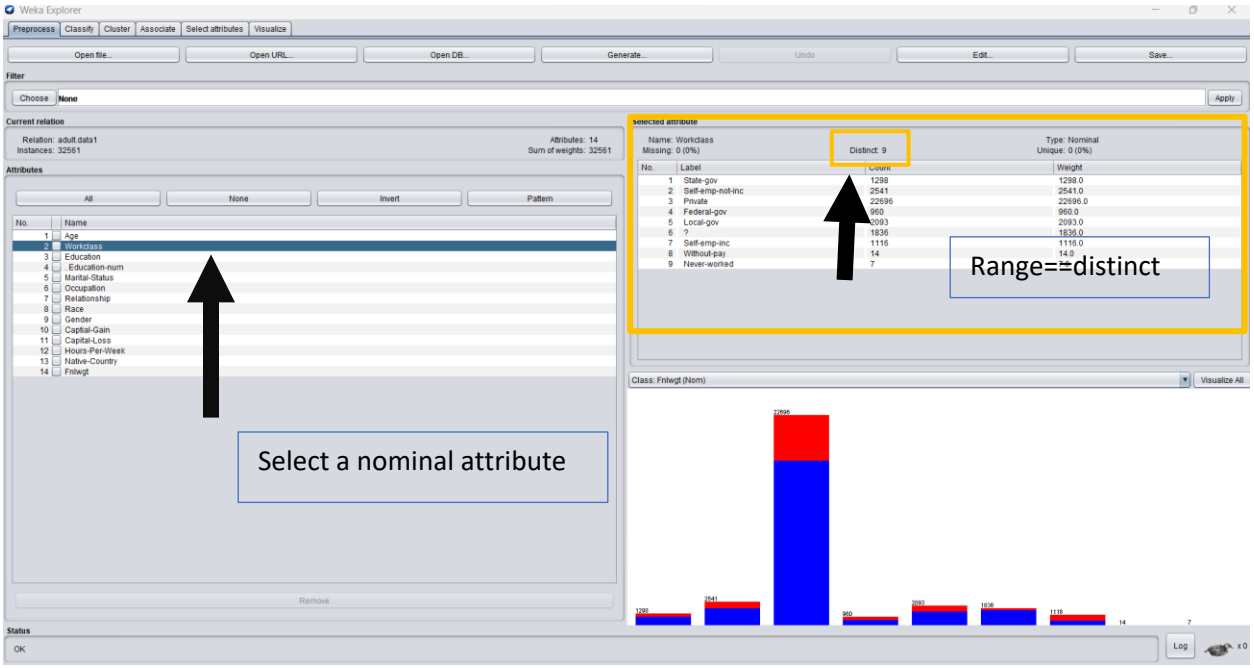

Figure4

6. Variance

6.1. open the dataset

6.2. select column and in any empty cell type VAR.S

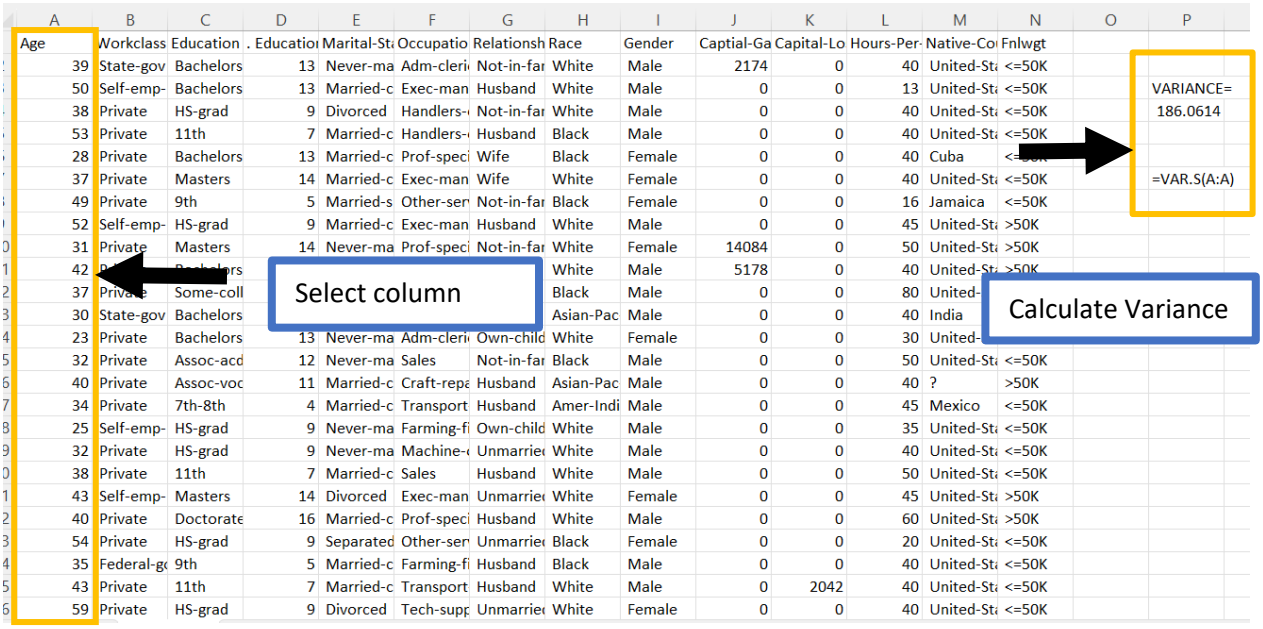

Figure5

7. Mean

7.1. Select attribute that you want to get mean fot it

7.2. From selected attribute panel you can get mean that have name (Mean).

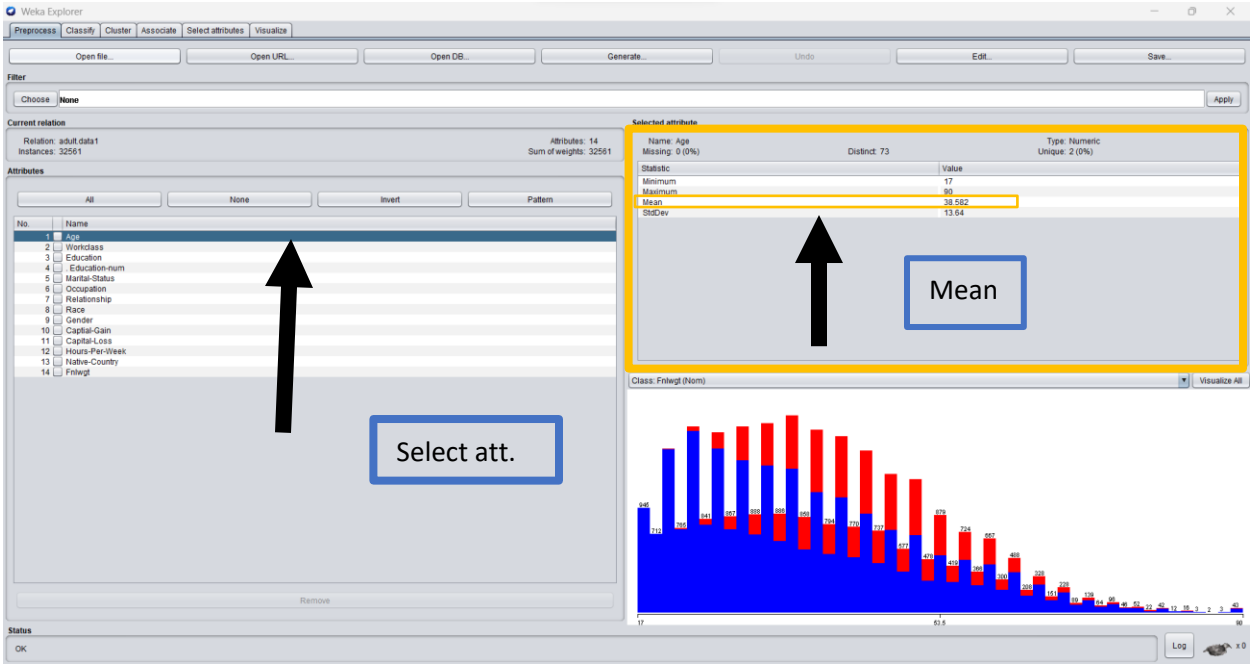

Figure6

8. Mood

8.1. open the dataset

