Αναφορά 2ης Άσκησης Τεχνικές Εξόρυξης Δεδομένων

1.Οπτικοποίηση Δεδομένων

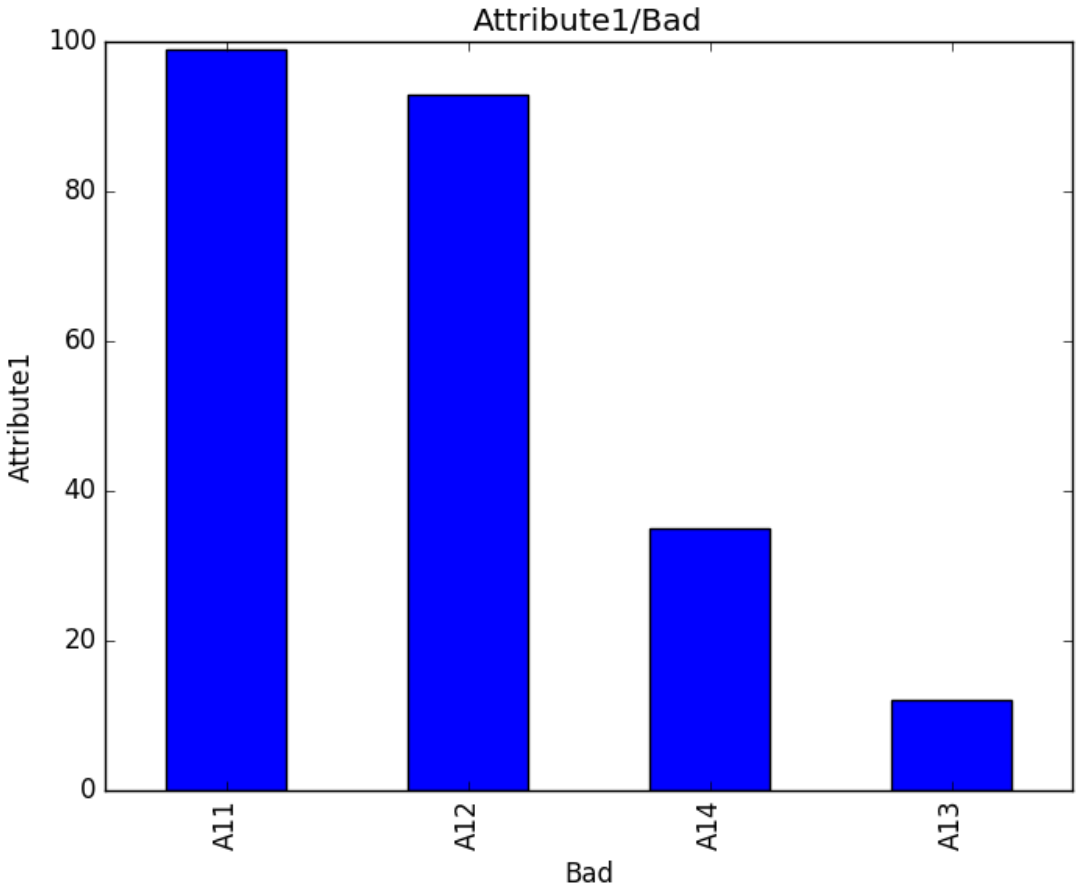
Σε αυτό το ερώτημα υλοποιήσαμε πλήρως τα histograms και box plots για τα categorical και numerical data αντίστοιχα.

