Exploratory Analysis over Census Data

Report by:
Vijetha Shenoy Badiadka - 1001822855
Kiran Venkatesh Kulkarni - 1001848434

Introduction:

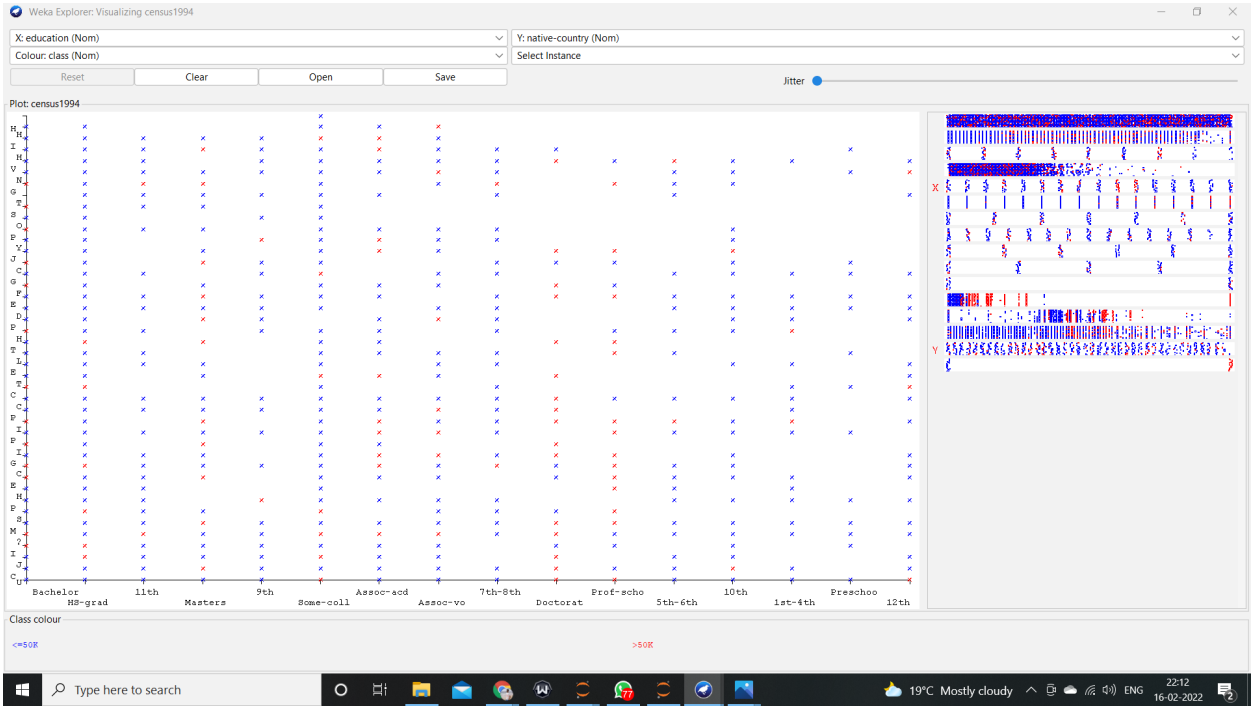
Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from the code (E.g: Java code). Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization.