8.2. select column and in any empty cell type MODE

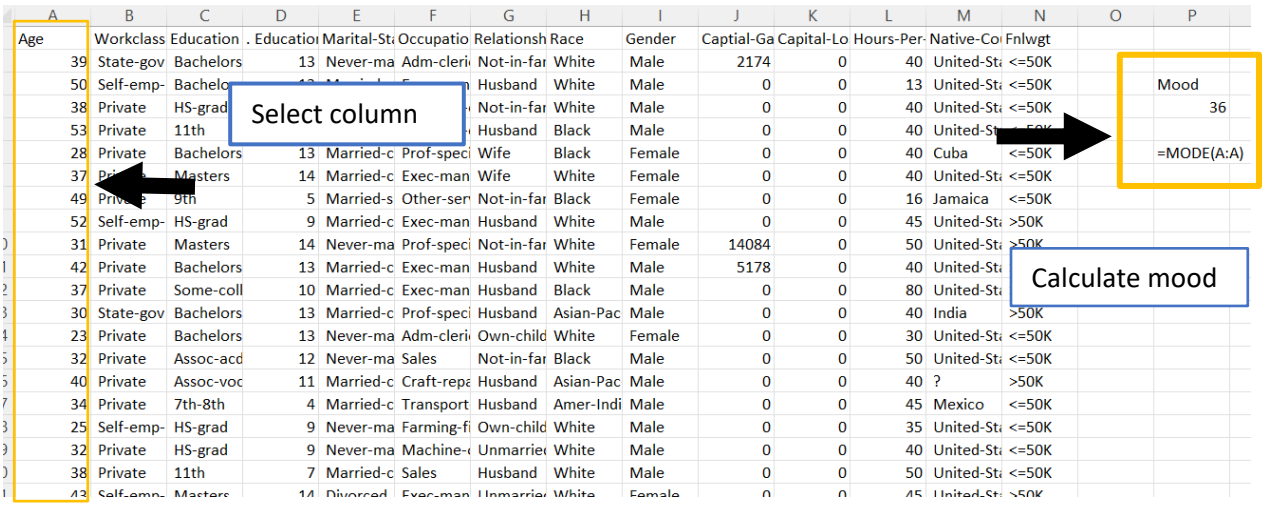

Figure7

9. Median

9.1. open the dataset

9.2. select column and in any empty cell type MEDIAN

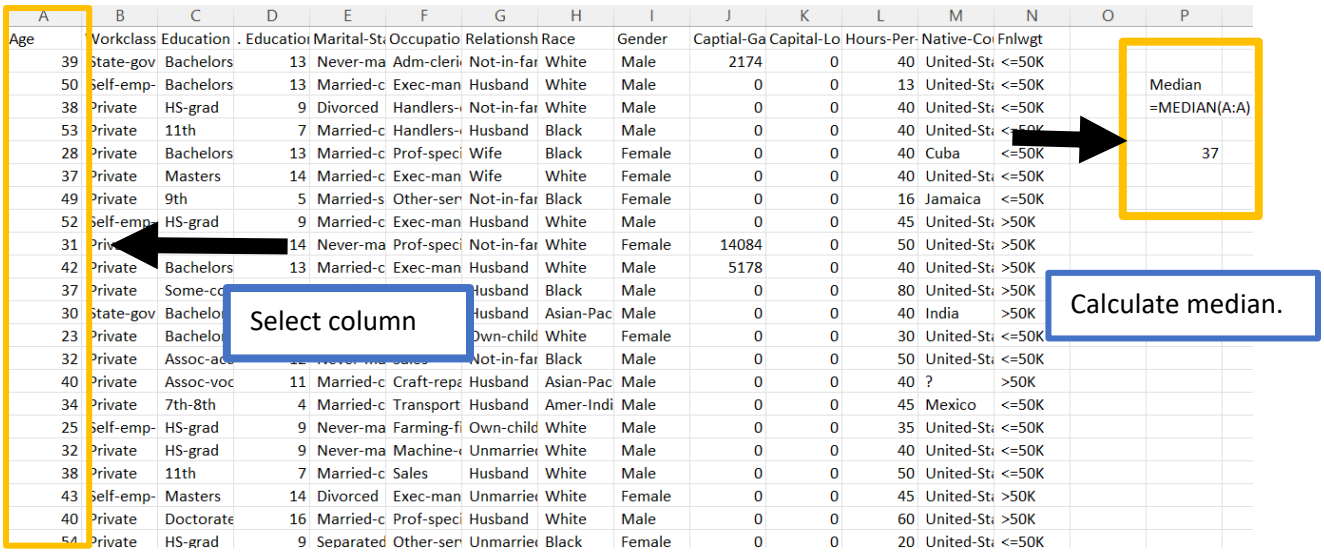

Figure8

10. Frequency

10.1. open the dataset

10.2. We have to copy the data you would like to get the frequency for it

10.3. select column and in any empty cell type COUNTIF

