

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.

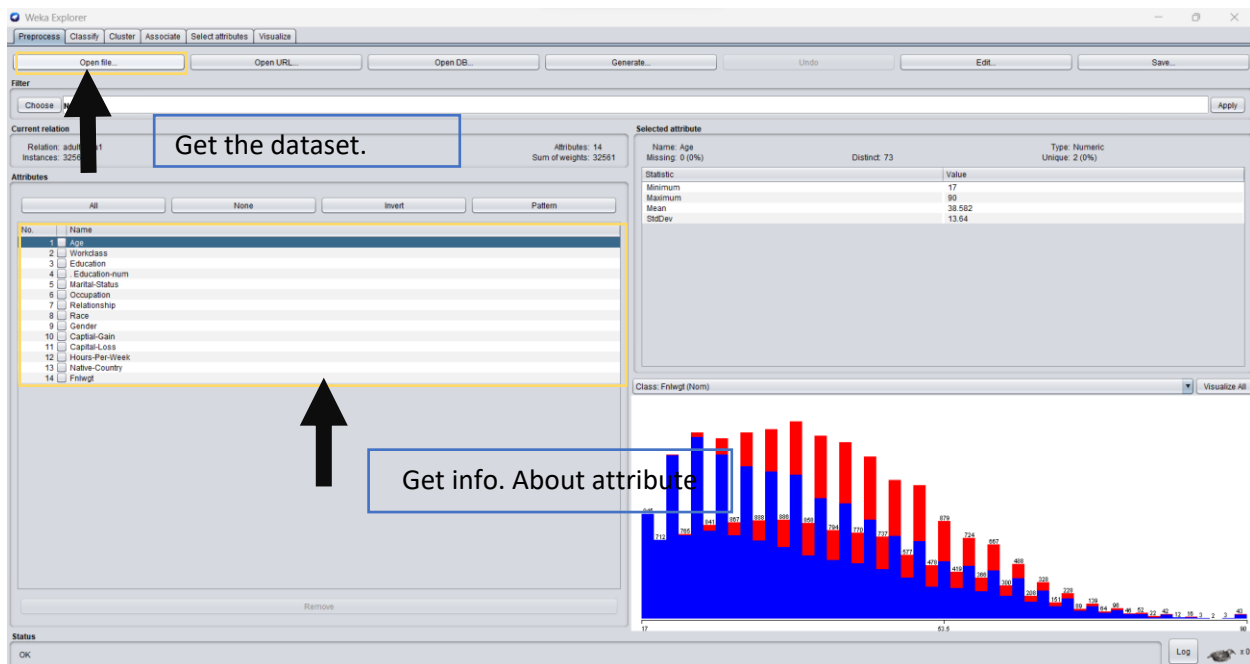


Figure 1

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.

