

CSC-6500 INTELLIGENT SYSTEMS: ALGORITHMS AND TOOLS

FINAL PROJECT

Name: Kalash Jain

Date: 12/05/2016

Access id: gb4957

Student id: 004550049

Abstract:

The aim of this project is to study the reasons of pollution in the state of Michigan in United States of America and how the occurrence of different pollutants is dependent on each other. Which pollutant is causing the maximum pollution in which part of a state and also in which year. Graphs are made to understand the trend of pollutants year by year. The project is aiming to meet all these goals.

Introduction:

As the Final Project of Intelligent Systems: Algorithms and Tools I have chosen the data of US pollution and in my project I have done operations on the Michigan Data. The data includes the pollution information of two counties of Michigan, first is Kent and other one is Wayne. At first the preprocessing of the data is done and several changes are made in order to make the data set ready for the data mining operation. After that the Association operation is performed in order to understand the relation between different pollutants and their interdependency, impact on the area et al.

Background:

The project is having the local data of two counties of Michigan. The information about the data is listed below:

Context

This dataset deals with pollution in the U.S. Pollution in the U.S. has been well documented by the U.S. EPA but it is a pain to download all the data and arrange them in a format that interests data scientists. Hence I gathered four major pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016 and place them neatly in a csv file.

Content

There is a total of 28 fields:

1. State Code : The code allocated by US EPA to each state
2. County code : The code of counties in a specific state allocated by US EPA

3. Site Num : The site number in a specific county allocated by US EPA
4. Address: Address of the monitoring site
5. State : State of monitoring site
6. County : County of monitoring site
7. City : City of the monitoring site
8. Date Local : Date of monitoring

The four pollutants (NO2, O3, SO2 and O3) each has 5 specific columns. For instance, for NO2:

- NO2 Units : The units measured for NO2
- NO2 Mean : The arithmetic mean of concentration of NO2 within a given day
- NO2 AQI : The calculated air quality index of NO2 within a given day
- NO2 1st Max Value : The maximum value obtained for NO2 concentration in a given day
- NO2 1st Max Hour : The hour when the maximum NO2 concentration was recorded in a given day

State	County	City	Date Local	NO2 Units	NO2 Mean	NO2 1st M	NO2 1st M	NO2 AQI	O3 Units	O3 Mean	O3 1st Ma	O3 1st Ma	O3 AQI	SO2 Units	SO2 Mean	SO2 1st
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/1/2000	Parts per l	19.04167	49	19	46	Parts per i	0.0225	0.04	10	34	Parts per i	3
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/1/2000	Parts per l	19.04167	49	19	46	Parts per i	0.0225	0.04	10	34	Parts per i	3
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/1/2000	Parts per l	19.04167	49	19	46	Parts per i	0.0225	0.04	10	34	Parts per i	2.975
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/1/2000	Parts per l	19.04167	49	19	46	Parts per i	0.0225	0.04	10	34	Parts per i	2.975
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/2/2000	Parts per l	22.95833	36	19	34	Parts per i	0.013375	0.032	10	27	Parts per i	1.958333
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/2/2000	Parts per l	22.95833	36	19	34	Parts per i	0.013375	0.032	10	27	Parts per i	1.958333
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/2/2000	Parts per l	22.95833	36	19	34	Parts per i	0.013375	0.032	10	27	Parts per i	1.9375
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/2/2000	Parts per l	22.95833	36	19	34	Parts per i	0.013375	0.032	10	27	Parts per i	1.9375
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/3/2000	Parts per l	38.125	51	8	48	Parts per i	0.007958	0.016	9	14	Parts per i	5.25
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/3/2000	Parts per l	38.125	51	8	48	Parts per i	0.007958	0.016	9	14	Parts per i	5.25
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/3/2000	Parts per l	38.125	51	8	48	Parts per i	0.007958	0.016	9	14	Parts per i	5.2
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/4/2000	Parts per l	40.26087	74	8	72	Parts per i	0.014167	0.033	9	28	Parts per i	7.083333
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/4/2000	Parts per l	40.26087	74	8	72	Parts per i	0.014167	0.033	9	28	Parts per i	7.083333
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/4/2000	Parts per l	40.26087	74	8	72	Parts per i	0.014167	0.033	9	28	Parts per i	7.05
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/4/2000	Parts per l	40.26087	74	8	72	Parts per i	0.014167	0.033	9	28	Parts per i	7.05
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/5/2000	Parts per l	48.45	61	22	58	Parts per i	0.006667	0.012	9	10	Parts per i	8.708333
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/5/2000	Parts per l	48.45	61	22	58	Parts per i	0.006667	0.012	9	10	Parts per i	8.708333
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/5/2000	Parts per l	48.45	61	22	58	Parts per i	0.006667	0.012	9	10	Parts per i	8.7
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/5/2000	Parts per l	48.45	61	22	58	Parts per i	0.006667	0.012	9	10	Parts per i	8.7
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/6/2000	Parts per l	39.95	73	8	71	Parts per i	0.01175	0.025	10	21	Parts per i	6.761905
4	13	3002 1645 E R O i Arizona	Maricopa Phoenix	1/6/2000	Parts per l	39.95	73	8	71	Parts per i	0.01175	0.025	10	21	Parts per i	6.761905

This was the initial content of the data.

For the purpose of experimentation with the data set some attributes were deleted and some new attributes were added.

Also only the state of Michigan was chosen as my project was to work on the local data of Michigan. Two counties of Michigan are given, Kent and Wayne.

By using the value of the (pollutant)mean eg. NO2 Mean, I have calculated the NO2 presence, that is calculate in excel using the mean condition. I first calculated the mean of NO2 Mean and then compared each value of NO2 Mean with its actual mean and if it is \geq Mean then “YES” otherwise “NO”.

	A	B	C	D	E	F	G	H	I	J	K	L
	Address	County	City	Date Local	NO2 Units	NO2 Presence	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI	O3 Units	O3 Presence
1	1179 MONROE NW	Kent	Grand Rapids	3/31/2002	Parts per billion	NO	14.041667	34	6	32	Parts per million	
2	1179 MONROE NW	Kent	Grand Rapids	3/31/2002	Parts per billion	NO	14.041667	34	6	32	Parts per million	
3	1179 MONROE NW	Kent	Grand Rapids	3/31/2002	Parts per billion	NO	14.041667	34	6	32	Parts per million	
4	1179 MONROE NW	Kent	Grand Rapids	3/31/2002	Parts per billion	NO	14.041667	34	6	32	Parts per million	
5	1179 MONROE NW	Kent	Grand Rapids	3/31/2002	Parts per billion	NO	14.041667	34	6	32	Parts per million	
6	1179 MONROE NW	Kent	Grand Rapids	4/1/2002	Parts per billion	YES	15.583333	30	6	28	Parts per million	
7	1179 MONROE NW	Kent	Grand Rapids	4/1/2002	Parts per billion	YES	15.583333	30	6	28	Parts per million	
8	1179 MONROE NW	Kent	Grand Rapids	4/1/2002	Parts per billion	YES	15.583333	30	6	28	Parts per million	
9	1179 MONROE NW	Kent	Grand Rapids	4/1/2002	Parts per billion	YES	15.583333	30	6	28	Parts per million	
10	1179 MONROE NW	Kent	Grand Rapids	4/2/2002	Parts per billion	YES	16.5	31	16	29	Parts per million	
11	1179 MONROE NW	Kent	Grand Rapids	4/2/2002	Parts per billion	YES	16.5	31	16	29	Parts per million	
12	1179 MONROE NW	Kent	Grand Rapids	4/2/2002	Parts per billion	YES	16.5	31	16	29	Parts per million	
13	1179 MONROE NW	Kent	Grand Rapids	4/2/2002	Parts per billion	YES	16.5	31	16	29	Parts per million	
14	1179 MONROE NW	Kent	Grand Rapids	4/3/2002	Parts per billion	NO	14.666667	30	23	28	Parts per million	
15	1179 MONROE NW	Kent	Grand Rapids	4/3/2002	Parts per billion	NO	14.666667	30	23	28	Parts per million	
16	1179 MONROE NW	Kent	Grand Rapids	4/3/2002	Parts per billion	NO	14.666667	30	23	28	Parts per million	
17	1179 MONROE NW	Kent	Grand Rapids	4/3/2002	Parts per billion	NO	14.666667	30	23	28	Parts per million	
18	1179 MONROE NW	Kent	Grand Rapids	4/4/2002	Parts per billion	YES	19.583333	35	3	33	Parts per million	
19	1179 MONROE NW	Kent	Grand Rapids	4/4/2002	Parts per billion	YES	19.583333	35	3	33	Parts per million	
20	1179 MONROE NW	Kent	Grand Rapids	4/4/2002	Parts per billion	YES	19.583333	35	3	33	Parts per million	
21	1179 MONROE NW	Kent	Grand Rapids	4/4/2002	Parts per billion	YES	19.583333	35	3	33	Parts per million	
22	1179 MONROE NW	Kent	Grand Rapids	4/5/2002	Parts per billion	YES	16.5	26	22	25	Parts per million	
23	1179 MONROE NW	Kent	Grand Rapids	4/5/2002	Parts per billion	YES	16.5	26	22	25	Parts per million	

By turning the data in this format the weka operations will be efficient to perform and the data is now ready to be loaded into the weka software.

