DATA MINING (CSE 5334-004)

Professor : Dr Elizabeth Diaz

ASSIGNMENT - 1

Team : 5

Rachana Ramireddy :	1002028071
Darshana Madalani :	1002033083
Venkata Harika Bandi :	1002032077

Report for Assignment - 1 using WEKA

Introduction:

WEKA is open-source software that can be used to focus on certain data mining applications.

We can utilize WEKA tool to visualize our dataset in a useful context.

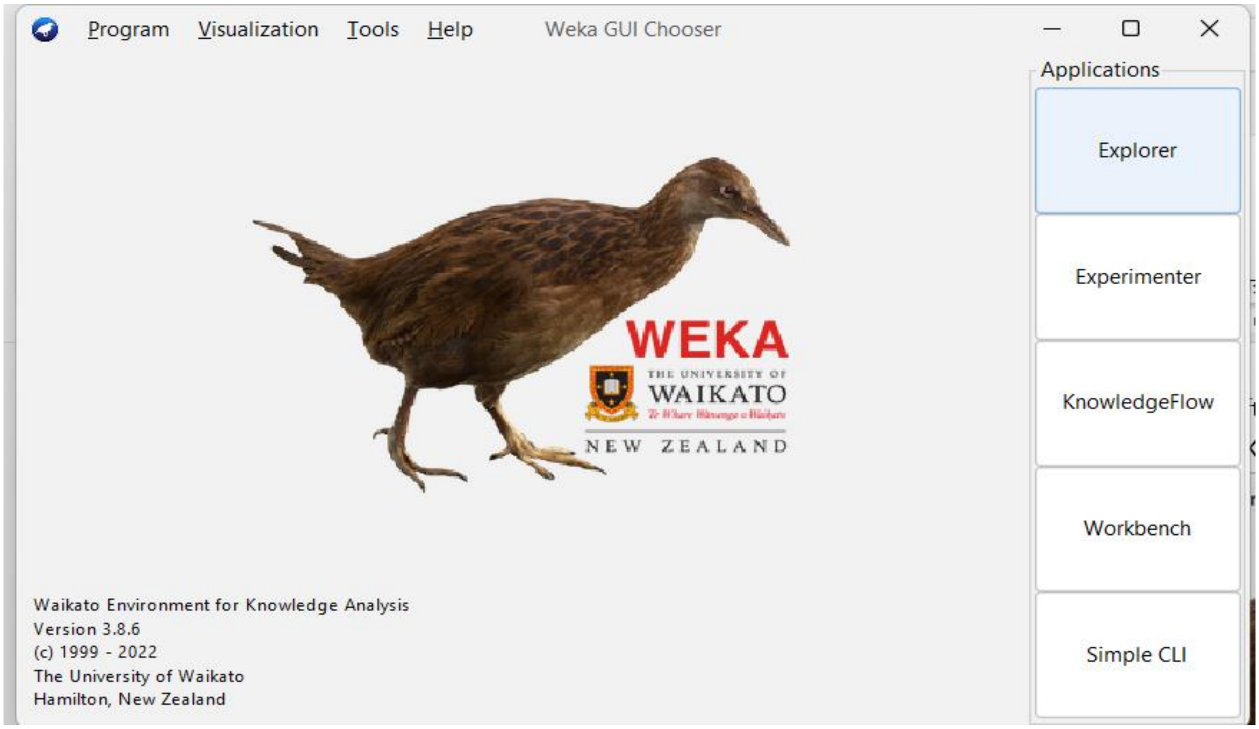
Dataset:

The dataset contains 25 attributes, mentioned as below:-

bmi
Age
asa_status
baseline_cancer
baseline_charlson
baseline_cvd
baseline_dementia
baseline_diabetes
baseline_digestive
baseline_osteoart
baseline_psych
baseline_pulmonary
ahrq_ccs
ccsComplicationRate
ccsMort30Rate
complication_rsi
dow
gender hour
month
moonphase
mort30
mortality_rsi
race
complication

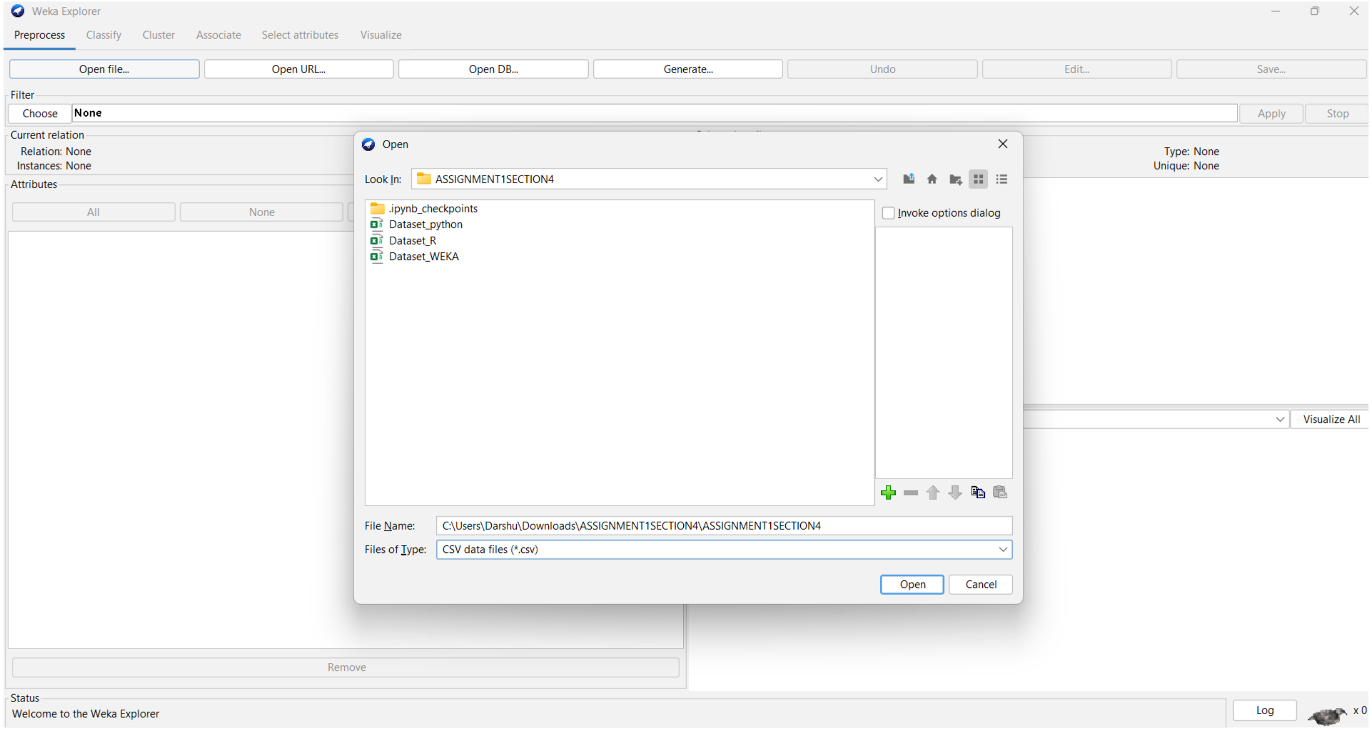
Retrieving the Data:

- First we need to install WEKA software from https://sourceforge.net/projects/weka/
- Below is the home screen of WEKA tool :



Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

- For getting/loading csv file in WEKA click on Explorer -> Open File, Choose your .csv file from your device.



Filter: None

Relation: None
Instances: None

Type: None
Unique: None

Look In: ASSIGNMENT1SECTION4

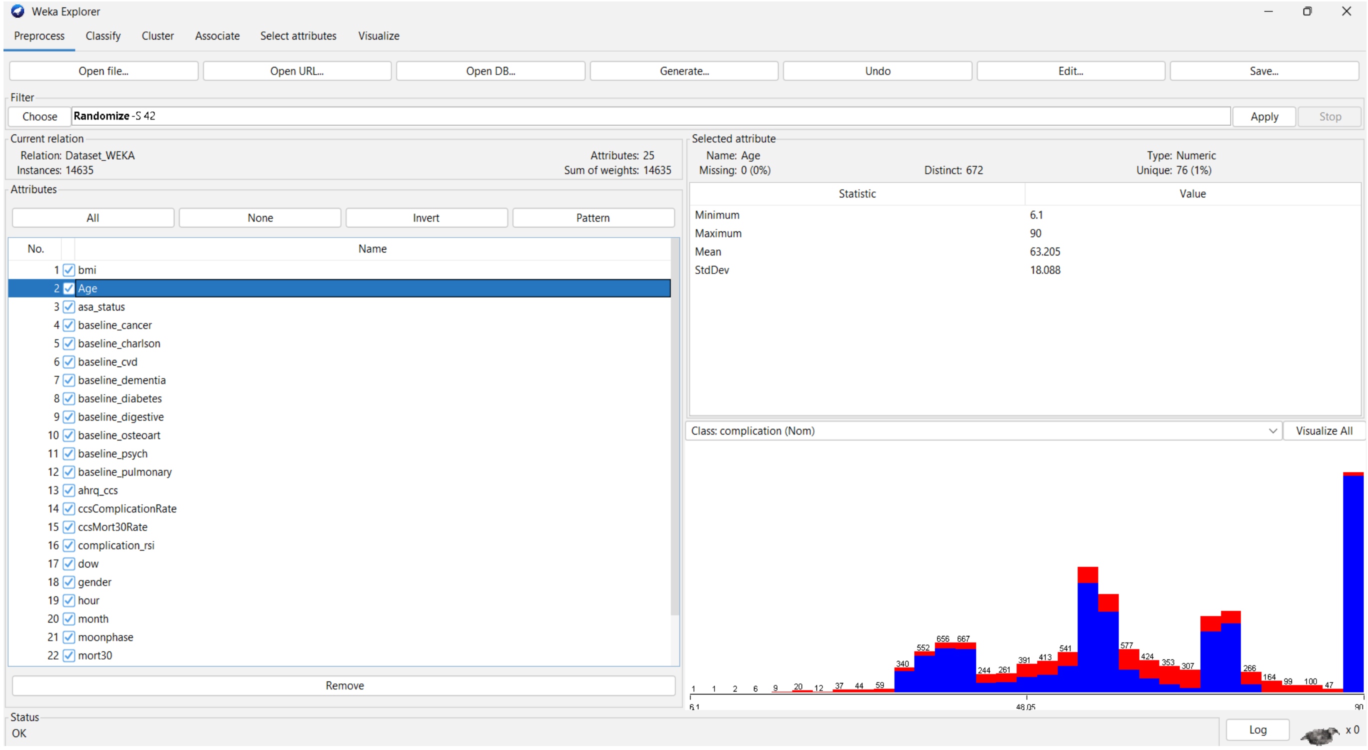
.ipynb_checkpoints
Dataset_python
Dataset_R
Dataset_WEKA

□ Invoke options dialog

File Name: C:\Users\Darshu\Downloads\ASSIGNMENT1SECTION4\ASSIGNMENT1SECTION4
Files of Type: CSV data files (*.csv)

Welcome to the Weka Explorer

- All of the attributes will be displayed in the "Attributes" window, such as below image:



Filter

Randomize -S 42

Current relation

Relation: Dataset_WEKA
Instances: 14635
Attributes: 25
Sum of weights: 14635

Attributes

No.	Name
1	☑ bmi
2	☑ Age
3	☑ asa_status
4	☑ baseline_cancer
5	☑ baseline_charlson
6	☑ baseline_cvd
7	☑ baseline_dementia
8	☑ baseline_diabetes
9	☑ baseline_digestive
10	☑ baseline_osteoart
11	☑ baseline_psych
12	☑ baseline_pulmonary
13	☑ ahrq_ccs
14	☑ ccsComplicationRate
15	☑ ccsMort30Rate
16	☑ complication_rsi
17	☑ dow
18	☑ gender
19	☑ hour
20	☑ month
21	☑ moonphase
22	☑ mort30