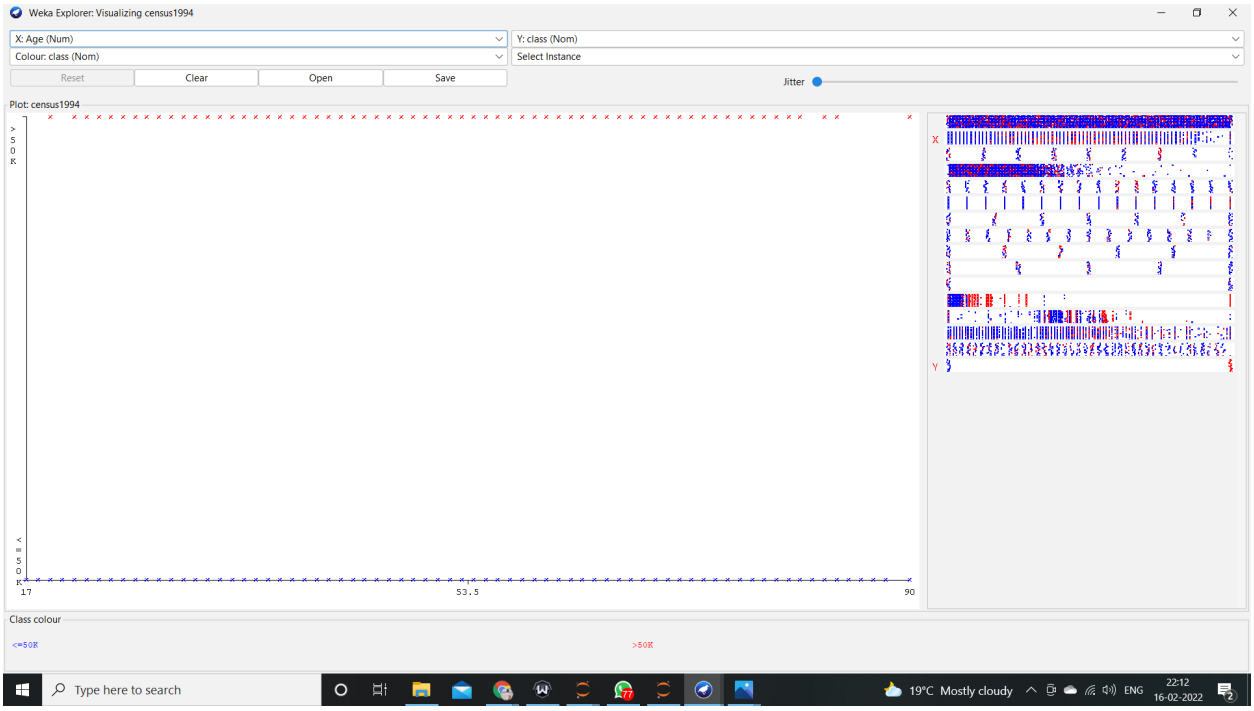
Retrieving the data

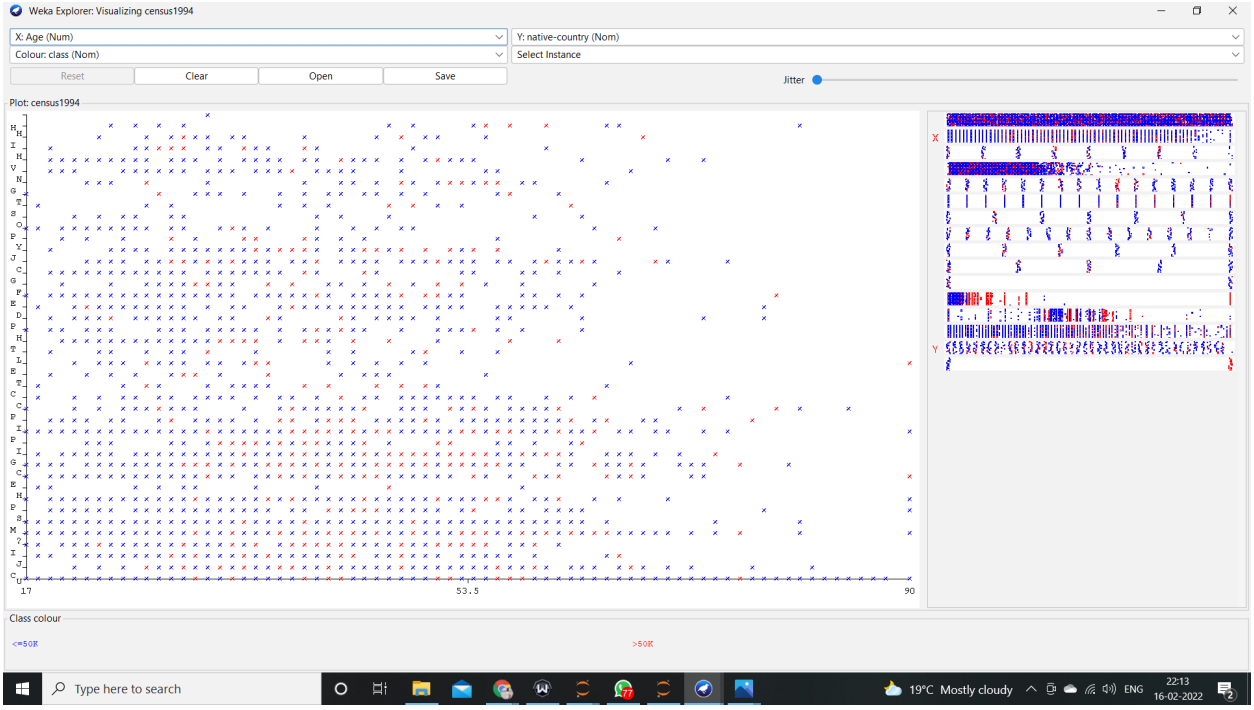
We load the dataset into Weka and then model it.