1. The initial looks like below in the weka software.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize | Auto-WEKA

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose None | Apply

Current relation: Relation: michigan pollution4 | Attributes: 28 | Instances: 8182 | Sum of weights: 8182

Attributes:

- All | None | Invert | Pattern
- No. | Name
- 1 | Address
- 2 | County
- 3 | City
- 4 | Date Local
- 5 | NO2 Units
- 6 | NO2 Presence
- 7 | NO2 Mean
- 8 | NO2 1st Max Value
- 9 | NO2 1st Max Hour
- 10 | NO2 AQI
- 11 | O3 Units
- 12 | O3 Presence
- 13 | O3 Mean
- 14 | O3 1st Max Value
- 15 | O3 1st Max Hour
- 16 | O3 AQI

Remove

Selected attribute:

Name: Address | Missing: 0 (0%) | Type: Nominal | Unique: 0 (0%)

No.	Label	Count	Weight
1	1179 MONROE NW	3524	3524.0
2	2451 MARQUETTE	4658	4658.0

Class: CO AQI (Num) | Visualize All

Status: OK | Log | x 0

2. Now in weka some more useless attributes are removed manually.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize | Auto-WEKA

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply

Current relation
 Relation: michigan pollution4
 Instances: 8182
 Attributes: 28
 Sum of weights: 8182

Attributes

No.	Name
1	Address
2	County
3	City
4	Date Local
5	NO2 Units
6	NO2 Presence
7	NO2 Mean
8	NO2 1st Max Value
9	NO2 1st Max Hour
10	NO2 AQI
11	O3 Units
12	O3 Presence
13	O3 Mean
14	O3 1st Max Value
15	O3 1st Max Hour
16	O3 AQI

Remove

Selected attribute

Name: CO Units
 Missing: 0 (0%)
 Distinct: 1
 Type: Nominal
 Unique: 0 (0%)

No.	Label	Count	Weight
1	Parts per million	8182	8182.0

Class: CO AQI (Num) Visualize All

Status: OK Log

3. Now the missing values are treated before starting the actual process of association.
 There are missing values in the Air Quality Index (AQI) attribute of pollutant CO and SO2. The missing values are set to 0.0
 - (i) CO missing values

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize | Auto-WEKA

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply

Current relation
 Relation: michigan pollution4-weka.filters.unsupervised.attribute.Remove-R1,3,5,11,17...
 Instances: 8182
 Attributes: 22
 Sum of weights: 8182

Attributes

No.	Name
7	NO2 AQI
8	O3 Presence
9	O3 Mean
10	O3 1st Max Value
11	O3 1st Max Hour
12	O3 AQI
13	SO2 Presence
14	SO2 Mean
15	SO2 1st Max Value
16	SO2 1st Max Hour
17	SO2 AQI
18	CO Presence
19	CO Mean
20	CO 1st Max Value
21	CO 1st Max Hour
22	CO AQI

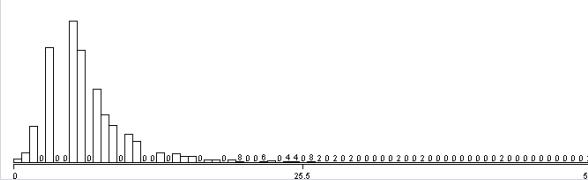
Remove

Selected attribute

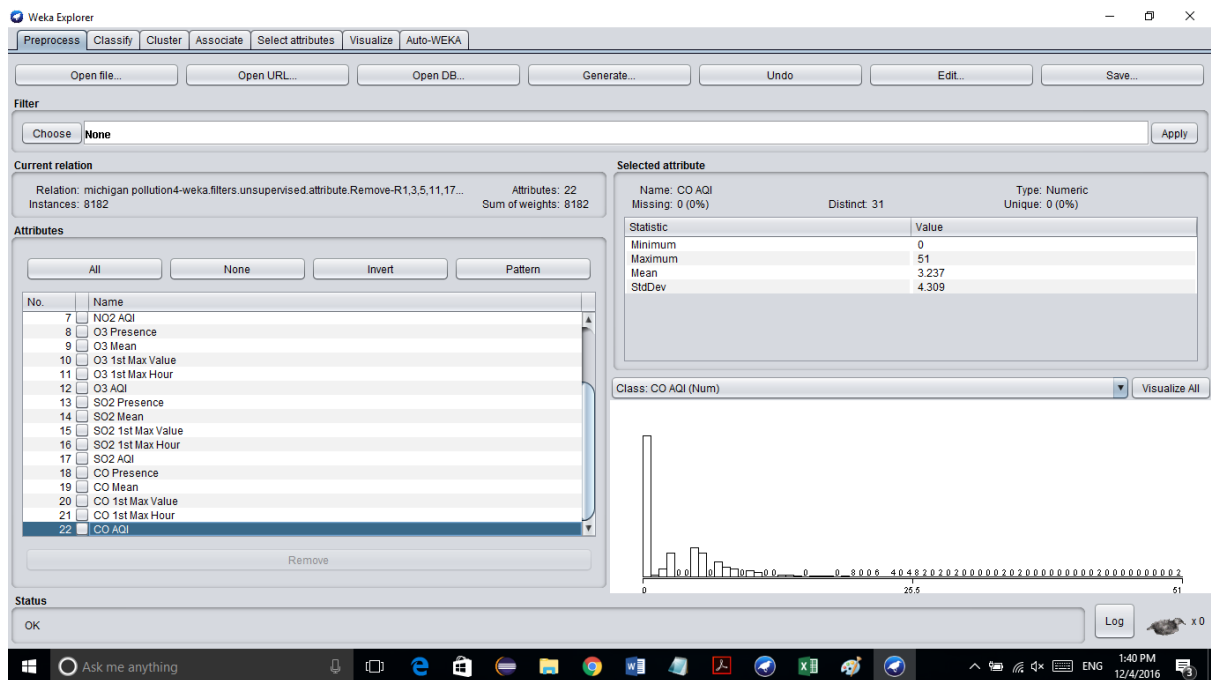
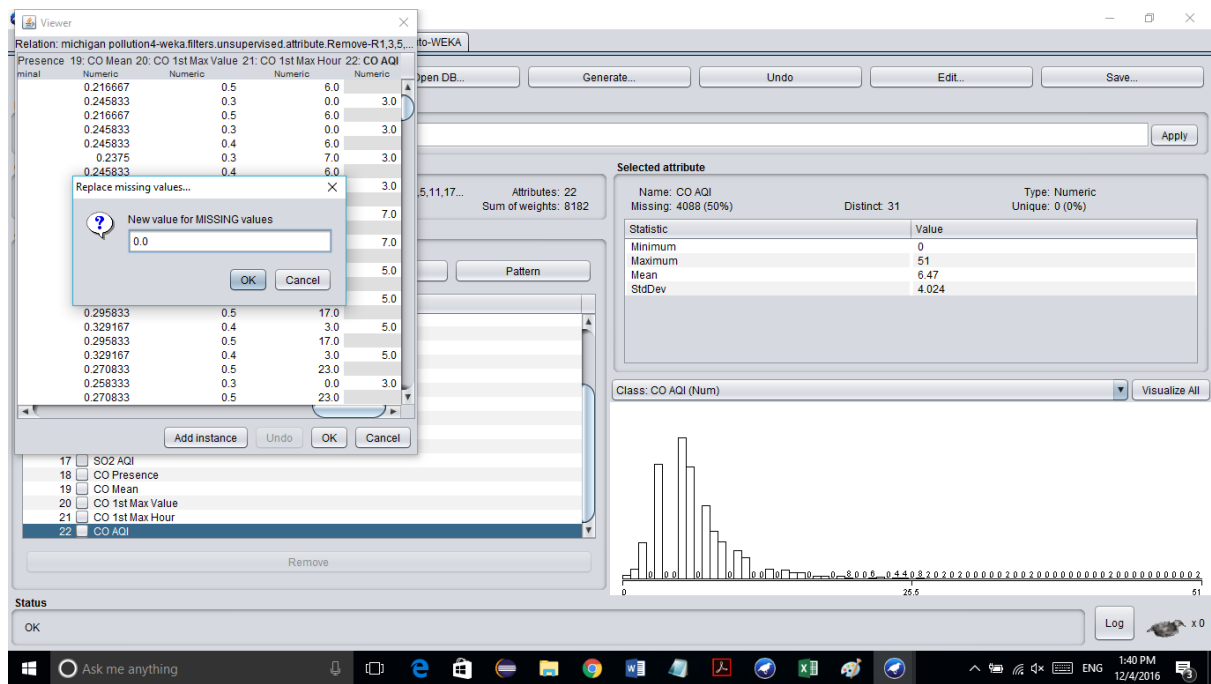
Name: CO AQI
 Missing: 4088 (50%)
 Distinct: 31
 Type: Numeric
 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	51
Mean	6.47
StdDev	4.024