Selected attribute

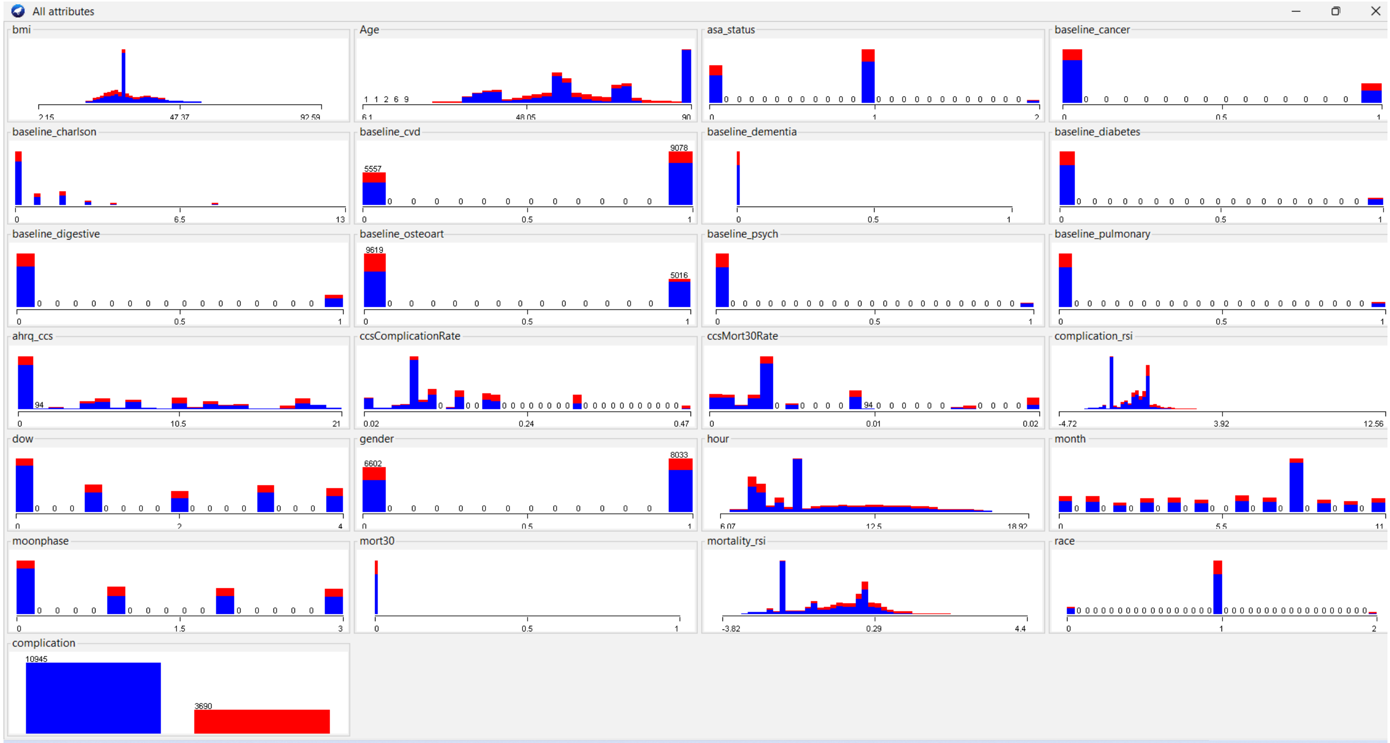
Name: Age
Missing: 0 (0%)
Distinct: 672
Type: Numeric
Unique: 76 (1%)

Statistic	Value
Minimum	6.1
Maximum	90
Mean	63.205
StdDev	18.088

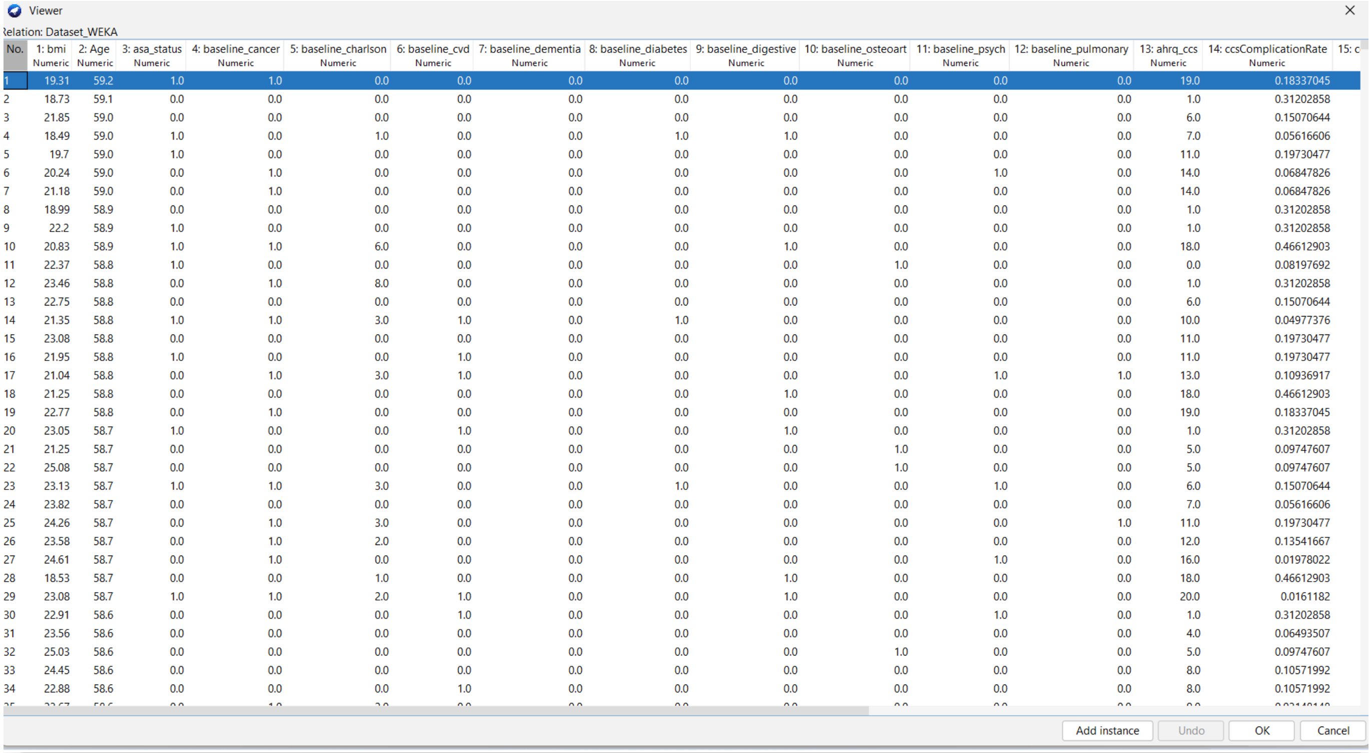
Class: complication (Nom)

Status
OK

- By clicking on "Visualize All" button, we can get the different combined version of data in small graphs as below:

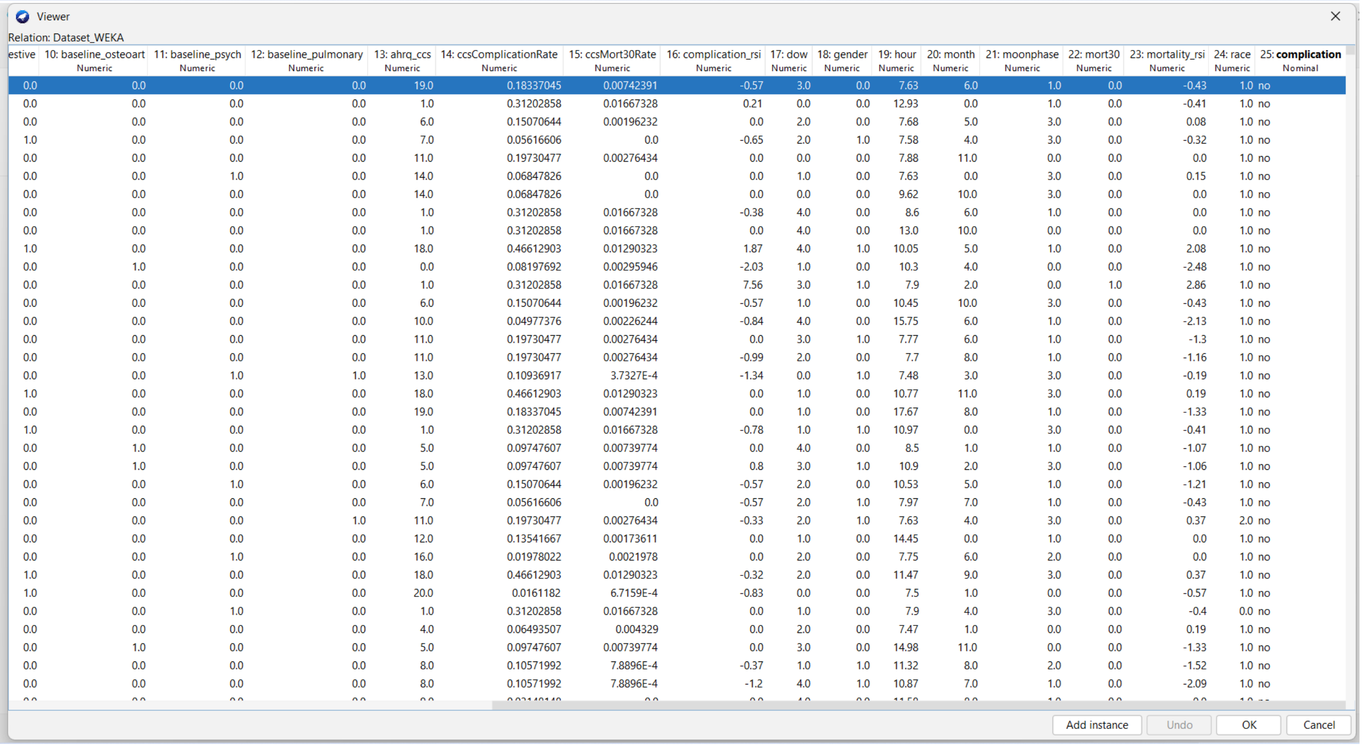


Glimpse of the data:



Relation: Dataset_WEKA

No.	1: bmi	2: Age	3: asa_status	4: baseline_cancer	5: baseline_charlson	6: baseline_cvd	7: baseline_dementia	8: baseline_diabetes	9: baseline_digestive	10: baseline_osteoart	11: baseline_psych	12: baseline_pulmonary	13: ahrq_ccs	14: ccsComplicationRate	15: c
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	
1	19.31	59.2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.0	0.18337045	
2	18.73	59.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.31202858	
3	21.85	59.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.15070644	
4	18.49	59.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	7.0	0.05616606	
5	19.7	59.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.19730477	
6	20.24	59.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	14.0	0.06847826	
7	21.18	59.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	0.06847826	
8	18.99	58.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.31202858	
9	22.2	58.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.31202858	
10	20.83	58.9	1.0	1.0	6.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	18.0	0.46612903	
11	22.37	58.8	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.08197692	
12	23.46	58.8	0.0	1.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.31202858	
13	22.75	58.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.15070644	
14	21.35	58.8	1.0	1.0	3.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	10.0	0.04977376	
15	23.08	58.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.19730477	
16	21.95	58.8	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.19730477	
17	21.04	58.8	0.0	1.0	3.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	13.0	0.10936917	
18	21.25	58.8	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	18.0	0.46612903	
19	22.77	58.8	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.0	0.18337045	
20	23.05	58.7	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.31202858	
21	21.25	58.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	5.0	0.09747607	
22	25.08	58.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	5.0	0.09747607	
23	23.13	58.7	1.0	1.0	3.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	6.0	0.15070644	
24	23.82	58.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	0.05616606	
25	24.26	58.7	0.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	11.0	0.19730477	
26	23.58	58.7	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.0	0.13541667	
27	24.61	58.7	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	16.0	0.01978022	
28	18.53	58.7	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	18.0	0.46612903	
29	23.08	58.7	1.0	1.0	2.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	20.0	0.0161182	
30	22.91	58.6	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.31202858	
31	23.56	58.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.06493507	
32	25.03	58.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	5.0	0.09747607	
33	24.45	58.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0	0.10571992	
34	22.88	58.6	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0	0.10571992	