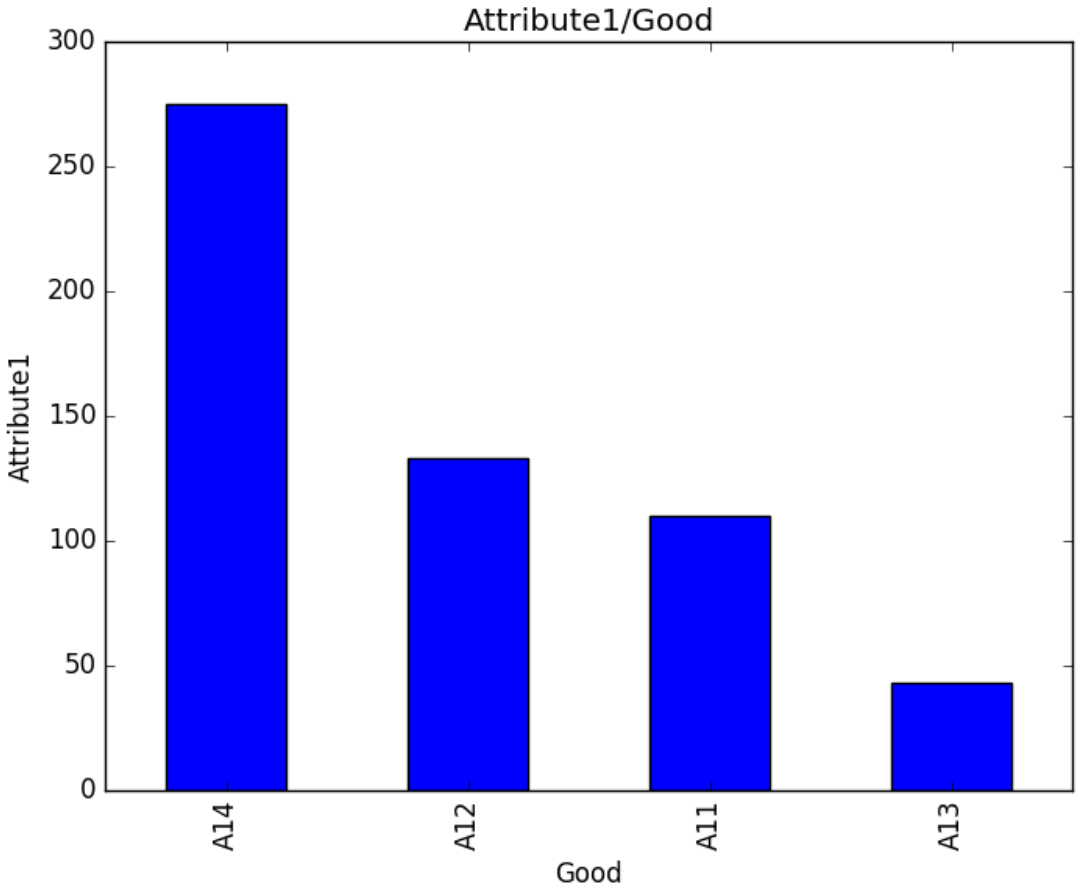
Συμπέρασμα:

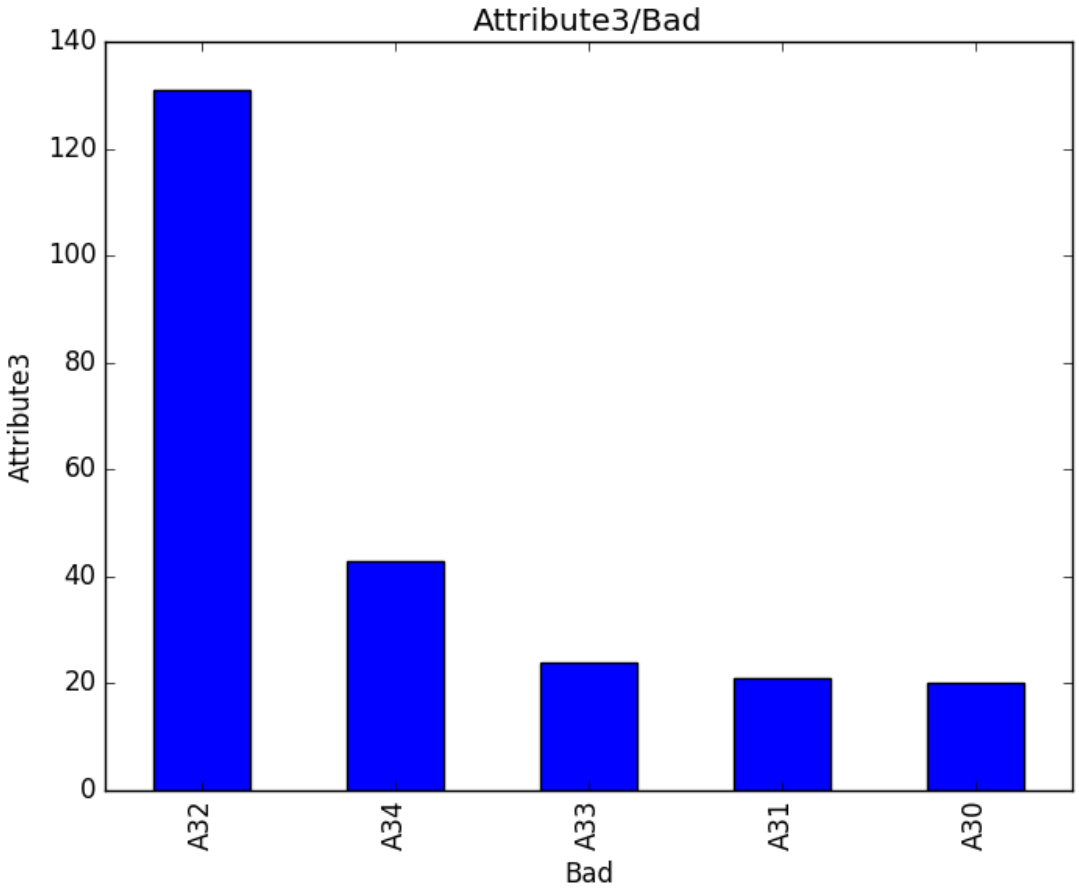
Με βάση την εντροπία, περιμένουμε ως πιο χρήσιμα τα στοιχεία των οποίων τα δεδομένα είναι πιο «μοιρασμένα». Πιο συγκεκριμένα αν οι μπάρες, για παράδειγμα, ενός Attribute είναι ίδιες για όλα του τα στοιχεία(entropy = 1), τότε το Attribute αυτό αποτελεί πολύ καλό στοιχείο για το split του δένδρου αποφάσεων.