Class: CO AQI (Num) Visualize All



Status: OK Log



(ii) SO2 missing values

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize | Auto-WEKA

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **None** Apply

Current relation

Relation: michigan pollution4-weka.filters.unsupervised.attribute.Remove-R1,3,5,11,17... Attributes: 22 Sum of weights: 8182
Instances: 8182

Attributes

All None Invert Pattern

No.	Name
6	NO2 1st Max Hour
7	NO2 AQI
8	O3 Presence
9	O3 Mean
10	O3 1st Max Value
11	O3 1st Max Hour
12	O3 AQI
13	SO2 Presence
14	SO2 Mean
15	SO2 1st Max Value
16	SO2 1st Max Hour
17	SO2 AQI
18	CO Presence
19	CO Mean
20	CO 1st Max Value
21	CO 1st Max Hour

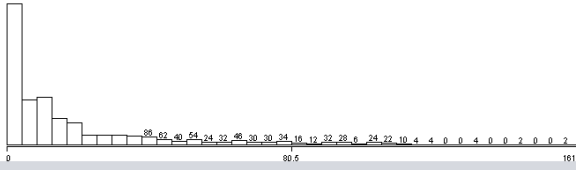
Remove

Selected attribute

Name: SO2 AQI
Missing: 4091 (50%)
Distinct: 94
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	161
Mean	17.971
StdDev	24.238

Class: CO AQI (Num) Visualize All



Status

OK Log x 0

Viewer

Relation: michigan pollution4-weka.filters.unsupervised.attribute.Remove-R1,3,5,11,17... Attributes: 22 Sum of weights: 8182

	Nominal	Numeric	Numeric	Numeric	Nominal			
2 Mean	15	SO2 1st Max Value	16	SO2 1st Max Hour	17	SO2 AQI	18	CO Presence
66667	4.0	15.0	6.0	NO				
66667	4.0	15.0	6.0	NO				
21375	3.0	17.0		NO				
21375	3.0	17.0		NO				
83333	4.0	12.0	6.0	NO				
83333	4.0	12.0	6.0	NO				
20375	2.6	8.0		NO				
20375				NO				
2125				NO				
2125				NO				
21125				NO				
21125				NO				
08333				NO				
08333				NO				
2175				NO				
2175				NO				
2.0	7.0	20.0	10.0	NO				
2.0	7.0	20.0	10.0	NO				
1.975	4.3	20.0		NO				
1.975	4.3	20.0		NO				
08333	4.0	8.0	6.0	NO				
08333	4.0	8.0	6.0	NO				
21875	3.0	8.0		NO				

Replace missing values... New value for MISSING values: 0.0

16 SO2 1st Max Hour
17 SO2 AQI
18 CO Presence
19 CO Mean
20 CO 1st Max Value
21 CO 1st Max Hour

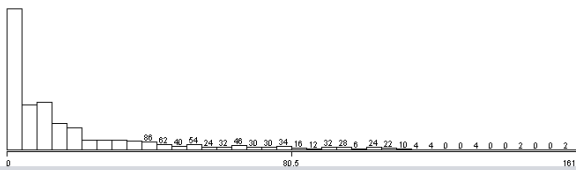
Add instance Undo OK Cancel

Selected attribute

Name: SO2 AQI
Missing: 4091 (50%)
Distinct: 94
Type: Numeric
Unique: 0 (0%)

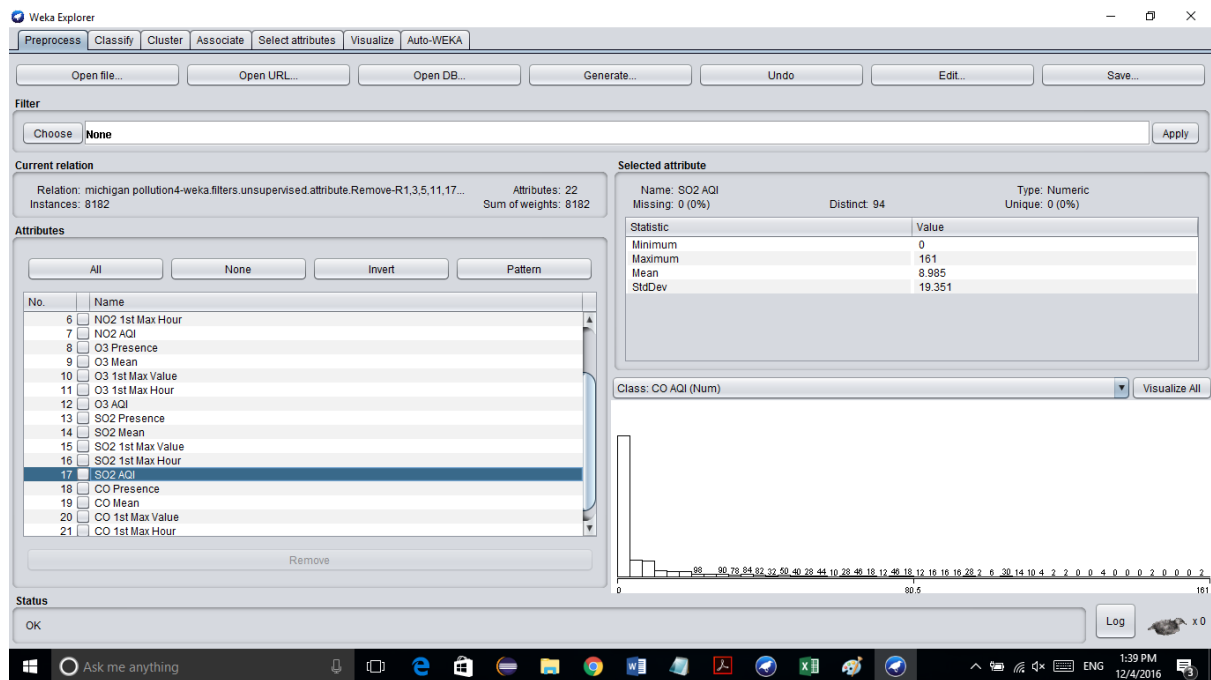
Statistic	Value
Minimum	0
Maximum	161
Mean	17.971
StdDev	24.238