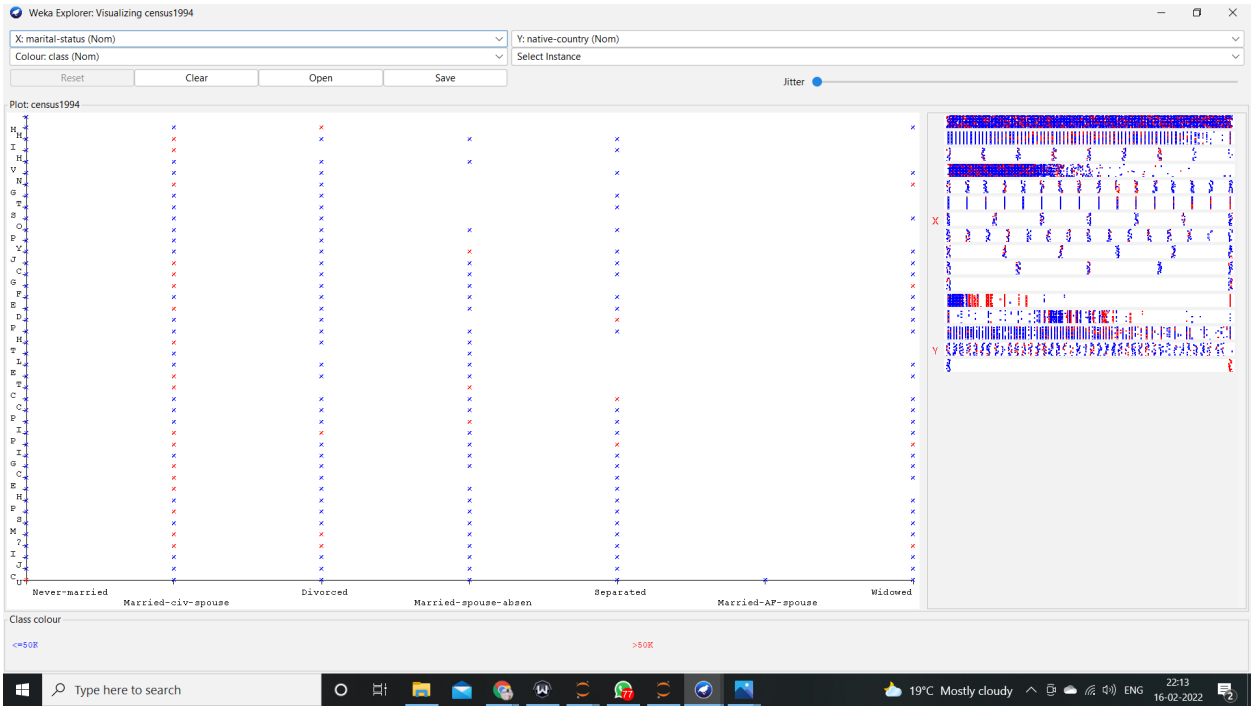
Visualization using Weka:

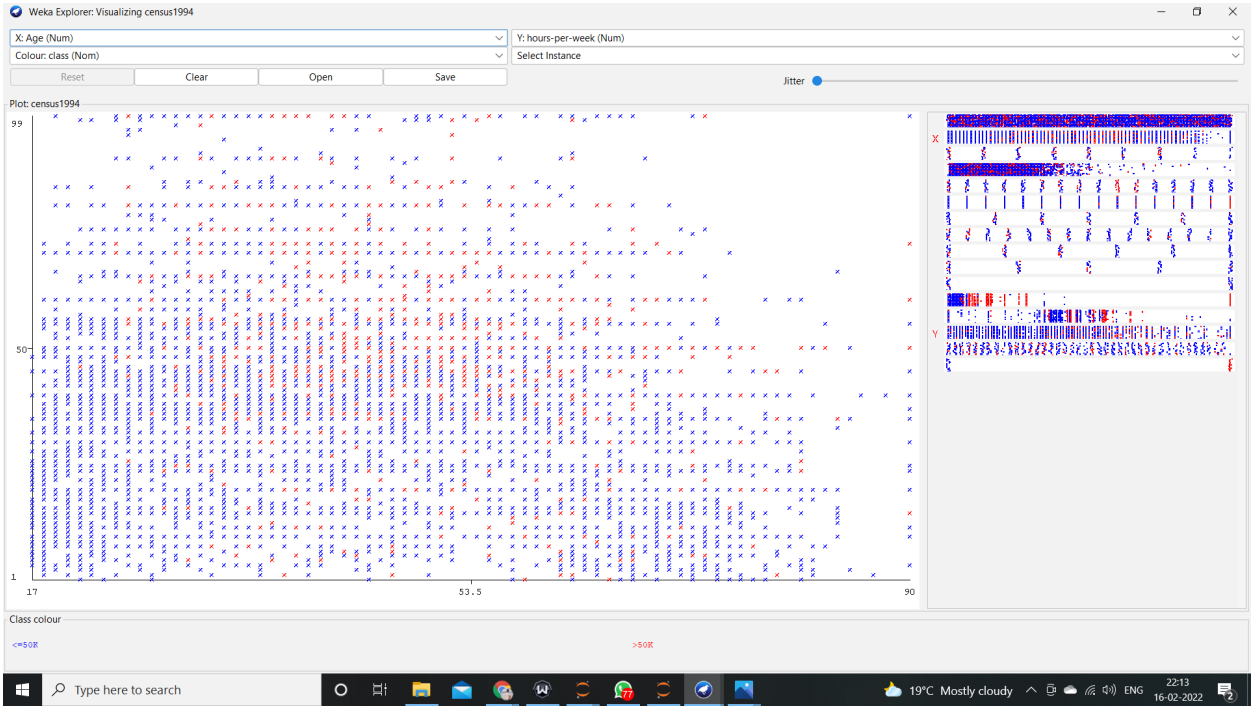
Snapshots:

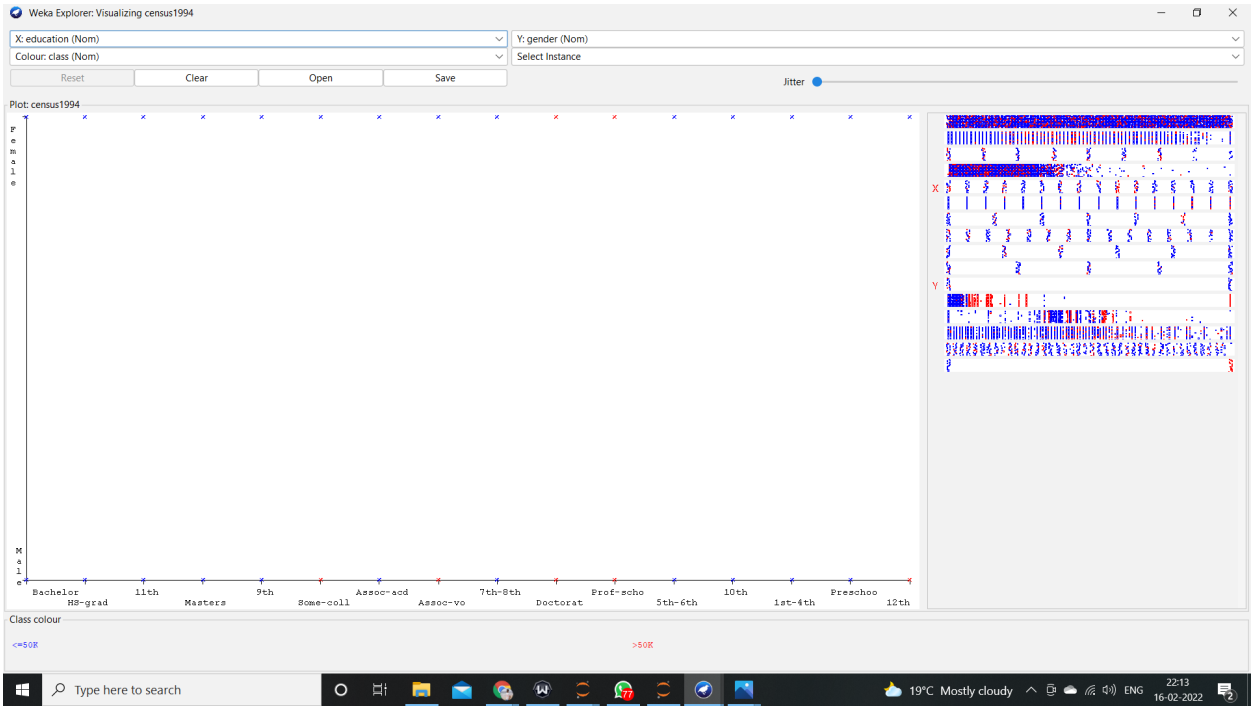


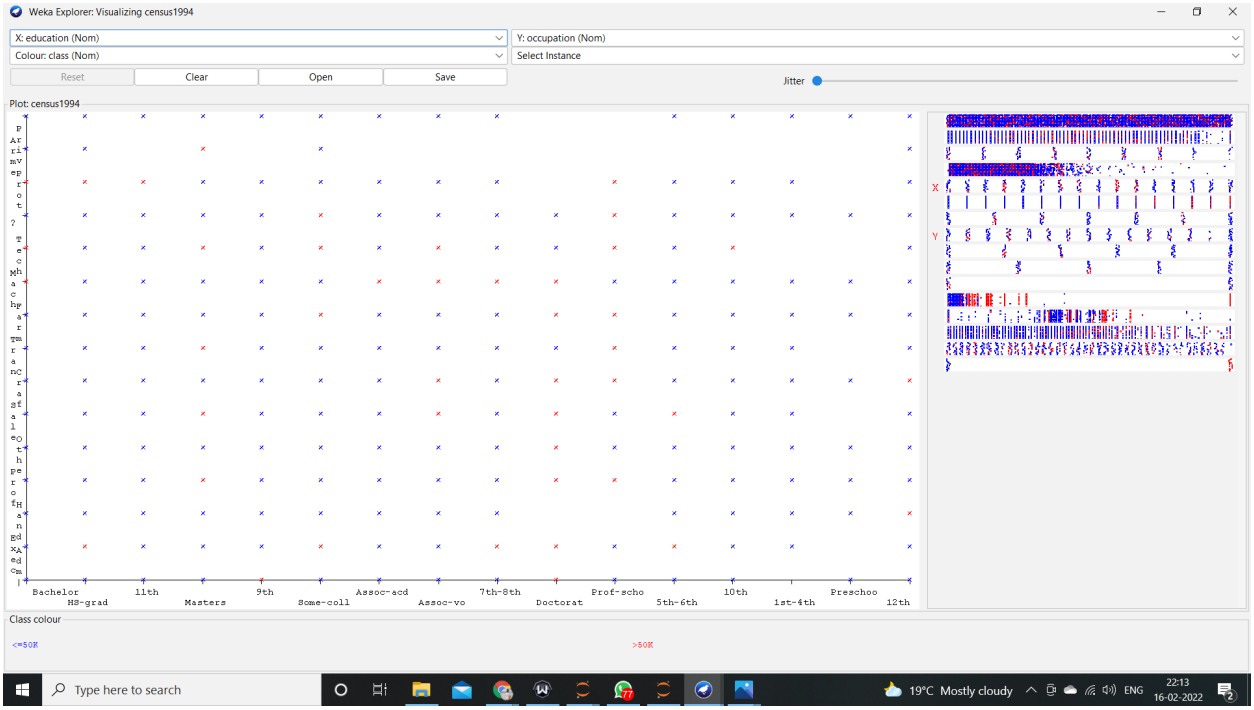

Weka Explorer: Visualizing census1994
X: Age (Num)
Y: class (Nom)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: census1994
>50K
<=50K
17
53.5
90
X
Y
Class colour
<=50K
>50K
Type here to search
19°C Mostly cloudy