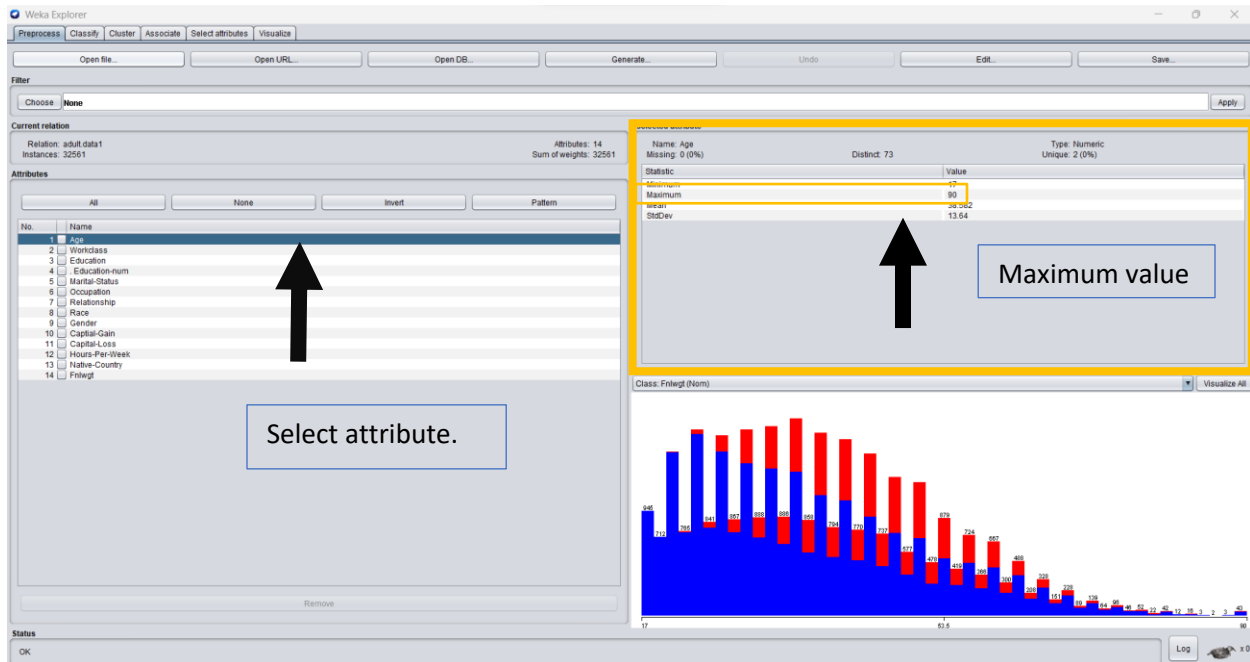


Figure 2

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.