Class: CO AQI (Num) Visualize All

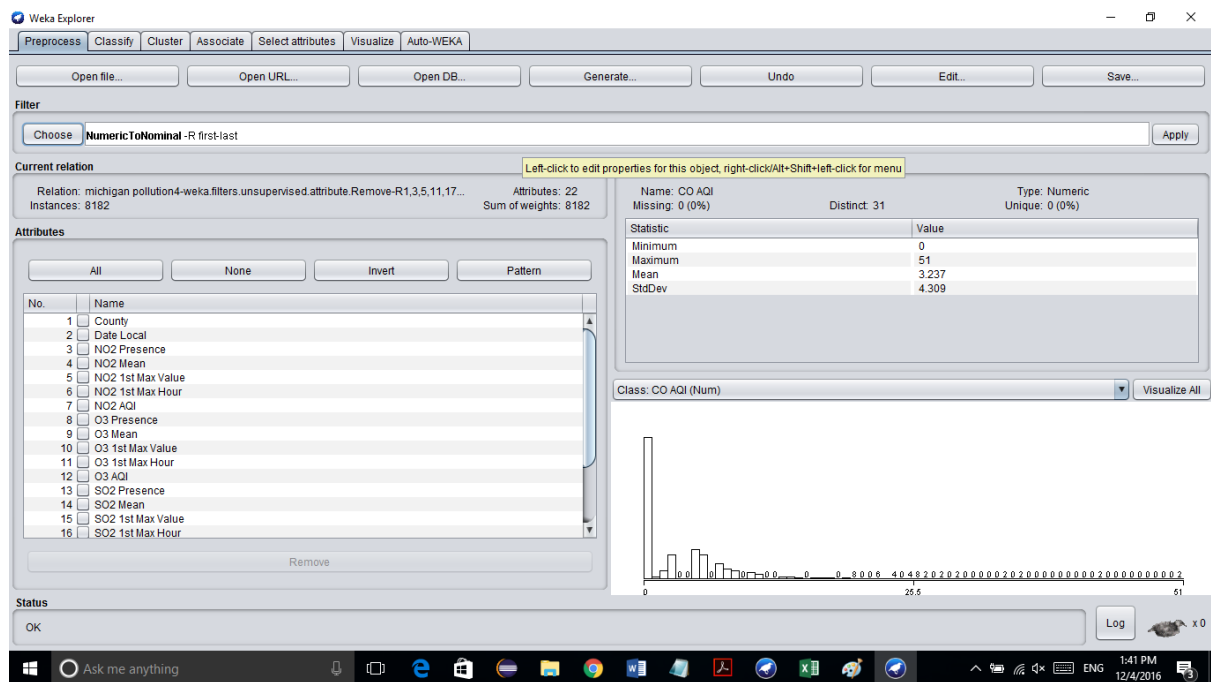


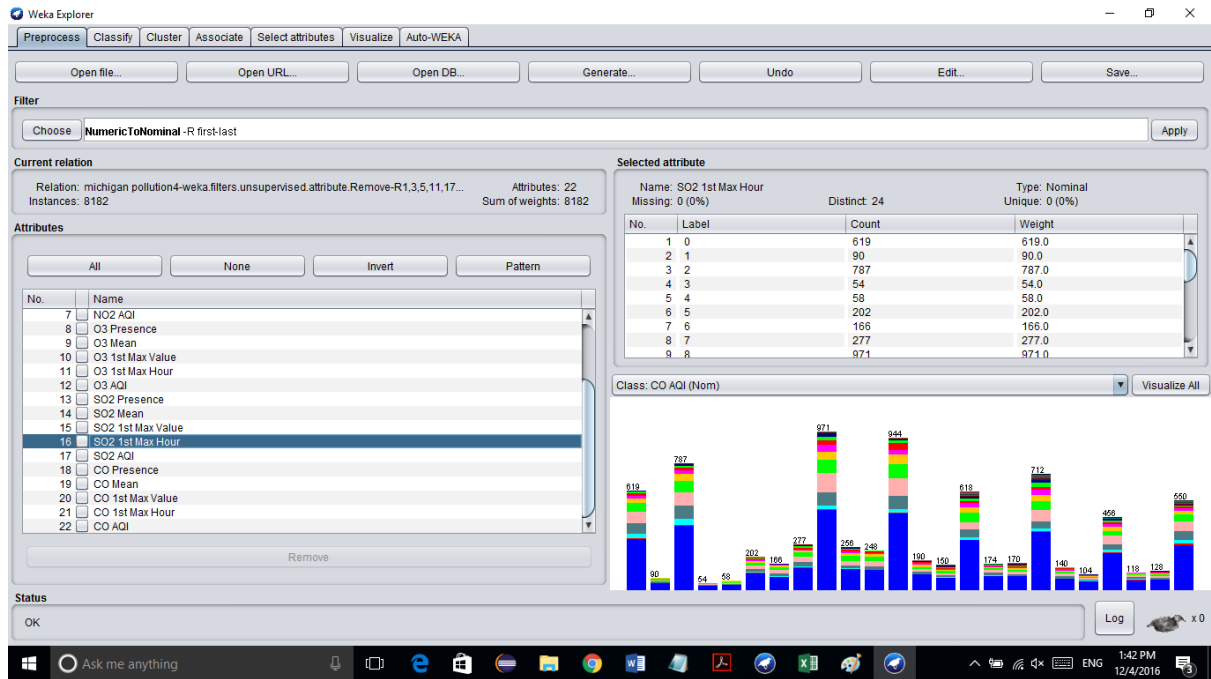
Status

OK Log x 0

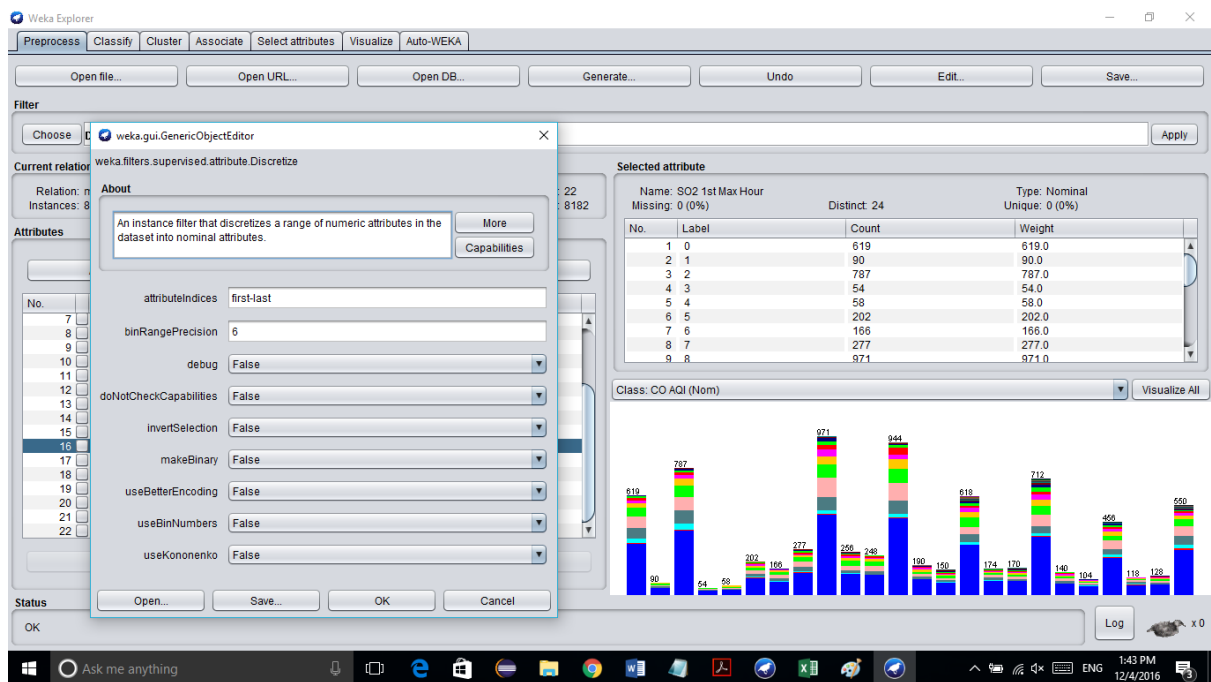


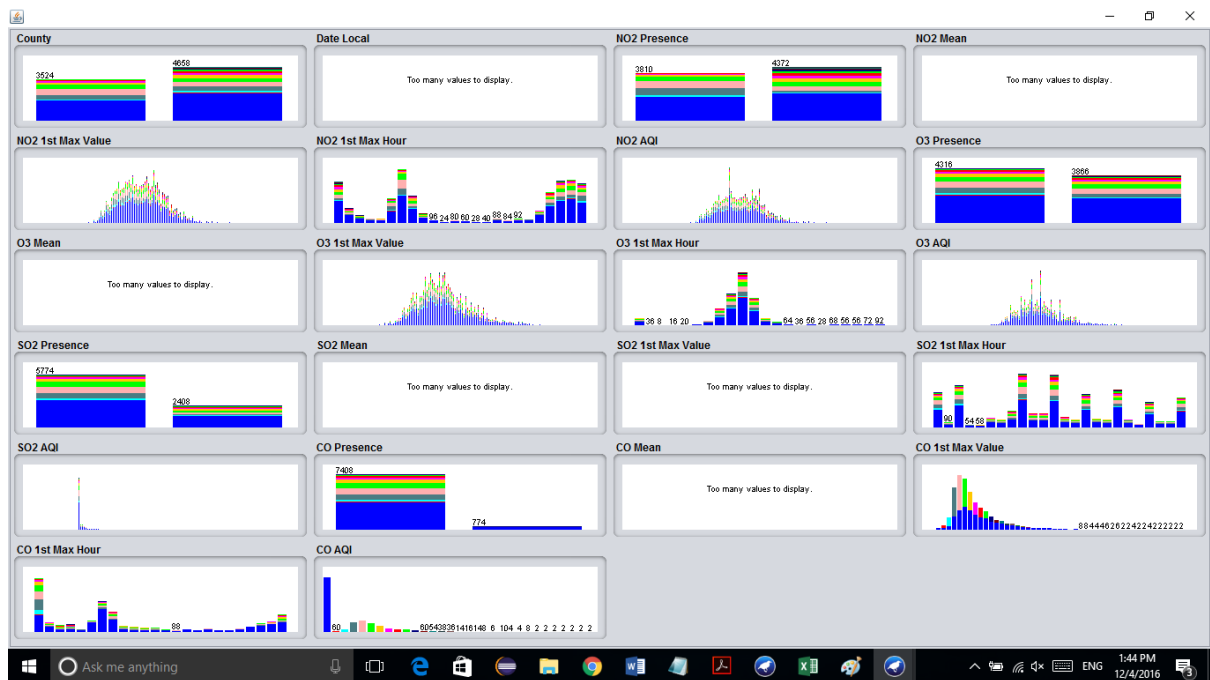
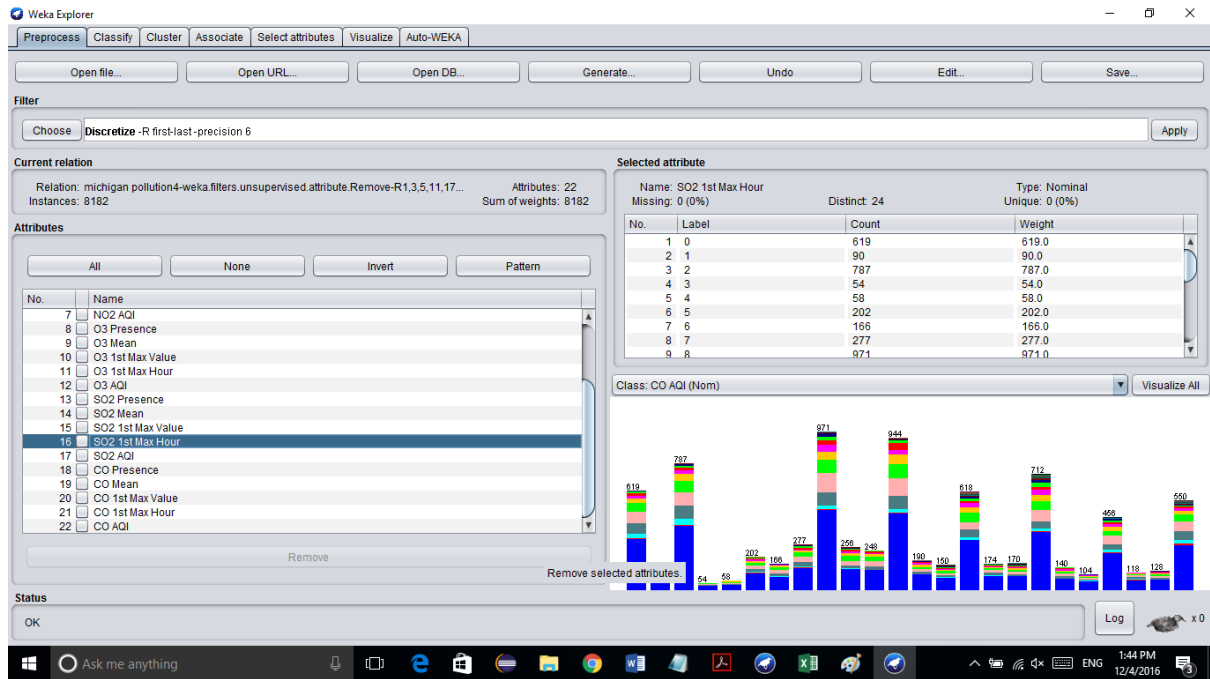
- Now in order to apply the association rule mining the data must get ready for the association operation. At first the numeric values are converted to the nominal values using the unsupervised filter of attributes.





- Now after numerical to nominal conversion all the attributes are set to the nominal values. The attributes are now discretized and using the supervised filter of attribute and the following output is being generated.