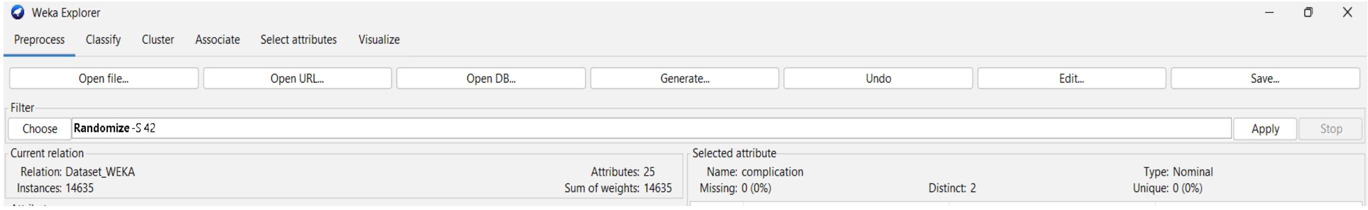


Relation: Dataset_WEKA

estive	10: baseline_osteoart	11: baseline_psych	12: baseline_pulmonary	13: ahrq_ccs	14: ccsComplicationRate	15: ccsMort30Rate	16: complication_rsi	17: dow	18: gender	19: hour	20: month	21: moonphase	22: mort30	23: mortality_rsi	24: race	25: **complication**
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
0.0	0.0	0.0	0.0	19.0	0.18337045	0.00742391	-0.57	3.0	0.0	7.63	6.0	1.0	0.0	-0.43	1.0	no
0.0	0.0	0.0	0.0	1.0	0.31202858	0.01667328	0.21	0.0	0.0	12.93	0.0	1.0	0.0	-0.41	1.0	no
0.0	0.0	0.0	0.0	6.0	0.15070644	0.00196232	0.0	2.0	0.0	7.68	5.0	3.0	0.0	0.08	1.0	no
1.0	0.0	0.0	0.0	7.0	0.05616606	0.0	-0.65	2.0	1.0	7.58	4.0	3.0	0.0	-0.32	1.0	no
0.0	0.0	0.0	0.0	11.0	0.19730477	0.00276434	0.0	0.0	0.0	7.88	11.0	0.0	0.0	0.0	1.0	no
0.0	0.0	1.0	0.0	14.0	0.06847826	0.0	0.0	1.0	0.0	7.63	0.0	3.0	0.0	0.15	1.0	no
0.0	0.0	0.0	0.0	14.0	0.06847826	0.0	0.0	0.0	0.0	9.62	10.0	3.0	0.0	0.0	1.0	no
0.0	0.0	0.0	0.0	1.0	0.31202858	0.01667328	-0.38	4.0	0.0	8.6	6.0	1.0	0.0	0.0	1.0	no
0.0	0.0	0.0	0.0	1.0	0.31202858	0.01667328	0.0	4.0	0.0	13.0	10.0	0.0	0.0	0.0	1.0	no
1.0	0.0	0.0	0.0	18.0	0.46612903	0.01290323	1.87	4.0	1.0	10.05	5.0	1.0	0.0	2.08	1.0	no
0.0	1.0	0.0	0.0	0.0	0.08197692	0.00295946	-2.03	1.0	0.0	10.3	4.0	0.0	0.0	-2.48	1.0	no
0.0	0.0	0.0	0.0	1.0	0.31202858	0.01667328	7.56	3.0	1.0	7.9	2.0	0.0	1.0	2.86	1.0	no
0.0	0.0	0.0	0.0	6.0	0.15070644	0.00196232	-0.57	1.0	0.0	10.45	10.0	3.0	0.0	-0.43	1.0	no
0.0	0.0	0.0	0.0	10.0	0.04977376	0.00226244	-0.84	4.0	0.0	15.75	6.0	1.0	0.0	-2.13	1.0	no
0.0	0.0	0.0	0.0	11.0	0.19730477	0.00276434	0.0	3.0	1.0	7.77	6.0	1.0	0.0	-1.3	1.0	no
0.0	0.0	0.0	0.0	11.0	0.19730477	0.00276434	-0.99	2.0	0.0	7.7	8.0	1.0	0.0	-1.16	1.0	no
0.0	0.0	1.0	1.0	13.0	0.10936917	3.7327E-4	-1.34	0.0	1.0	7.48	3.0	3.0	0.0	-0.19	1.0	no
1.0	0.0	0.0	0.0	18.0	0.46612903	0.01290323	0.0	1.0	0.0	10.77	11.0	3.0	0.0	0.19	1.0	no
0.0	0.0	0.0	0.0	19.0	0.18337045	0.00742391	0.0	1.0	0.0	17.67	8.0	1.0	0.0	-1.33	1.0	no
1.0	0.0	0.0	0.0	1.0	0.31202858	0.01667328	-0.78	1.0	1.0	10.97	0.0	3.0	0.0	-0.41	1.0	no
0.0	1.0	0.0	0.0	5.0	0.09747607	0.00739774	0.0	4.0	0.0	8.5	1.0	1.0	0.0	-1.07	1.0	no
0.0	1.0	0.0	0.0	5.0	0.09747607	0.00739774	0.8	3.0	1.0	10.9	2.0	3.0	0.0	-1.06	1.0	no
0.0	0.0	1.0	0.0	6.0	0.15070644	0.00196232	-0.57	2.0	0.0	10.53	5.0	1.0	0.0	-1.21	1.0	no
0.0	0.0	0.0	0.0	7.0	0.05616606	0.0	-0.57	2.0	1.0	7.97	7.0	1.0	0.0	-0.43	1.0	no
0.0	0.0	0.0	1.0	11.0	0.19730477	0.00276434	-0.33	2.0	1.0	7.63	4.0	3.0	0.0	0.37	2.0	no
0.0	0.0	0.0	0.0	12.0	0.13541667	0.00173611	0.0	1.0	0.0	14.45	0.0	1.0	0.0	0.0	1.0	no
0.0	0.0	1.0	0.0	16.0	0.01978022	0.0021978	0.0	2.0	0.0	7.75	6.0	2.0	0.0	0.0	1.0	no
1.0	0.0	0.0	0.0	18.0	0.46612903	0.01290323	-0.32	2.0	0.0	11.47	9.0	3.0	0.0	0.37	1.0	no
1.0	0.0	0.0	0.0	20.0	0.0161182	6.7159E-4	-0.83	0.0	0.0	7.5	1.0	0.0	0.0	-0.57	1.0	no
0.0	0.0	1.0	0.0	1.0	0.31202858	0.01667328	0.0	1.0	0.0	7.9	4.0	3.0	0.0	-0.4	0.0	no
0.0	0.0	0.0	0.0	4.0	0.06493507	0.004329	0.0	2.0	0.0	7.47	1.0	0.0	0.0	0.19	1.0	no
0.0	1.0	0.0	0.0	5.0	0.09747607	0.00739774	0.0	3.0	0.0	14.98	11.0	0.0	0.0	-1.33	1.0	no
0.0	0.0	0.0	0.0	8.0	0.10571992	7.8896E-4	-0.37	1.0	1.0	11.32	8.0	2.0	0.0	-1.52	1.0	no
0.0	0.0	0.0	0.0	8.0	0.10571992	7.8896E-4	-1.2	4.0	1.0	10.87	7.0	1.0	0.0	-2.09	1.0	no

Check for missing data:

Basically missing data is defined as the items that are missing from an instance. Outcomes will be significantly impacted by missing data. After importing .csv file in WEKA tool we can see that there are no missing values from below image for Attribute “Complication” :