ENG
22:12
16-02-2022

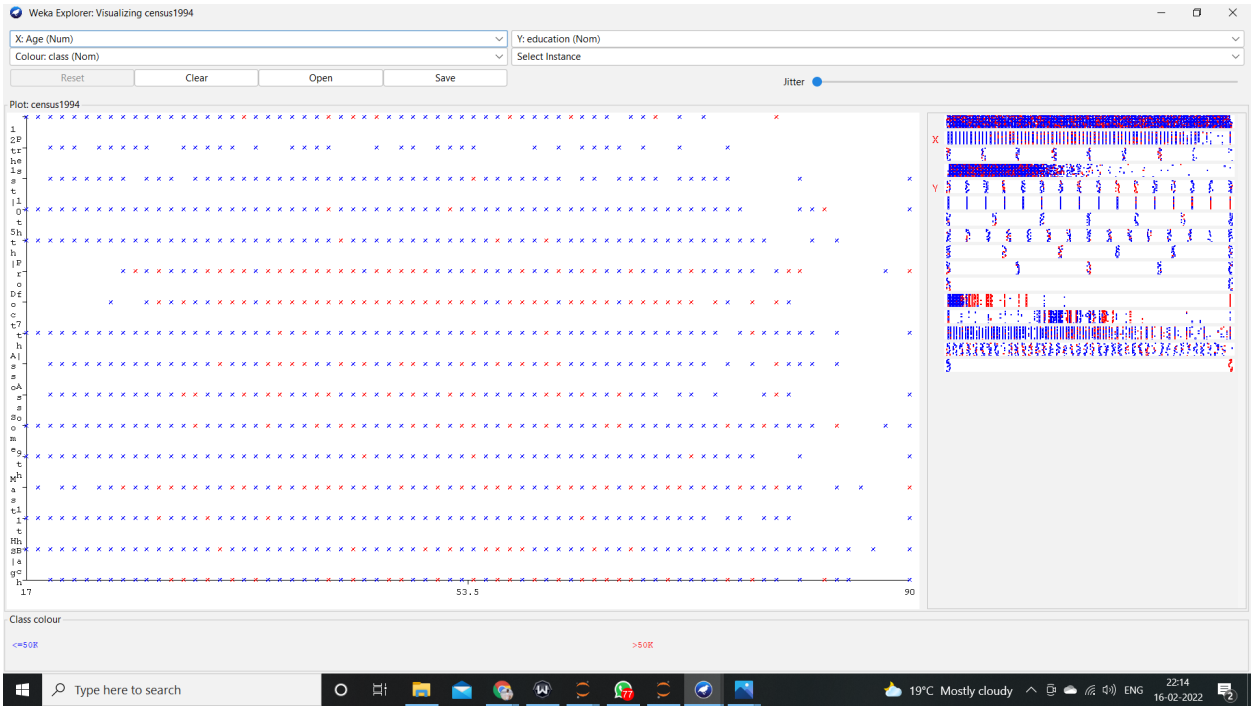

Weka Explorer: Visualizing census1994
X: Age (Num)
Y: native-country (Nom)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: census1994
17
53.5
90
X
Y
Class colour
<=50K
>50K
Type here to search
19°C Mostly cloudy
ENG
22:13
16-02-2022