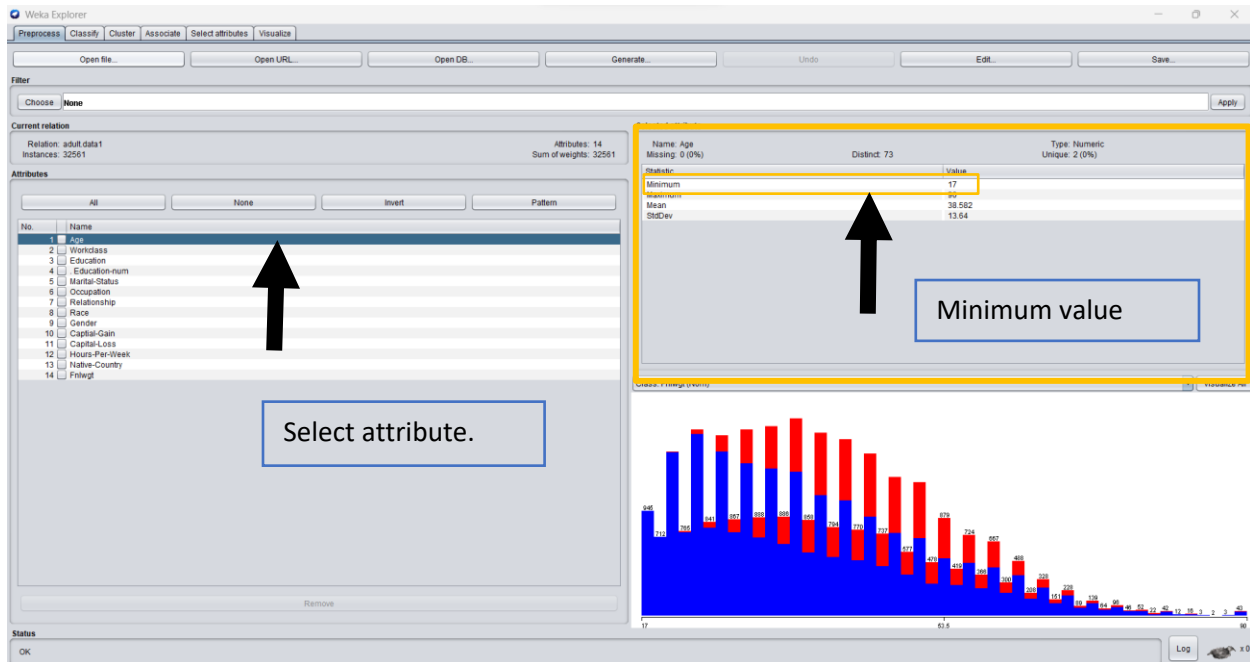


Figure 3

4. Range (non nominal)

4.1. range = max-min

5. Range(nominal)

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

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

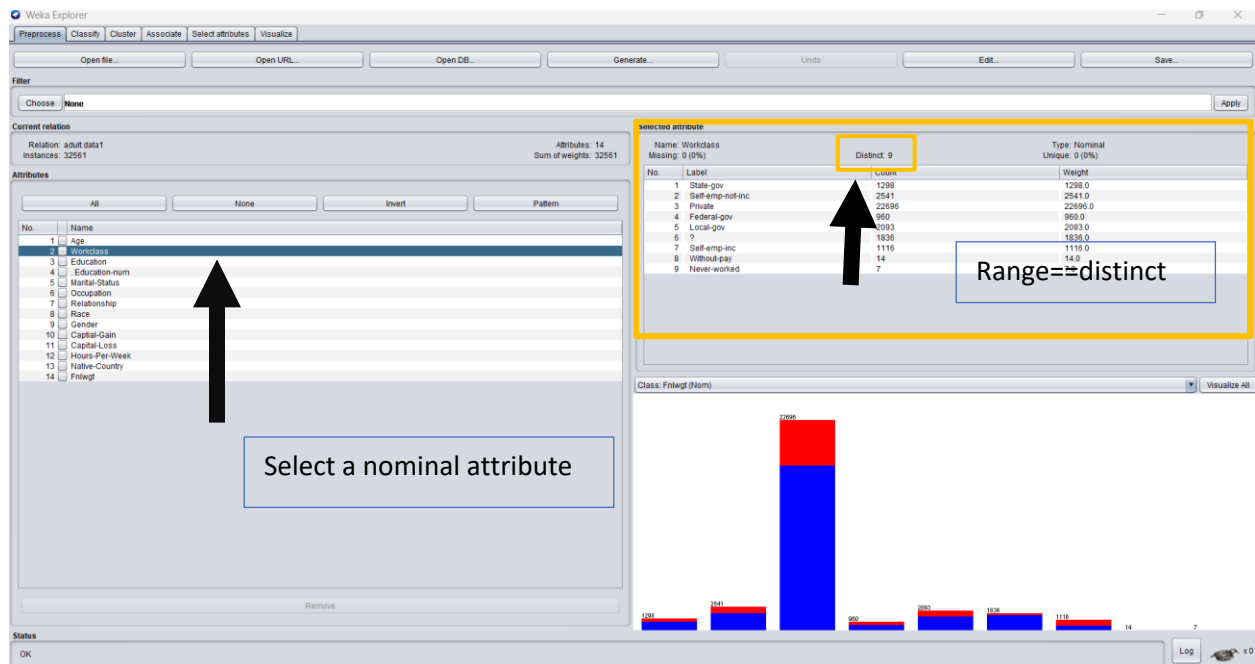


Figure4

6. Variance

6.1. open the dataset

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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Age	Workclass	Education	Education-Num	Marital-Status	Occupation	Relationship	Race	Gender	Capital-Gain	Capital-Loss	Hours-Per-Week	Native-Country	Foreign-Born		
1	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K		
2	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K		VARIANCE=
3	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K		186.0614
4	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K		=VAR.S(A:A)
5	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K		
6	37	Private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K		
7	49	Private	9th	5	Married-civ-spouse	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K		
8	52	Self-emp-inc	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K		
9	31	Private	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K		
10	42	Private	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	5178	0	40	United-States	>50K		
11	37	Private	Some-college	10	Married-civ-spouse	Adm-clerical	Black	Male	0	0	0	80	United-States	<=50K		
12	30	State-gov	Bachelors	13	Never-married	Adm-clerical	Asian-Pac-Islander	Male	0	0	0	40	India	<=50K		
13	23	Private	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K		
14	32	Private	Assoc-acad	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K		
15	40	Private	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K		
16	34	Private	7th-8th	4	Married-civ-spouse	Transportation	Husband	Amer-Indian-Alaska-Native	Male	0	0	45	Mexico	<=50K		
17	25	Self-emp-inc	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K		
18	32	Private	HS-grad	9	Never-married	Machine-operating	Unmarried	White	Male	0	0	40	United-States	<=50K		
19	38	Private	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K		
20	43	Self-emp-inc	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K		
21	40	Private	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K		
22	54	Private	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K		
23	35	Federal-gov	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K		
24	43	Private	11th	7	Married-civ-spouse	Transportation	Husband	White	Male	0	2042	40	United-States	<=50K		
25	59	Private	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K		