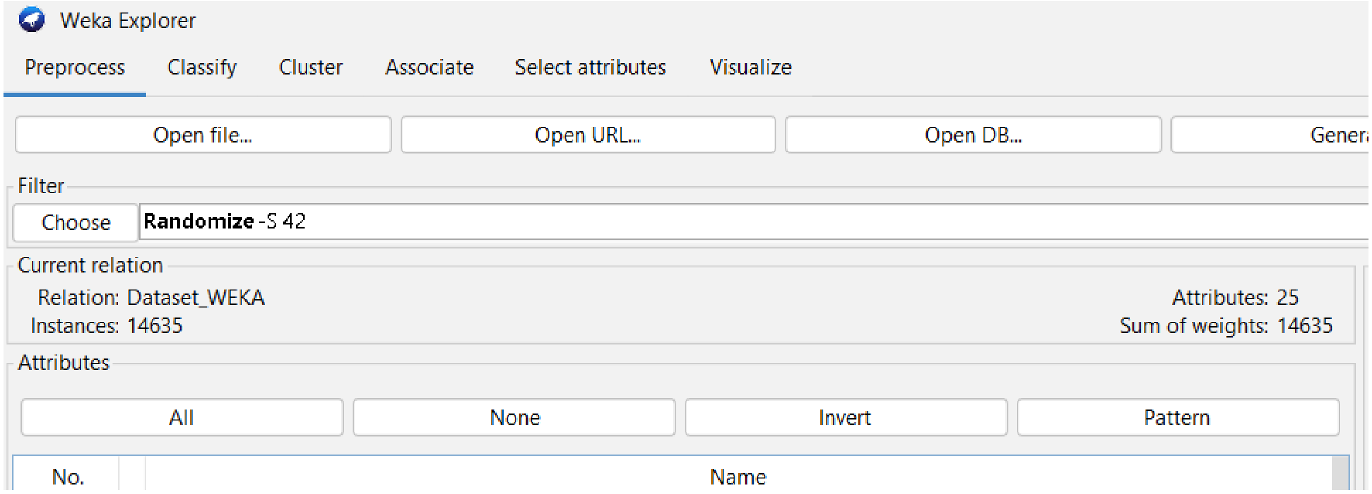


Filter: Randomize -S 42

Current relation
Relation: Dataset_WEKA
Instances: 14635
Attributes: 25
Sum of weights: 14635

Selected attribute
Name: complication
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

- **Details about Rows and Columns of Dataset:**
  - Number of Rows/Instances shown in below image : 14635
  - Number of Columns/Attributes shown in below image : 25



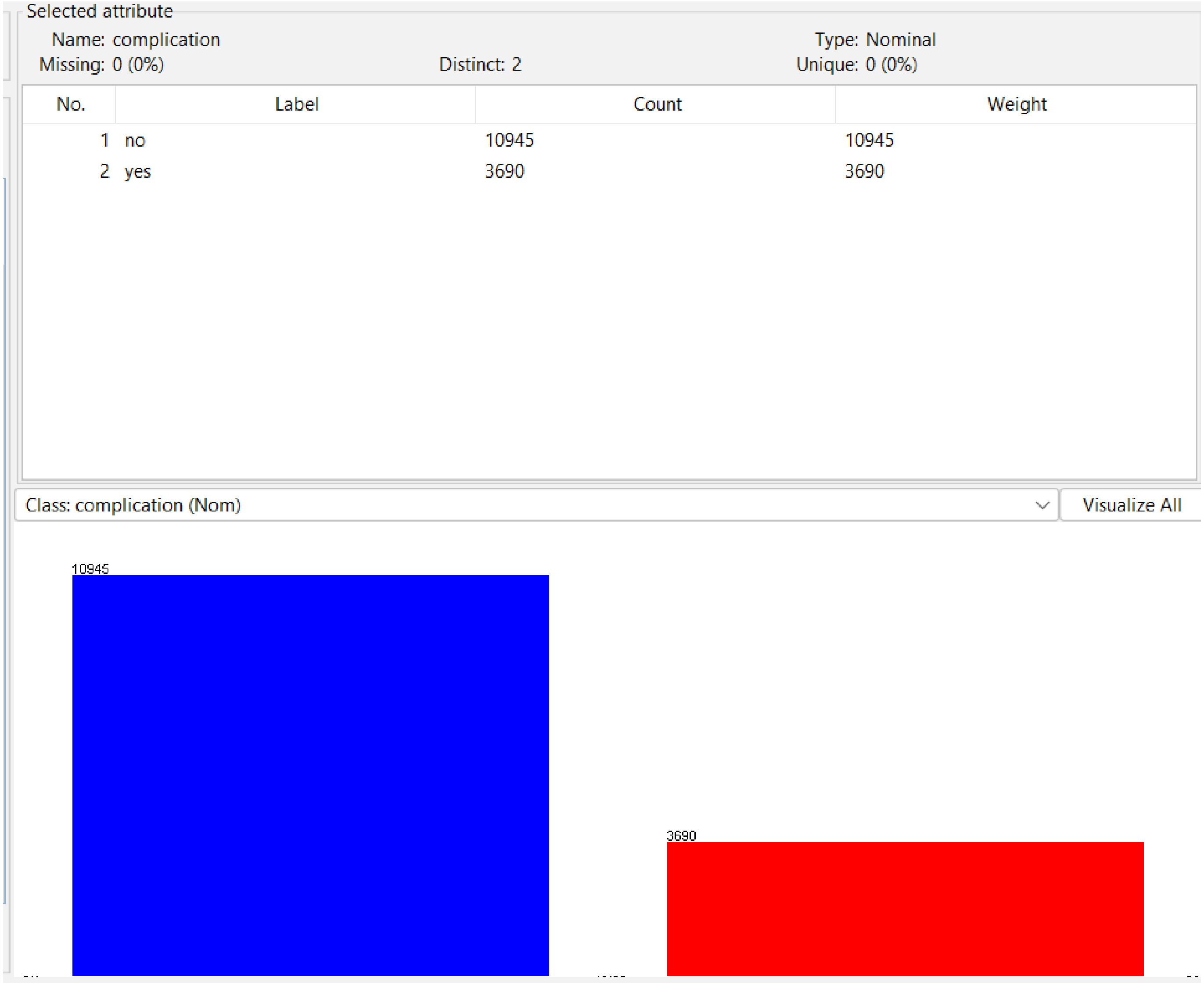
Filter: Randomize -S 42

Current relation
Relation: Dataset_WEKA
Instances: 14635
Attributes: 25
Sum of weights: 14635

Data Exploration:

- **Attribute : baseline_cancer**

  There are 10706 people who has no baseline_cancer (no Symptoms) and 3839 people has a baseline_cancer.