Weka Explorer: Visualizing census1994
X: marital-status (Nom)
Y: native-country (Nom)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: census1994
Never-married
Married-civ-spouse
Divorced
Married-spouse-absen
Separated
Married-AF-spouse
Widowed
Class colour
<=50K
>50K
Type here to search
19°C Mostly cloudy
ENG
22:13
16-02-2022


Weka Explorer: Visualizing census1994
X: Age (Num)
Y: hours-per-week (Num)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: census1994
99
50
1
17
53.5
90
Class colour
<=50K
>50K
Type here to search
19°C Mostly cloudy
ENG
22:13
16-02-2022


Weka Explorer: Visualizing census1994
X: education (Nom)
Y: gender (Nom)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: census1994
Female
Male
Bachelor
HS-grad
11th
Masters
9th
Some-coll
Assoc-acd
Assoc-vo
7th-8th
Doctorat
Prof-scho
5th-6th
10th
1st-4th
Preschoo
12th
X
Y
Class colour
<=50K
>50K
Type here to search
19°C Mostly cloudy
ENG
22:13
16-02-2022

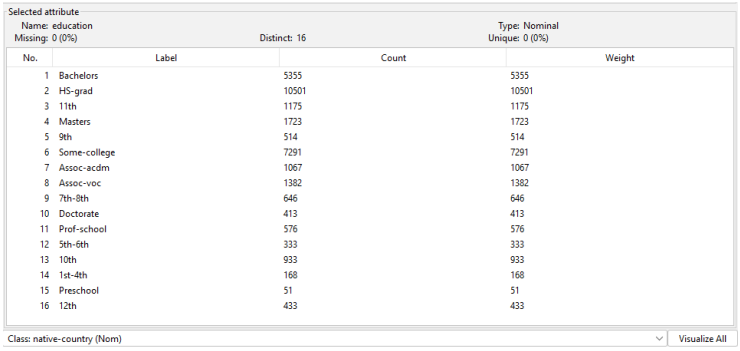

Weka Explorer: Visualizing census1994
X: education (Nom)
Y: occupation (Nom)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: census1994
Bachelor
HS-grad
11th
Masters
9th
Some-coll
Assoc-acd
Assoc-vo
7th-8th
Doctorat
Prof-scho
5th-6th
10th
1st-4th
Preschoo
12th
X
Y
Class colour
<=50K
>50K
Type here to search
19°C Mostly cloudy
ENG
22:13
16-02-2022



References:

https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/

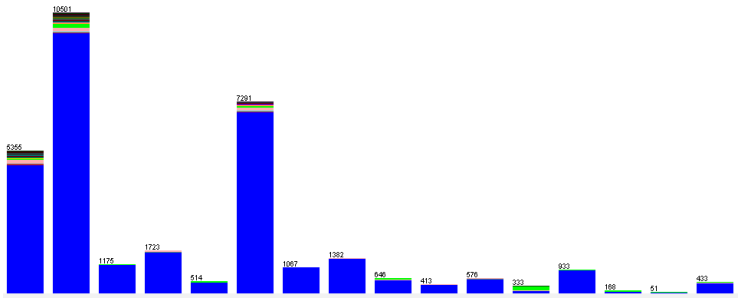
Task 1 a. Capital Gain vs Education Level:



Selected attribute
Name: education
Missing: 0 (0%)
Distinct: 16
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Bachelors	5355	5355
2	HS-grad	10501	10501
3	11th	1175	1175
4	Masters	1723	1723
5	9th	514	514
6	Some-college	7291	7291
7	Assoc-acdm	1067	1067
8	Assoc-voc	1382	1382
9	7th-8th	646	646
10	Doctorate	413	413
11	Prof-school	576	576
12	5th-6th	333	333
13	10th	933	933
14	1st-4th	168	168
15	Preschool	51	51
16	12th	433	433

Class: native-country (Nom)
Visualize All



As seen in from the above graph, capital gain vs education level is plotted. Native country is used as a class to identify people from various countries.

- Task 2-e: Find out top 5 native-countries besides United-States, print their names and number of surveys belonging to each.
-

Selected attribute

Name: native-country
Missing: 0 (0%)
Distinct: 42
Type: Nominal
Unique: 1 (0%)

No.	Label	Count	Weight
1	United-States	29170	29170
2	Cuba	95	95
3	Jamaica	81	81
4	India	100	100
5	?	583	583
6	Mexico	643	643
7	South	80	80
8	Puerto-Rico	114	114
9	Honduras	13	13
10	England	90	90
11	Canada	121	121
12	Germany	137	137
13	Iran	43	43
14	Philippines	198	198
15	Italy	73	73
16	Poland	60	60
17	Columbia	59	59

Class: Age (Num) Visualize All

As seen from the above graph and the table, the top 5 countries besides the US are **Cuba, Jamaica, India and Mexico**.