Figure5

7. Mean

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

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

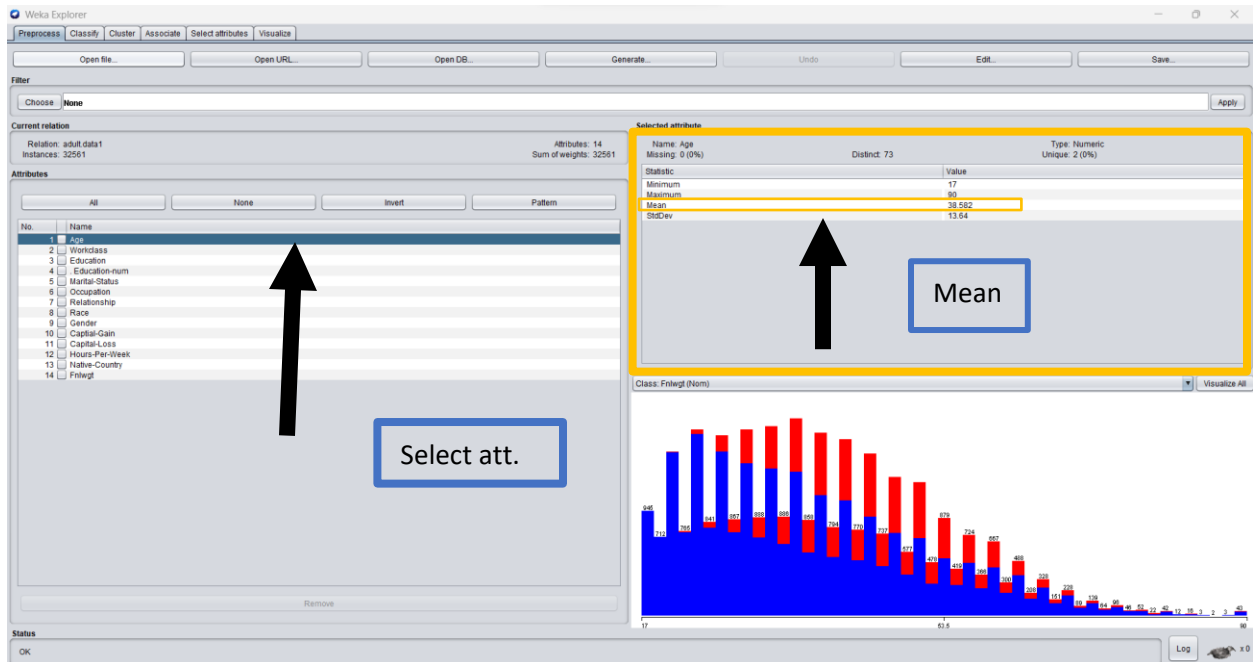


Figure6

8. Mood

8.1. open the dataset

8.2. select column and in any empty cell type MODE

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Age	Workclass	Education	Education	Marital-St	Occupatio	Relationsh	Race	Gender	Capital-Ga	Capital-Lo	Hours-Per	Native-Co	Fnlwgt		
39	State-gov	Bachelors	13	Never-ma	Adm-cler	Not-in-far	White	Male	2174	0	40	United-St	<=50K		
50	Self-emp	Bachelors	13	Married-c	Exec-man	Husband	White	Male	0	0	13	United-St	<=50K		
38	Private	HS-grad		Not-in-far	White	Male	0	0	0	0	40	United-St	<=50K		
53	Private	11th		Husband	Black	Male	0	0	0	0	40	United-St	<=50K		
28	Private	Bachelors	13	Married-c	Prof-speci	Wife	Black	Female	0	0	40	Cuba	<=50K		
37	Private	Masters	14	Married-c	Exec-man	Wife	White	Female	0	0	40	United-St	<=50K		
49	Private	9th	5	Married-s	Other-ser	Not-in-far	Black	Female	0	0	16	Jamaica	<=50K		
52	Self-emp	HS-grad	9	Married-c	Exec-man	Husband	White	Male	0	0	45	United-St	>50K		
31	Private	Masters	14	Never-ma	Prof-speci	Not-in-far	White	Female	14084	0	50	United-St	>50K		
42	Private	Bachelors	13	Married-c	Exec-man	Husband	White	Male	5178	0	40	United-St	<=50K		
37	Private	Some-coll	10	Married-c	Exec-man	Husband	Black	Male	0	0	80	United-St	<=50K		
30	State-gov	Bachelors	13	Married-c	Prof-speci	Husband	Asian-Pac	Male	0	0	40	India	>50K		
23	Private	Bachelors	13	Never-ma	Adm-cler	Own-child	White	Female	0	0	30	United-St	<=50K		
32	Private	Assoc-acd	12	Never-ma	Sales	Not-in-far	Black	Male	0	0	50	United-St	<=50K		
40	Private	Assoc-voc	11	Married-c	Craft-rep	Husband	Asian-Pac	Male	0	0	40	?	>50K		
34	Private	7th-8th	4	Married-c	Transport	Husband	Amer-Indi	Male	0	0	45	Mexico	<=50K		
25	Self-emp	HS-grad	9	Never-ma	Farming-f	Own-child	White	Male	0	0	35	United-St	<=50K		
32	Private	HS-grad	9	Never-ma	Machine-o	Unmarrie	White	Male	0	0	40	United-St	<=50K		
38	Private	11th	7	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K		
43	Self-emp	Masters	14	Divorced	Exec-man	Unmarrie	White	Female	0	0	45	United-St	<=50K		

← Select column

Mood
36
=MODE(A:A)