Selected attribute

Name: complication	Type: Nominal

Missing: 0 (0%)	Distinct: 2	Unique: 0 (0%)

No.	Label	Count	Weight
1	no	10945	10945
2	yes	3690	3690

Class: complication (Nom)

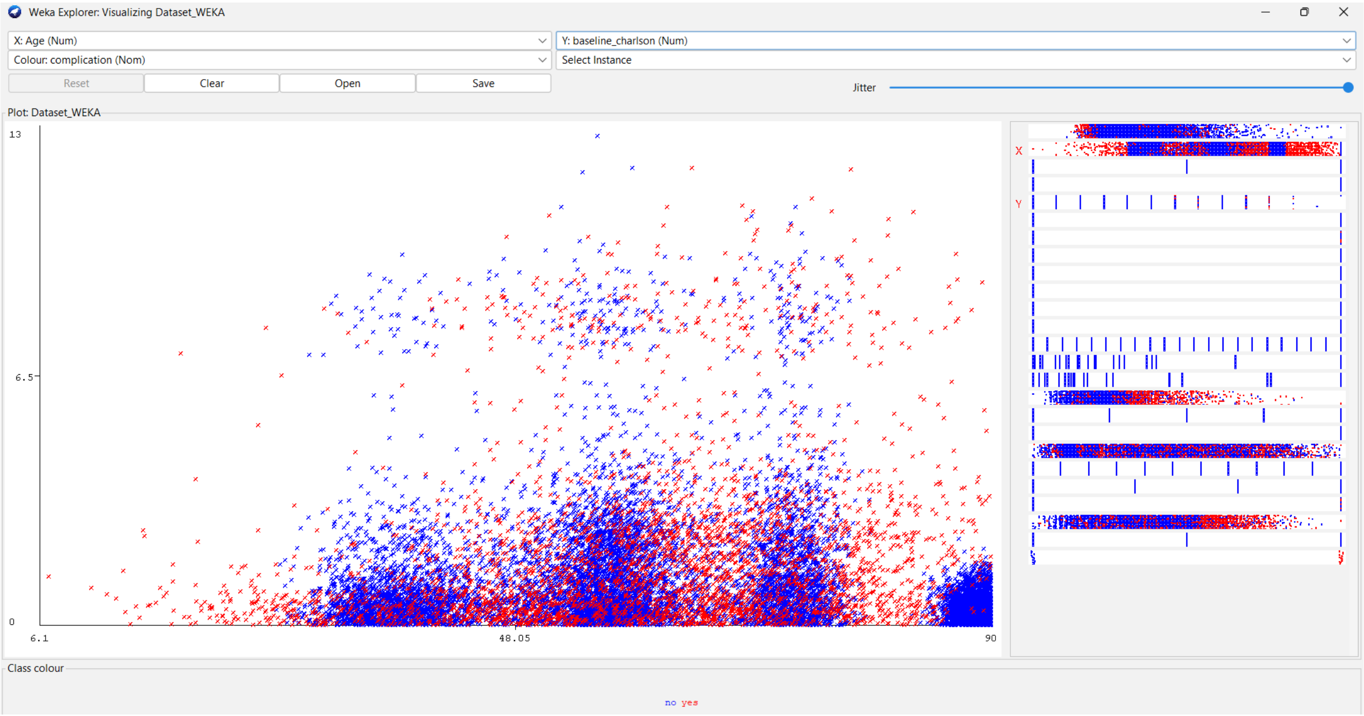
❖ We are presenting the graph based on Age and baseline_charlson.

- X axis : Age
  Minimum age : 6.1
  Maximum age : 90
- Y axis: baseline_charlson
  Minimum value : 0
  Maximum value : 13

- The graph illustrates data that indicates the number of individuals who have baseline_charlson using the Age column/attribute. We could observe that just one MAN (blue color) has **baseline_charlson** with value 13, and that person's age is 57.1 . We can conclude that there is just one woman(red color), age 6.1, whose **baseline_charlson** value is 1.

  Note : Red color marks represent woman and blue color marks represent man in the graph below.

- The graph below illustrates the data insights.



Weka Explorer: Visualizing Dataset_WEKA

X: Age (Num)
Y: baseline_charlson (Num)
Colour: complication (Nom)
Select Instance

- **We are presenting the graph based on Age and baseline_charlson.**
  - X axis : Age