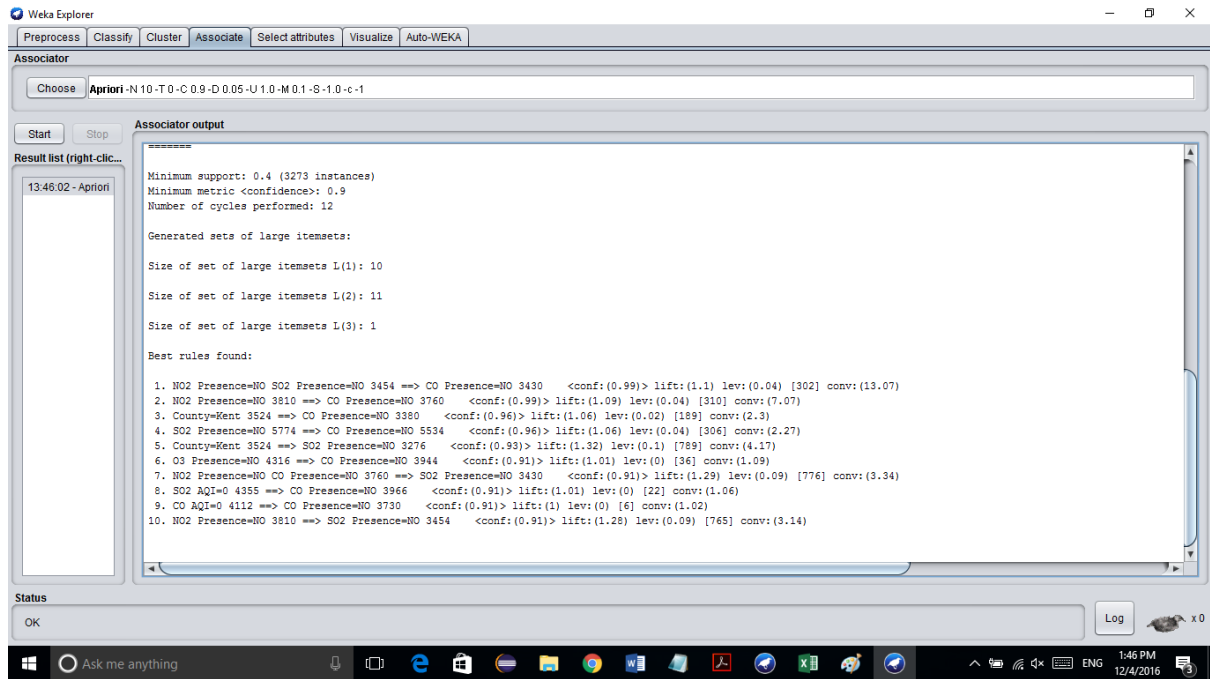




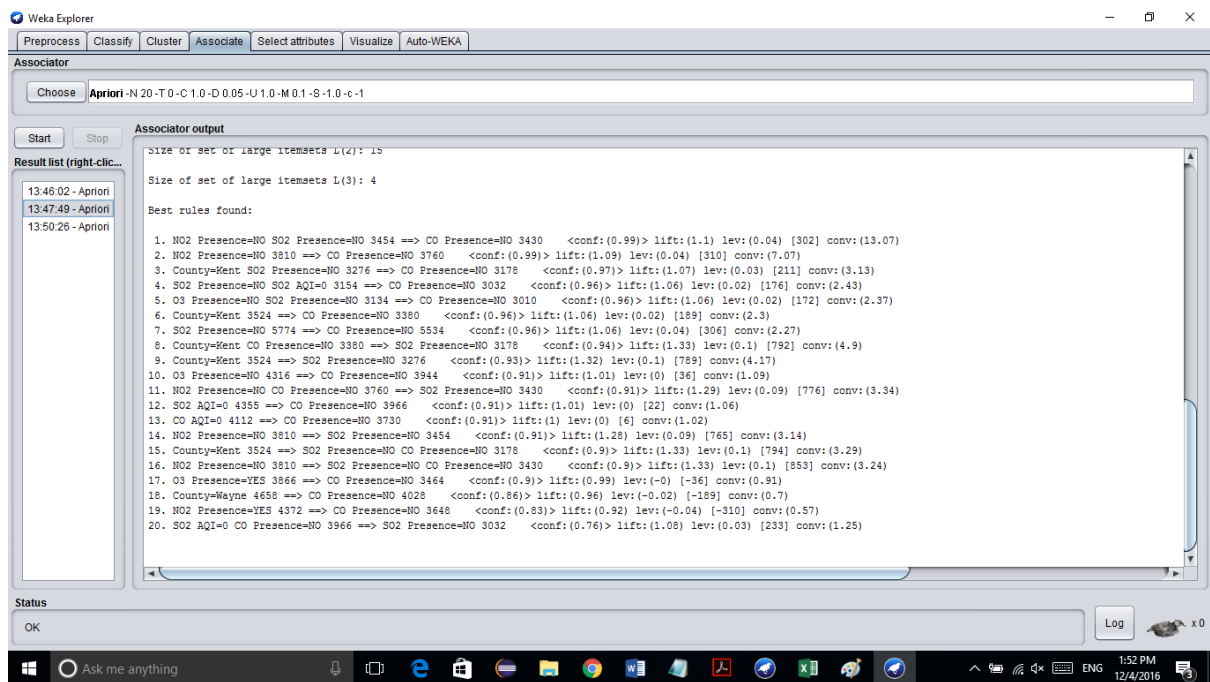
II ASSOCIATION RULE MINING:

The data set is now ready for the association rule mining algorithm and the Apriori algorithm is used and the rules are associated by changing the values of the attributes of the Apriori algorithm. The following set of rules are generated in the different run of the algorithm.

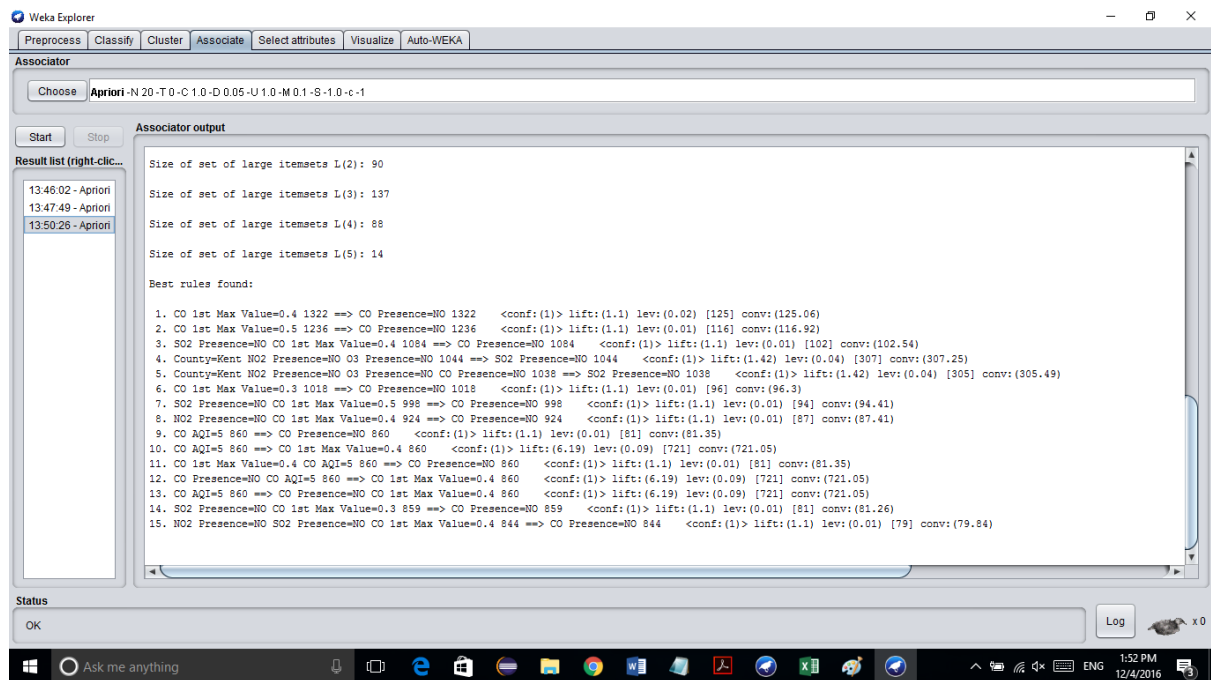
- (i) Minimum support: 0.4 (3273 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12



- (ii) Minimum support: 0.35 (2864 instances)
 Minimum metric <confidence>: 0.7
 Number of cycles performed: 13



- (iii) Minimum support: 0.1 (818 instances)
 Minimum metric <confidence>: 1
 Number of cycles performed: 18



Results:

Based on the operation performed in the weka and excel above the rules that are mined are very useful in understanding the main cause of pollution and also the interdependency between the pollutants and the area to which they belong.

On seeing all the rules obtained by different combinations of the data association techniques some rules are very informative and they are as follows:

1. NO2 Presence=NO CO Presence=NO 3760 ==> SO2 Presence=NO 3430
<conf:(0.91)> lift:(1.29) lev:(0.09) [776] conv:(3.34)

Significance: This rule specifies that if NO2 is not present and CO is also not present then SO2 will also not be present in any of the two counties.

2. County=Kent SO2 Presence=NO 3276 ==> CO Presence=NO 3178 <conf:(0.97)>
lift:(1.07) lev:(0.03) [211] conv:(3.13)

Significance: If the county is Kent and if SO2 is not present then the presence of CO will not be there.

3. NO2 Presence=YES SO2 Presence=NO 2320 ==> CO Presence=NO 2104
<conf:(0.91)> lift:(1) lev:(0) [3] conv:(1.01)

Significance: If NO2 is present and SO2 is not present then CO will also not be present. Thus it shows that the presence or absence of NO2 is not dependent on either SO2 or CO.

4. O3 Presence=YES SO2 Presence=NO 2640 ==> CO Presence=NO 2524
<conf:(0.96)> lift:(1.06) lev:(0.02) [133] conv:(2.13)

Significance: This rule signifies that if O3 is present and SO2 is not present then CO will also not be present. Thus, presence or absence of O3 is also not dependent on SO2 and CO.

5. CO 1st Max Value=0.4 1322 ==> CO Presence=NO 1322 <conf:(1)> lift:(1.1) lev:(0.02) [125] conv:(125.06)

Significance: If the value of the maximum CO concentration recorded in a day is 0.41322 then CO will not be present and out of 8182 instances this pattern is observed in 1322 of them.