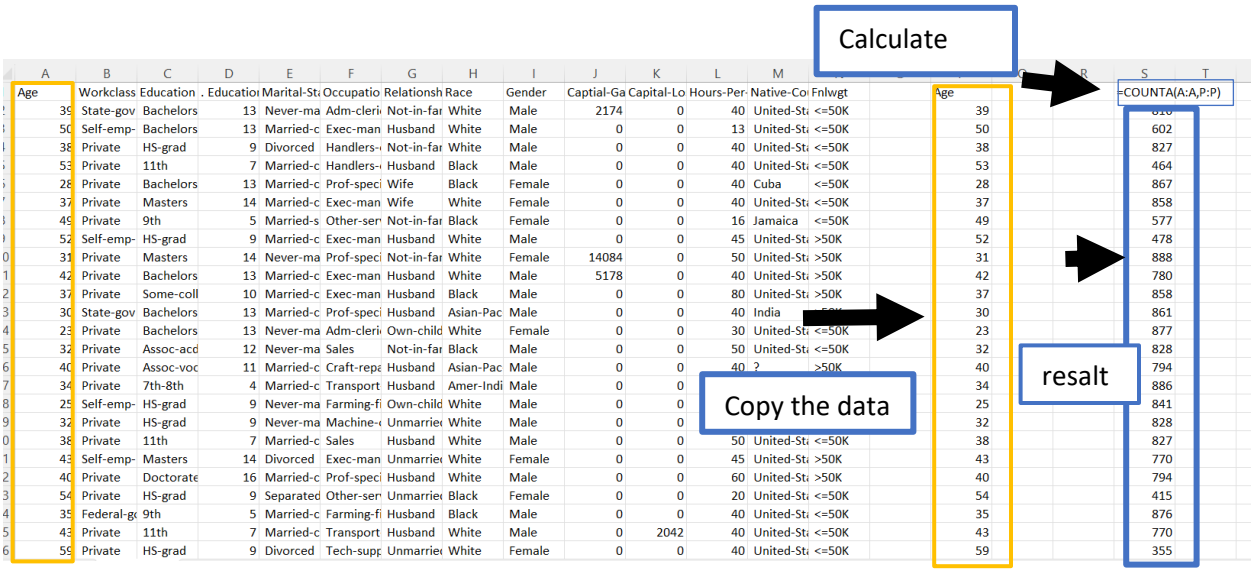

Figure9

Attribute Name	Location	Max	Min	Range	Variance	Mood	Meadian	Mean
Age	1st	90	17	73	186.0671095	36	53.5	38.582
Workclass	2nd	#	#	8	#	Private	#	#
Education	3th	#	#	16	#	HS-grad	#	#
Education-Num	4th	16	1	15	6.618831435	9	8.5	10.081
Marital-status	5th	#	#	7	#	Married-civ-spouse	#	#
Occupation	6th	#	#	14	#	Prof-specialty	#	#
Relationship	7th	#	#	6	#	Husband	#	#
Race	8th	#	#	5	#	White	#	#
Gender	9th	#	#	2	#	Male	#	#
Capital-gain	10th	99999	0	99999	54544177.45	0	49999.5	1077.615
Capital-loss	11th	4356	0	4356	162381.6909	0	2178	87.307
Hours-per-week	12th	99	1	98	152.4636717	40	50	40.437
Native-country	13th	#	#	41	#	United-States	#	#
Fnlwgt	14th	#	#	2	#	<=50K	#	#

Table 2

1.1.2 A3 Identify any outliers, clusters of similar instances, "interesting" attributes, and specific values of those attributes.