→ Calculate mood

Figure7

9. Median

9.1. open the dataset

9.2. select column and in any empty cell type MEDIAN

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Age	Workclass	Education	Education	Marital-Sta	Occupatio	Relationsh	Race	Gender	Capitla-Ga	Capital-Lo	Hours-Per	Native-Co	Fnlwgt		
39	State-gov	Bachelors	13	Never-ma	Adm-clerik	Not-in-far	White	Male	2174	0	40	United-Sti	<=50K		
50	Self-emp	Bachelors	13	Married-c	Exec-man	Husband	White	Male	0	0	13	United-Sti	<=50K		
38	Private	HS-grad	9	Divorced	Handlers-i	Not-in-far	White	Male	0	0	40	United-Sti	<=50K		
53	Private	11th	7	Married-c	Handlers-i	Husband	Black	Male	0	0	40	United-Sti	<=50K		
28	Private	Bachelors	13	Married-c	Prof-speci	Wife	Black	Female	0	0	40	Cuba	<=50K		
37	Private	Masters	14	Married-c	Exec-man	Wife	White	Female	0	0	40	United-Sti	<=50K		
49	Private	9th	5	Married-s	Other-ser	Not-in-far	Black	Female	0	0	16	Jamaica	<=50K		
52	Self-emp	HS-grad	9	Married-c	Exec-man	Husband	White	Male	0	0	45	United-Sti	>50K		
31	Private	Bachelors	14	Never-ma	Prof-speci	Not-in-far	White	Female	14084	0	50	United-Sti	>50K		
42	Private	Bachelors	13	Married-c	Exec-man	Husband	White	Male	5178	0	40	United-Sti	>50K		
37	Private	Some-col				Husband	Black	Male	0	0	80	United-Sti	>50K		
30	State-gov	Bachelors				Husband	Asian-Pac	Male	0	0	40	India	>50K		
23	Private	Bachelors				Own-child	White	Female	0	0	30	United-Sti	<=50K		
32	Private	Assoc-acc				Not-in-far	Black	Male	0	0	50	United-Sti	<=50K		
40	Private	Assoc-voc	11	Married-c	Craft-repa	Husband	Asian-Pac	Male	0	0	40	?	>50K		
34	Private	7th-8th	4	Married-c	Transport	Husband	Amer-Indi	Male	0	0	45	Mexico	<=50K		
25	Self-emp	HS-grad	9	Never-ma	Farming-fi	Own-child	White	Male	0	0	35	United-Sti	<=50K		
32	Private	HS-grad	9	Never-ma	Machine-o	Unmarrie	White	Male	0	0	40	United-Sti	<=50K		
38	Private	11th	7	Married-c	Sales	Husband	White	Male	0	0	50	United-Sti	<=50K		
43	Self-emp	Masters	14	Divorced	Exec-man	Unmarrie	White	Female	0	0	45	United-Sti	>50K		
40	Private	Doctorate	16	Married-c	Prof-speci	Husband	White	Male	0	0	60	United-Sti	>50K		
54	Private	HS-grad	9	Separated	Other-ser	Unmarrie	Black	Female	0	0	20	United-Sti	<=50K		

Median
=MEDIAN(A:A)

Select column

Calculate 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

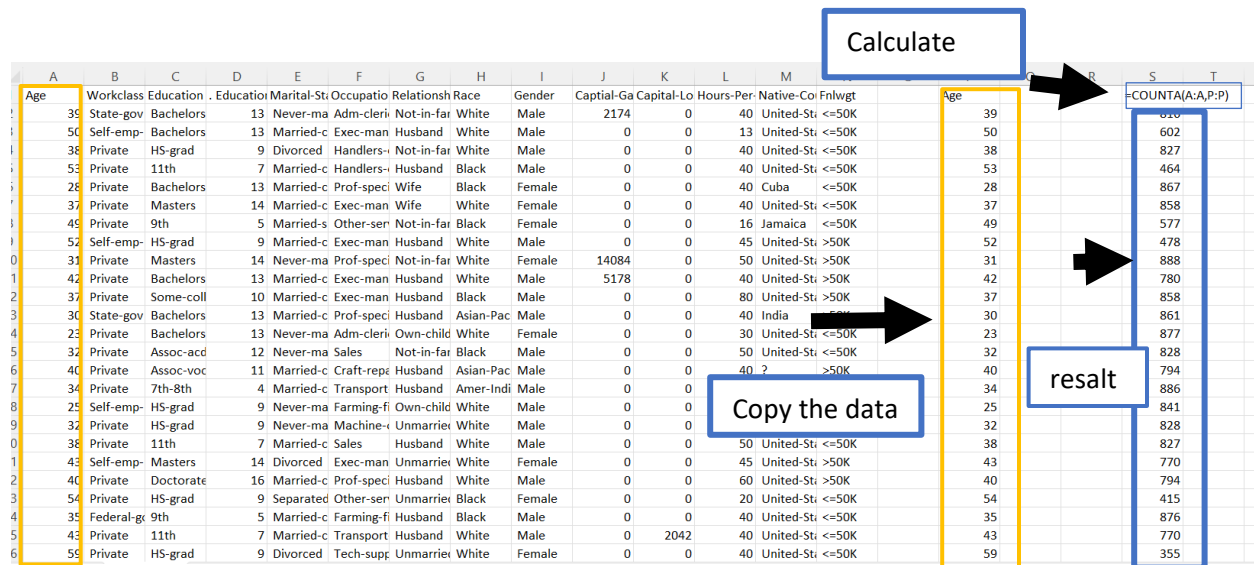


Figure9

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	#	#
Enlwgt	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.