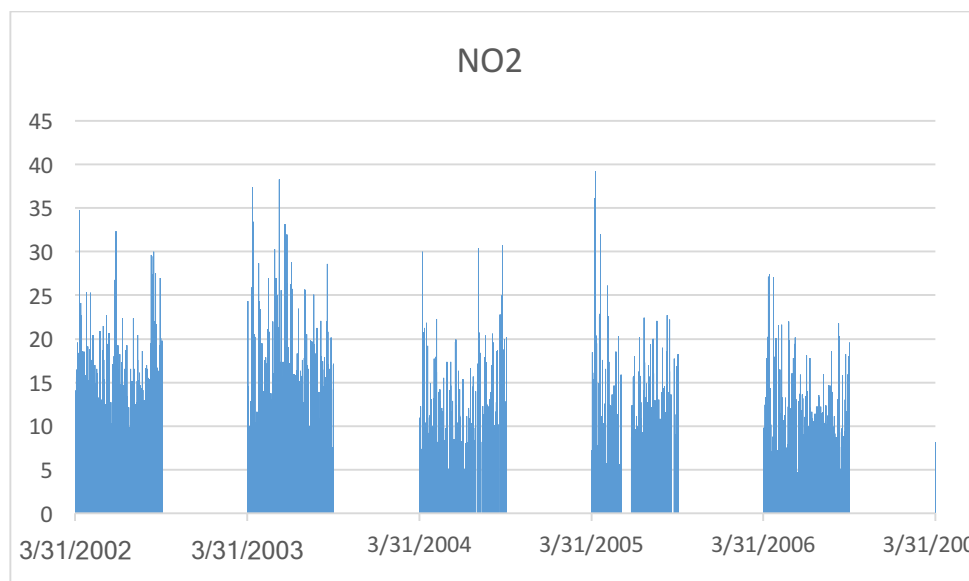
6. County=Kent NO2 Presence=NO O3 Presence=NO CO Presence=NO 1038 ==> SO2 Presence=NO 1038 <conf:(1)> lift:(1.42) lev:(0.04) [305] conv:(305.49)

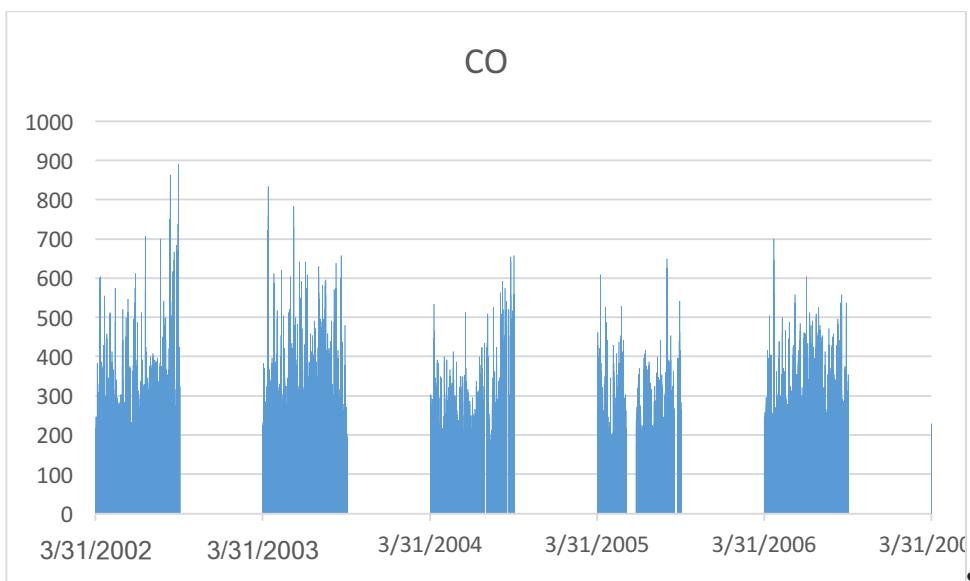
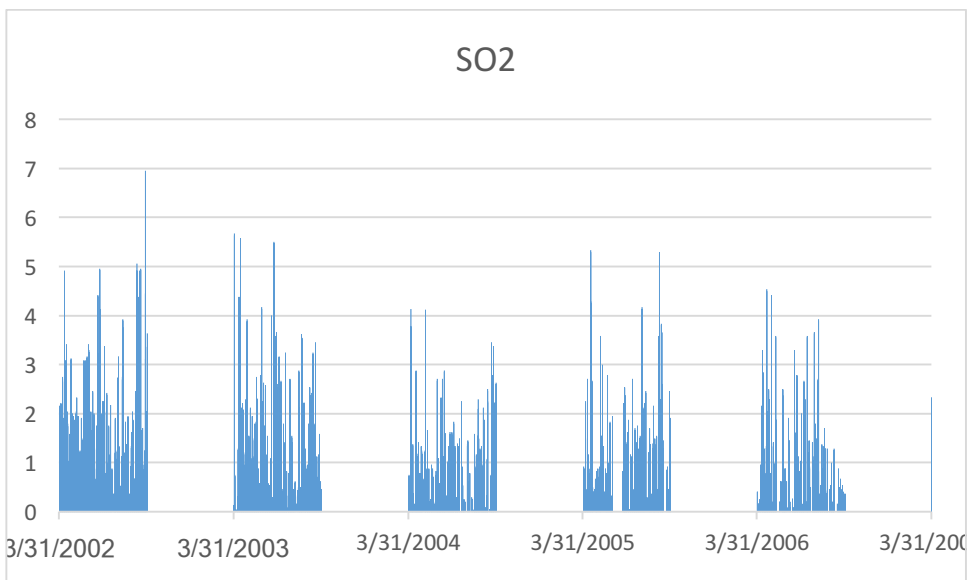
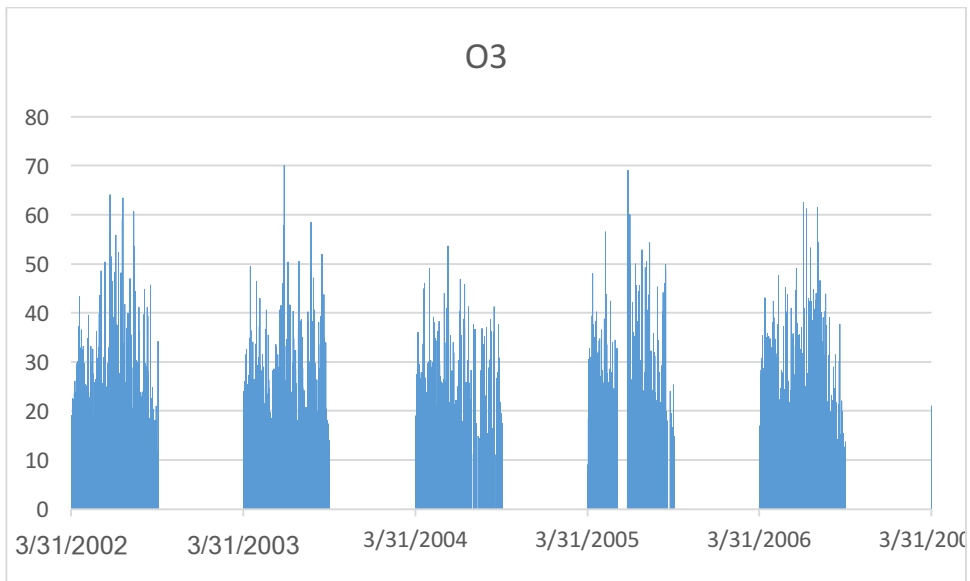
Significance: This rule is the biggest of all and covering all the pollutants. It says that in the Kent county, if NO2, O3, CO is not present then SO2 will also not be present and out of 3524 data entries of the Kent county 1038 are supporting this hypothesis.

Other than the Association operation on the weka some other excel operations were also performed on the data and the graphs are made for understanding the distribution of pollutants over the year.