    Minimum age : 6.1

    Maximum age : 90
  - Y axis: mortality_rsi

    Minimum value : -3.82

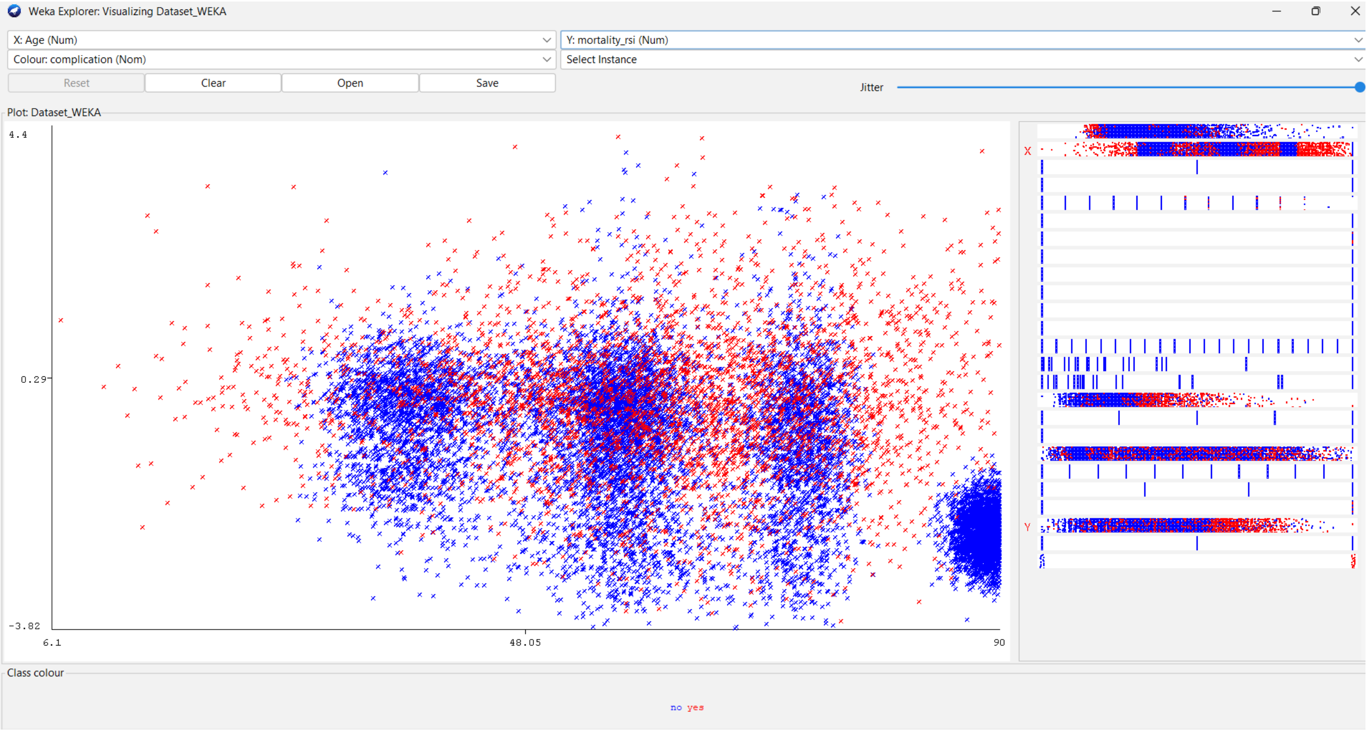
    Maximum value : 4.4

- The figure represents the data of the attribute mortality_rsi using the age attribute.

We can see that there are no blue marks that indicate MAN with a mortality_rsi having value greater than 3.62 and we can see that majority of MAN have mortality_rsi when the age is around 90 .

Note : Red color marks represent woman and blue color marks represent man in the graph below.

- The graph below illustrates the data insights.



Weka Explorer: Visualizing Dataset_WEKA

X: Age (Num)

Y: mortality_rsi (Num)

Colour: complication (Nom)

Select Instance

Plot: Dataset_WEKA

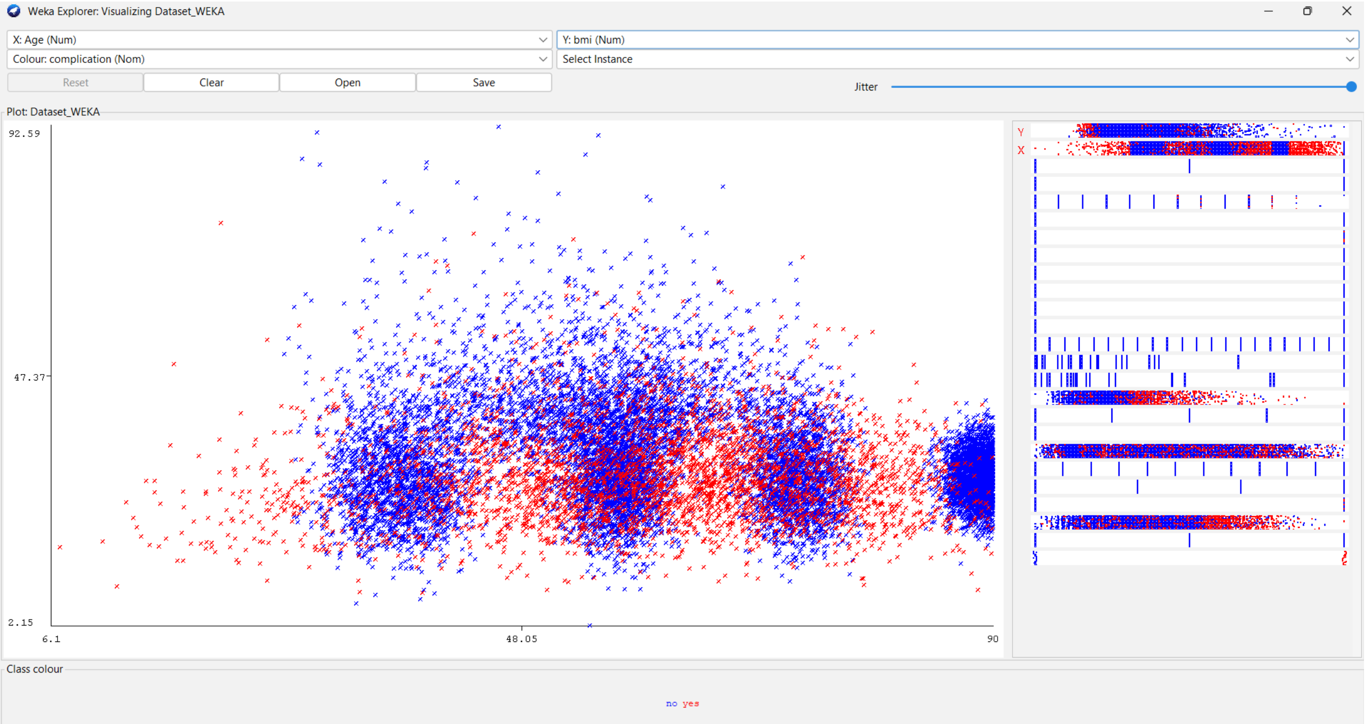
Class colour

❖ We are presenting the graph based on Age and baseline_charlson.

- X axis : Age
  Minimum age : 6.1
  Maximum age : 90
- Y axis: bmi
  Minimum value : 2.15
  Maximum value : 92.59
- There is only one woman with a BMI of 14.4 who is 6.1 years old. There is only one man who possesses the highest BMI, which is 92.59. The individual with the lowest BMI of 2.15 is also a man. In comparison to women, the majority of men are around the age of 90 having BMI between 2.15 to 48 .

Note : Red color marks represent woman and blue color marks represent man in the graph below.

- The graph below illustrates the data insights.



Weka Explorer: Visualizing Dataset_WEKA

X: Age (Num)

Y: bmi (Num)

Colour: complication (Nom)

Select Instance

Plot: Dataset_WEKA

Class colour

Contribution By Each Team Member:

We all met on Teams for discussing and we studied on WEKA tool, we have watched some tutorial video on internet then we have tried to find out the interesting insights of the data. I, Darshana Madalani completed the tasks in WEKA tool and generated this report.

Reference : https://www.tutorialspoint.com/weka/index.htm