1. Outliers

1.1. Select class attribute

1.2. From filter panel press "choose"

1.3. Go to Filter–>Unsupervised–>Attribute–>Interquartile Range.

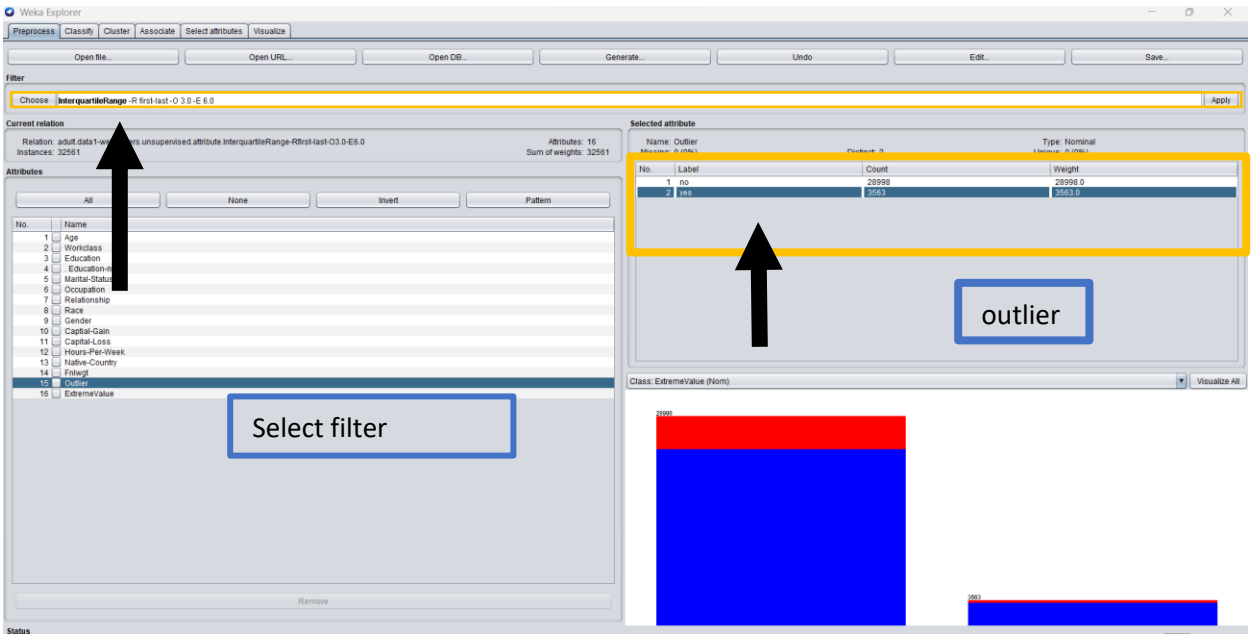

Figure10

2. Cluster

2.1. from cluster tab