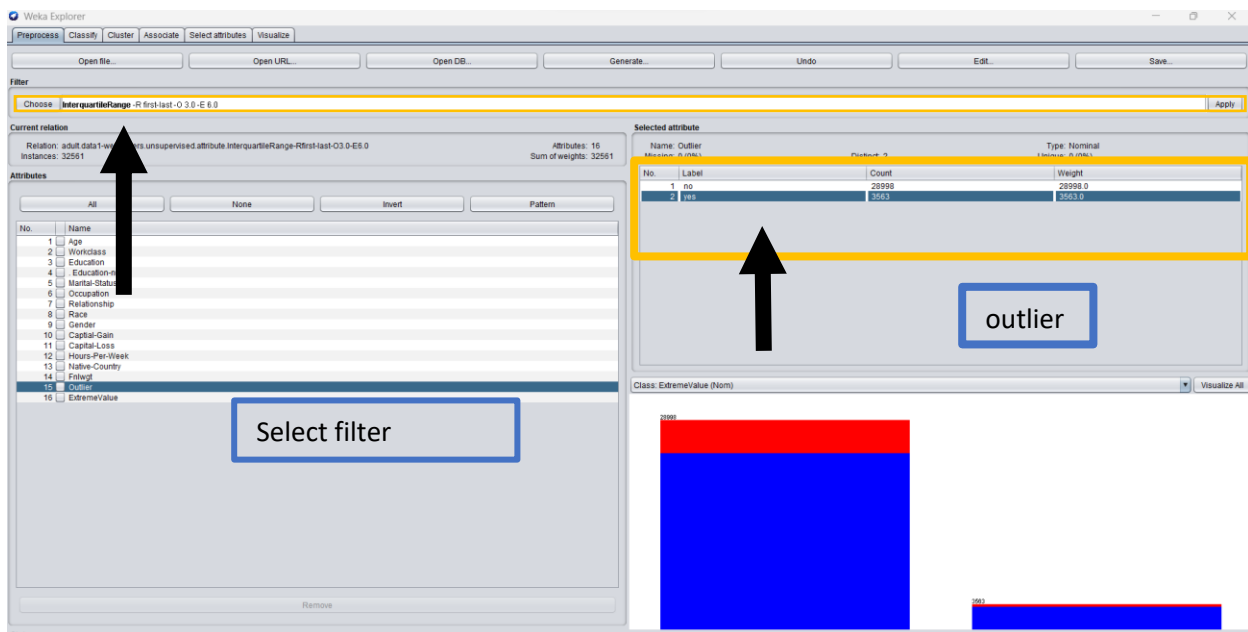


Figure10

2. Cluster

2.1. from cluster tab

2.2. choose EM class

2.3. hit start

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Choose' button is set to 'EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100'. The 'Cluster mode' section has 'Use training set' selected. The 'Start' button is highlighted with a yellow box and a black arrow pointing to it from a blue box labeled 'Start algorithm.'. The 'Result list' shows '15:52 EM'. The 'Clusterer output' section displays a table of cluster results for various countries, followed by summary statistics and a 'Clustered Instances' table. The 'Clustered Instances' table is highlighted with a yellow box, and a black arrow points to it from a blue box labeled 'Cluster value'.