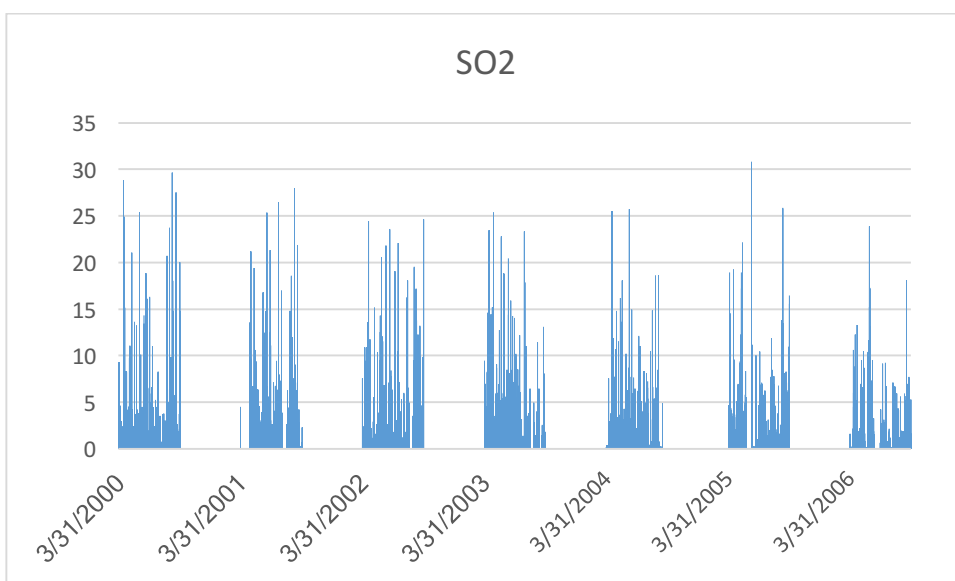
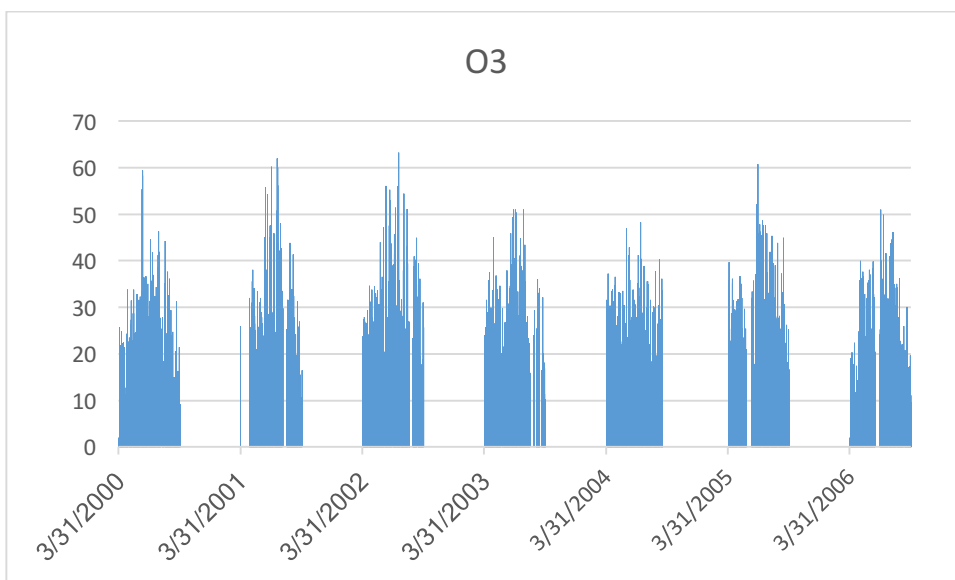
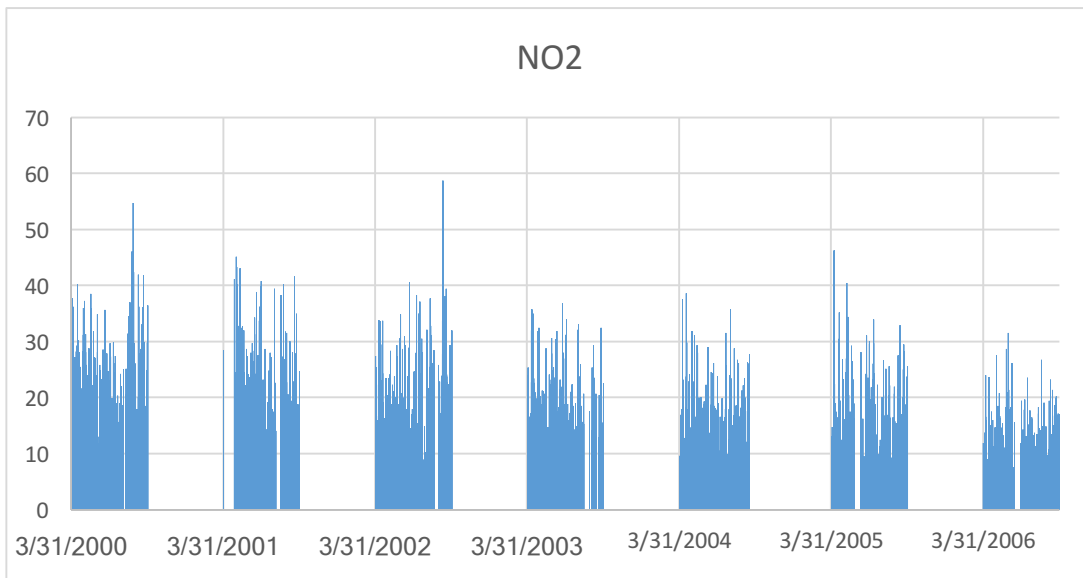
All the graphs are made with the years on the x- axis and the unit in Parts per Bollion on the y-axis. Those pollutants whose units were not ppb were first converted and then graphically represented.

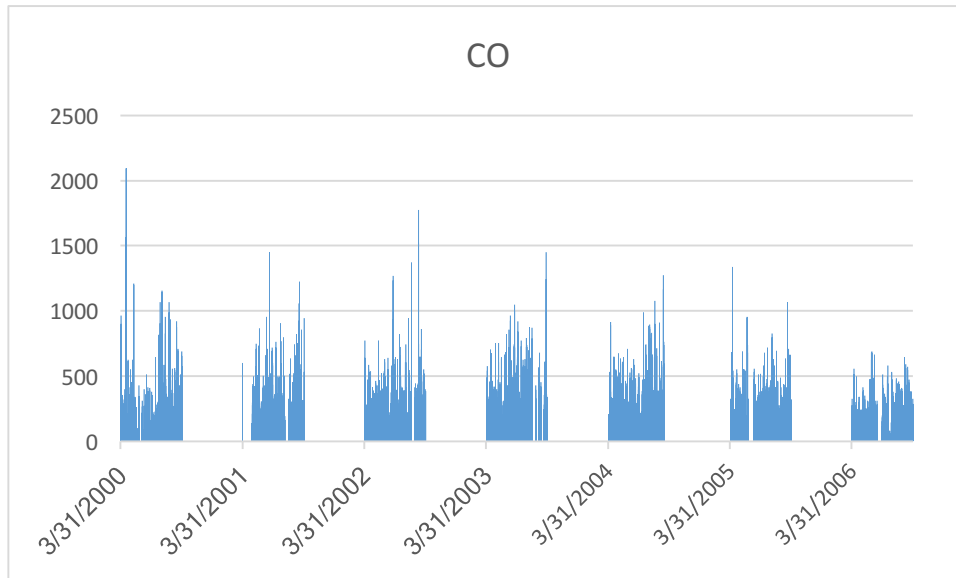
I Kent County:





I Wayne County:





The county wise details of the maximum mean value of a pollutant being recorded.

County	City	Date Local	Max Mean Value (ppb)	Pollutant
Kent	Grand Rapids	4/7/2005	39.208333	NO2
Kent	Grand Rapids	9/27/2002	891.667	CO
Kent	Grand Rapids	6/25/2003	70.167	O3
Kent	Grand Rapids	9/27/2002	6.956522	SO2

County	City	Date Local	Max Mean Value (ppb)	Pollutant
Wayne	Detroit	9/9/2002	58.7	NO2
Wayne	Detroit	4/15/2000	2095.833	CO
Wayne	Detroit	7/17/2002	63.167	O3
Wayne	Detroit	9/9/2002	13.791667	SO2

Thus by looking at the tables and graph above it is clear that:

- Detroit is more polluted than Grand Rapids.
- The maximum polluting pollutant in both the counties is CO.
- The least pollution is done by SO2 in both the counties.
- The rate of pollution of a pollutant is decreasing or increasing in a particular county in a month or a year, but no pattern is observed.

Related Work:

In order to understand the data set and the pollution criteria I visited some websites and understood the degree and units of various pollutants.

For the data understanding purpose and for excel and weka operations the following websites were visited:

<https://www.kaggle.com/sogun3/uspollution>

https://airnow.gov/index.cfm?action=aqi_brochure.index

<https://exceljet.net/excel-functions/excel-if-function>

<https://www.wunderground.com/history/?MR=1>

<http://www.cs.waikato.ac.nz/ml/weka/>

https://blackboard.wayne.edu/bbcswebdav/pid-5861272-dt-content-rid-10313276_2/courses/CSC_5800_1609_001/Weka%20-%20Association%20Analysis_Jayyousi.pdf

Conclusion:

This project aimed at finding the relation and association between different pollutants and also the characteristic value of a pollutant. As far as the primary goals are concerned, the work done is good enough to meet the aim. In future in this project more information is tried to be collected like vehicle, industry information which are spreading the pollution, and by using them the potential causes can be proposed and some ways can be suggested to eradicate or lessen the causes.