2.2. choose EM class

2.3. hit start

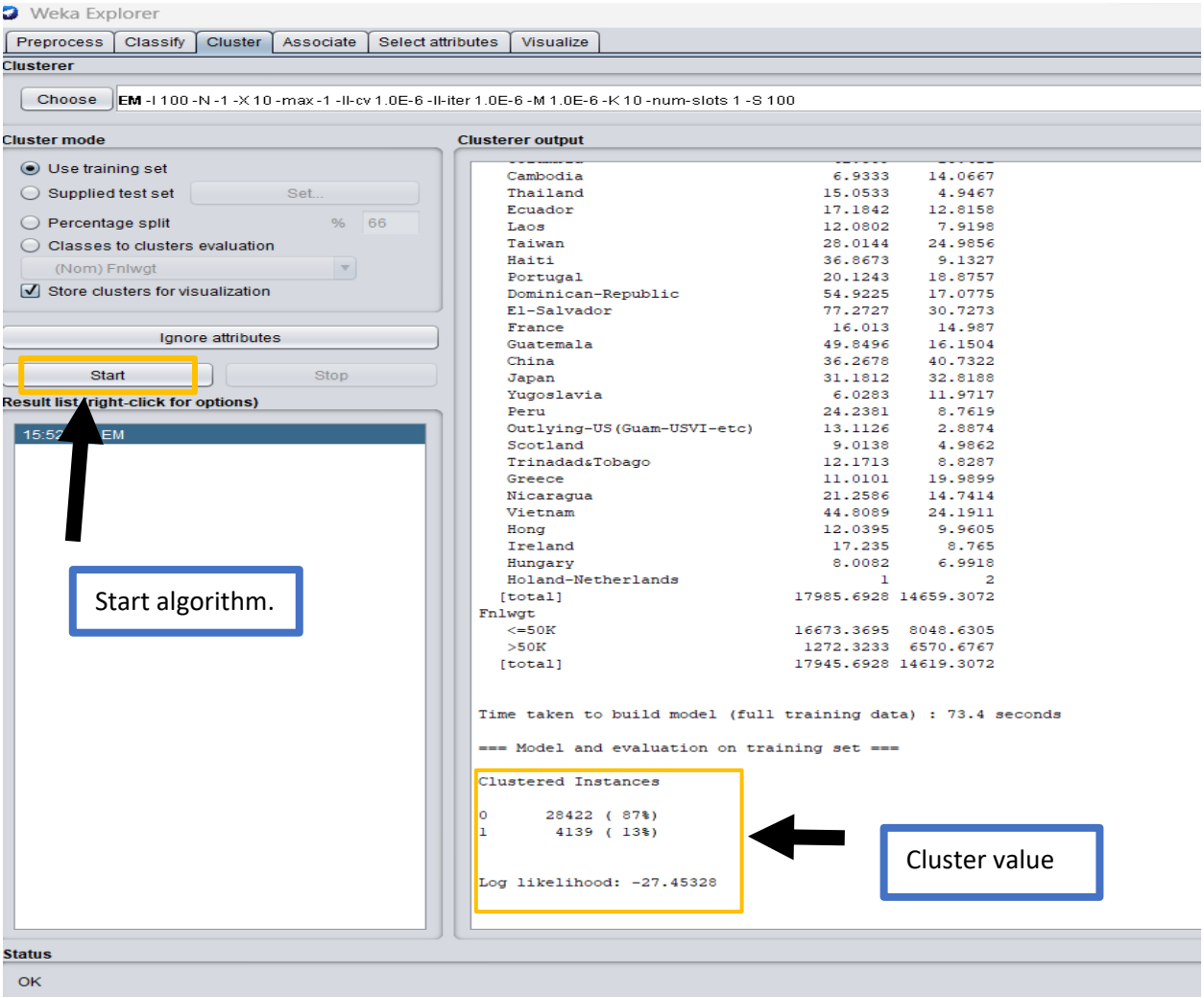

Figure11

3. Scatter Plots

3.1. from visualize tab

3.2. you can see the relation between attributes.

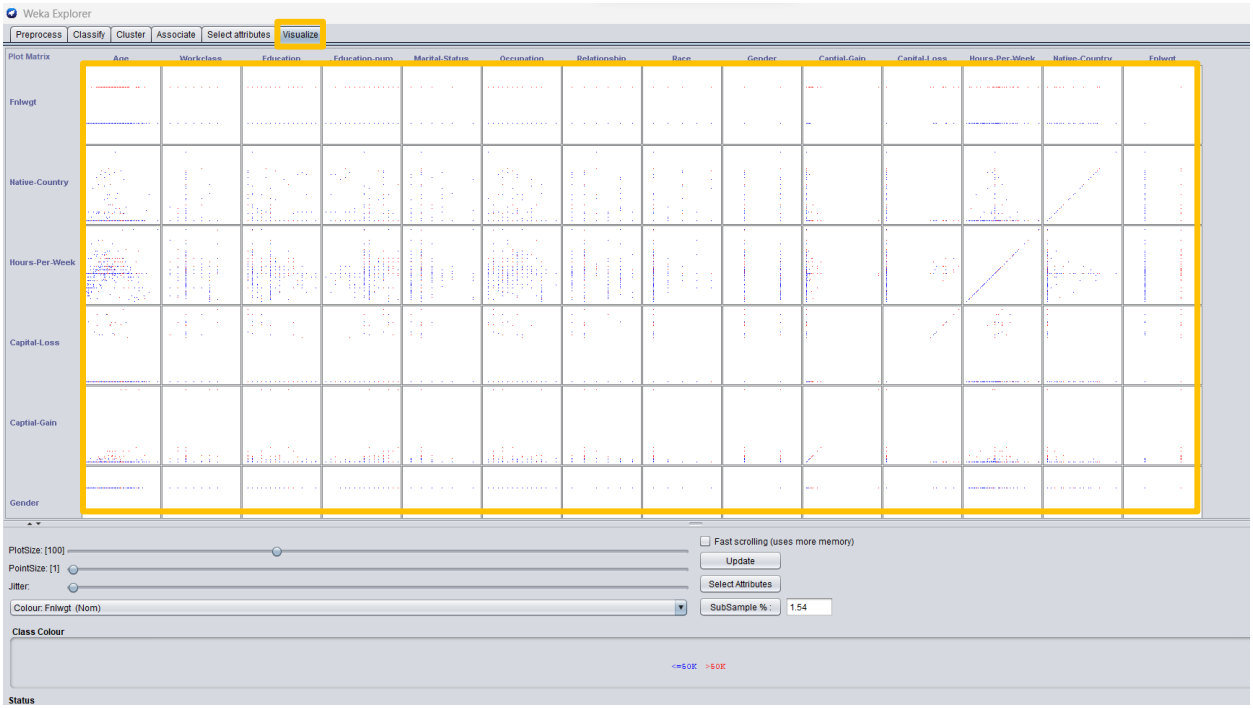

Figure12