- Find out Top-5 native-countries with the most number of samples belonging to class >50K

Selected attribute

Name: native-country Type: Nominal
Missing: 0 (0%) Distinct: 42 Unique: 1 (0%)

No.	Label	Count	Weight
1	United-States	29170	29170
2	Cuba	95	95
3	Jamaica	81	81
4	India	100	100
5	?	583	583
6	Mexico	643	643
7	South	80	80
8	Puerto-Rico	114	114
9	Honduras	13	13
10	England	90	90
11	Canada	121	121
12	Germany	137	137
13	Iran	43	43
14	Philippines	198	198
15	Italy	73	73
16	Poland	60	60
17	Columbia	59	59

Class: Age (Num) Visualize All

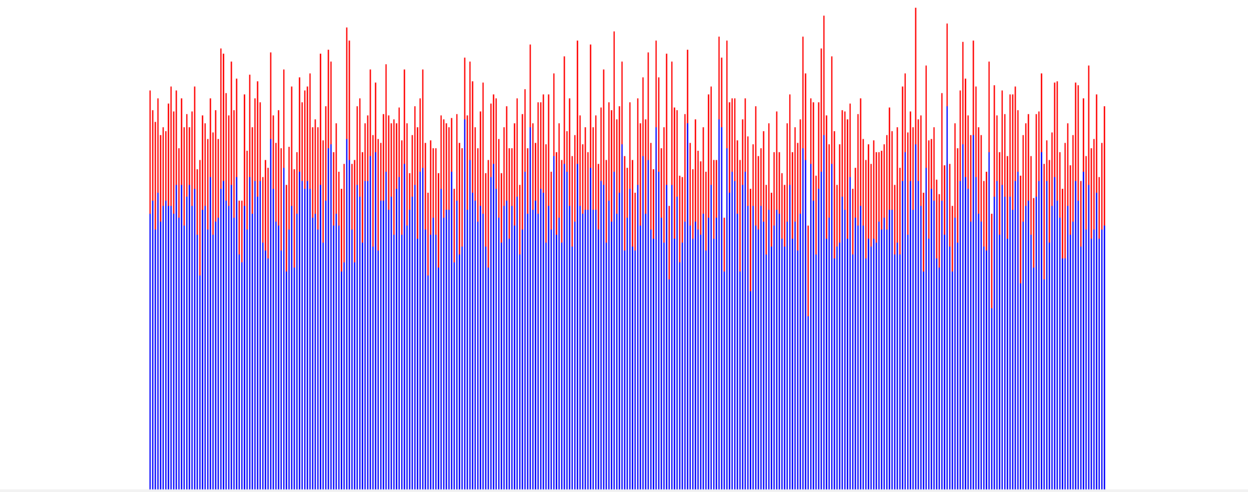
- Draw a vertical bar chart for total number of surveys taken for each gender for each month. Display months with their corresponding names.

Selected attribute

Name: ï»¿Date Type: Nominal
Missing: 0 (0%) Distinct: 365 Unique: 0 (0%)

No.	Label	Count	Weight
1	3/20/1994	97	97
2	1/14/1994	92	92
3	8/14/1994	89	89
4	3/17/1994	95	95
5	9/20/1994	86	86
6	11/28/1994	88	88
7	3/2/1994	87	87
8	11/27/1994	94	94
9	12/25/1994	98	98
10	10/10/1994	92	92
11	10/29/1994	97	97
12	11/10/1994	83	83
13	1/4/1994	95	95
14	2/15/1994	88	88
15	8/1/1994	91	91
16	9/1/1994	88	88
17	5/15/1994	92	92

Class: class (Nom) Visualize All



Red color in the graph denotes the male gender whereas the blue represents females. The plot shows per day.

- Draw a "vertical" bar chart that lists the top-5 native-countries based on the number of samples with class >50K.

Selected attribute

Name: native-country Type: Nominal
Missing: 0 (0%) Distinct: 42 Unique: 1 (0%)

No.	Label	Count	Weight
1	United-States	29170	29170
2	Cuba	95	95
3	Jamaica	81	81
4	India	100	100
5	?	583	583
6	Mexico	643	643
7	South	80	80
8	Puerto-Rico	114	114
9	Honduras	13	13
10	England	90	90
11	Canada	121	121
12	Germany	137	137
13	Iran	43	43
14	Philippines	198	198
15	Italy	73	73
16	Poland	60	60
17	Columbia	59	59

Class: native-country (Nom) Visualize All

- Draw a scatter plot for age vs hours per week.