Clusterer output

Cambodia	6.9333	14.0667
Thailand	15.0533	4.9467
Ecuador	17.1842	12.8158
Laos	12.0802	7.9198
Taiwan	28.0144	24.9856
Haiti	36.8673	9.1327
Portugal	20.1243	18.8757
Dominican-Republic	54.9225	17.0775
El-Salvador	77.2727	30.7273
France	16.013	14.987
Guatemala	49.8496	16.1504
China	36.2678	40.7322
Japan	31.1812	32.8188
Yugoslavia	6.0283	11.9717
Peru	24.2381	8.7619
Outlying-US (Guam-USVI-etc)	13.1126	2.8874
Scotland	9.0138	4.9862
Trinidad&Tobago	12.1713	8.8287
Greece	11.0101	19.9899
Nicaragua	21.2586	14.7414
Vietnam	44.8089	24.1911
Hong	12.0395	9.9605
Ireland	17.235	8.765
Hungary	8.0082	6.9918
Holand-Netherlands	1	2
[total]	17985.6928	14659.3072
Fnlgwt		
<=50K	16673.3695	8048.6305
>50K	1272.3233	6570.6767
[total]	17945.6928	14619.3072

Time taken to build model (full training data) : 73.4 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	28422	(87%)
1	4139	(13%)

Log likelihood: -27.45328

Figure11

3. Scatter Plots

3.1. from visualize tab

3.2. you can see the relation between attributes.

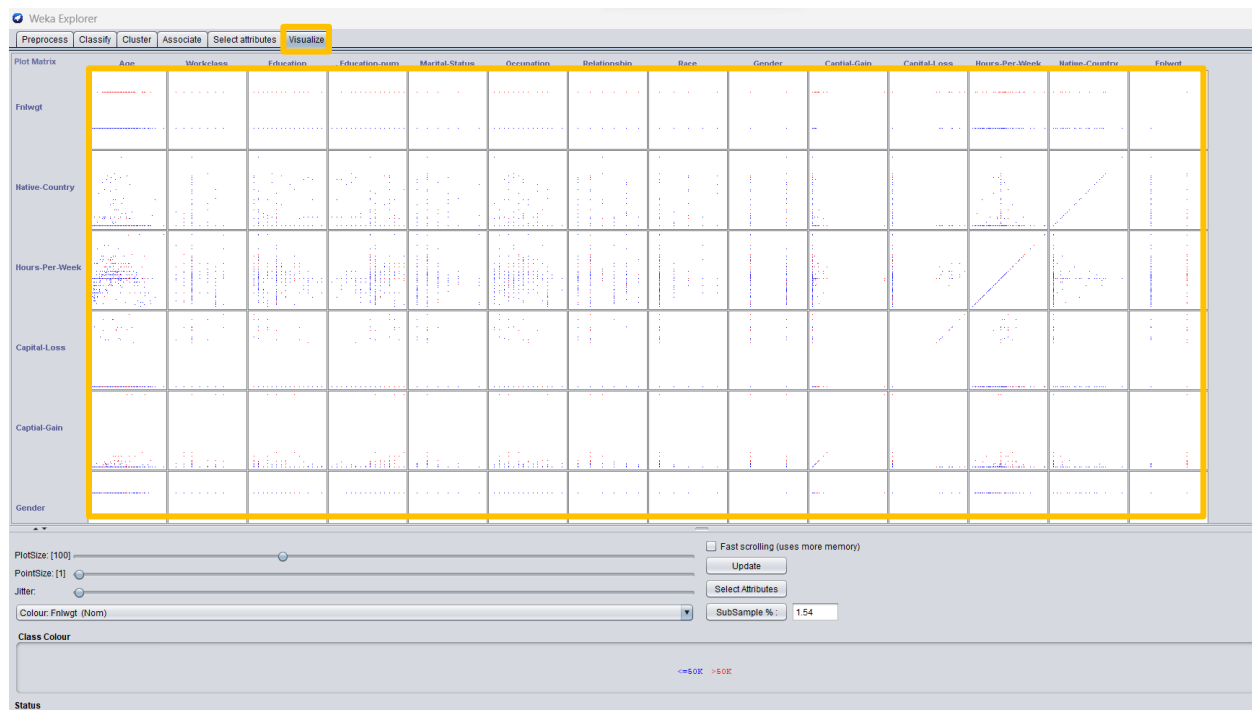


Figure12