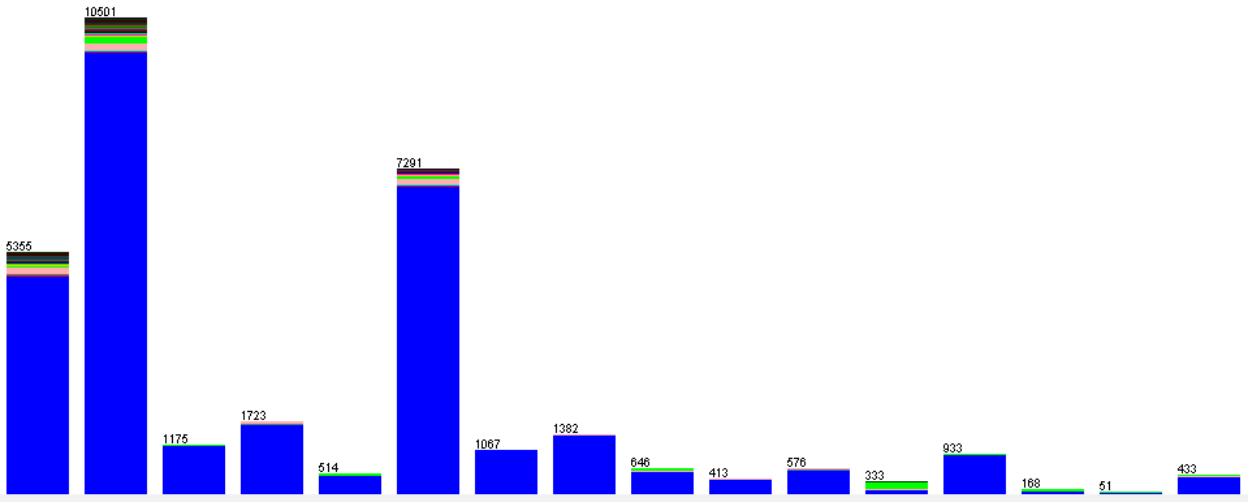
- Draw a line chart showing average capital gain for each education category.

Selected attribute

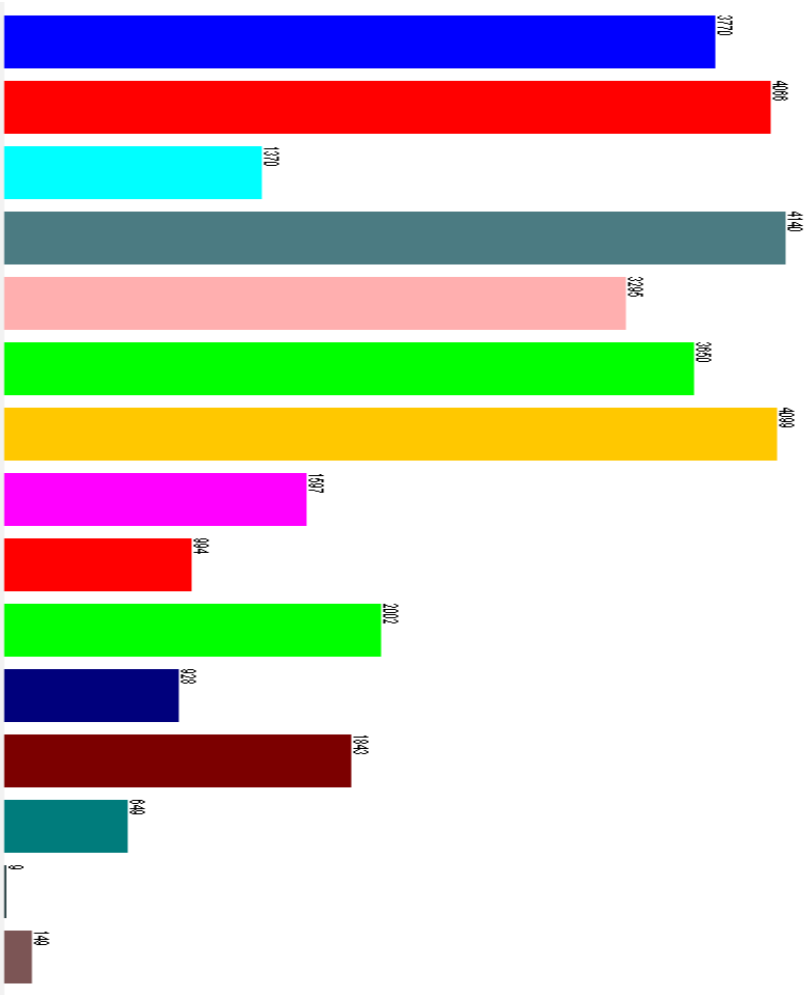
Name: education Type: Nominal
Missing: 0 (0%) Distinct: 16 Unique: 0 (0%)

No.	Label	Count	Weight
1	Bachelors	5355	5355
2	HS-grad	10501	10501
3	11th	1175	1175
4	Masters	1723	1723
5	9th	514	514
6	Some-college	7291	7291
7	Assoc-acdm	1067	1067
8	Assoc-voc	1382	1382
9	7th-8th	646	646
10	Doctorate	413	413
11	Prof-school	576	576
12	5th-6th	333	333
13	10th	933	933
14	1st-4th	168	168
15	Preschool	51	51
16	12th	433	433

Class: native-country (Nom) Visualize All



* Draw a 'horizontal' bar chart for the top-5 most common occupation.



- Draw a 'horizontal' bar chart for the top-5 most common workclass.