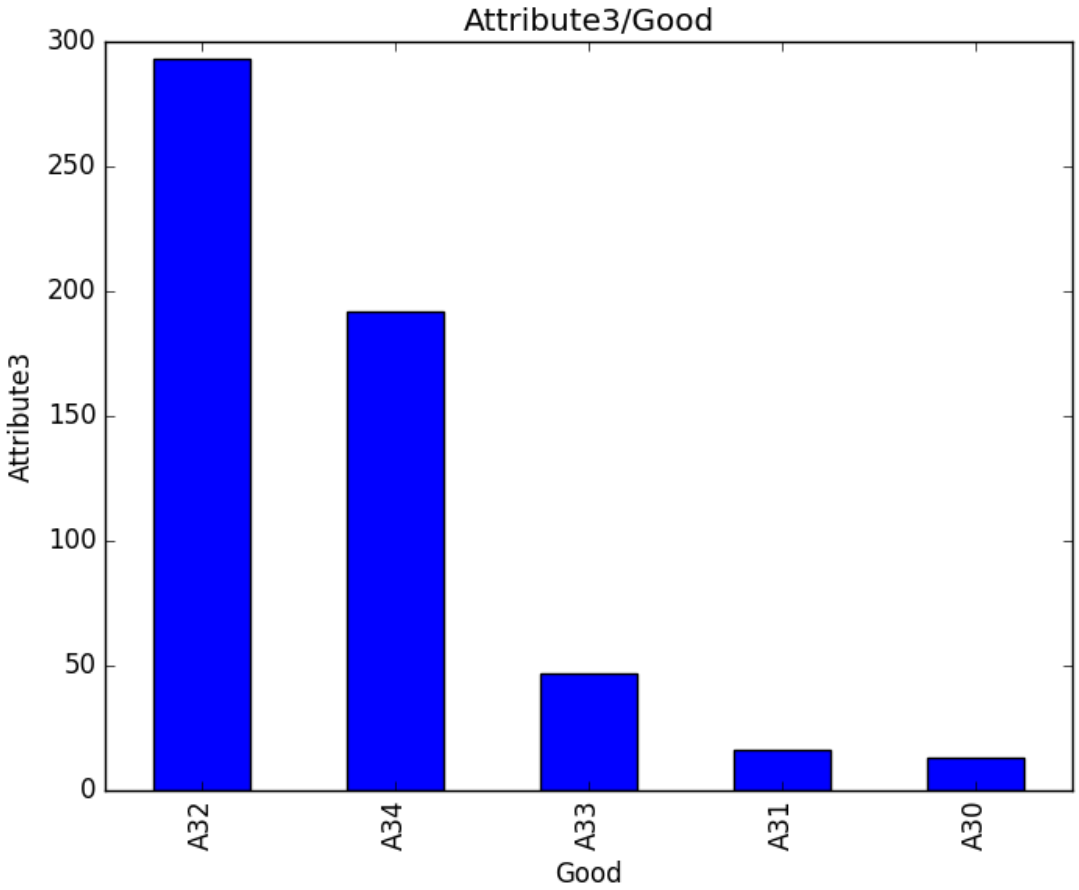
Παραθέτονται όλα τα διαγράμματα(histograms & box plots):

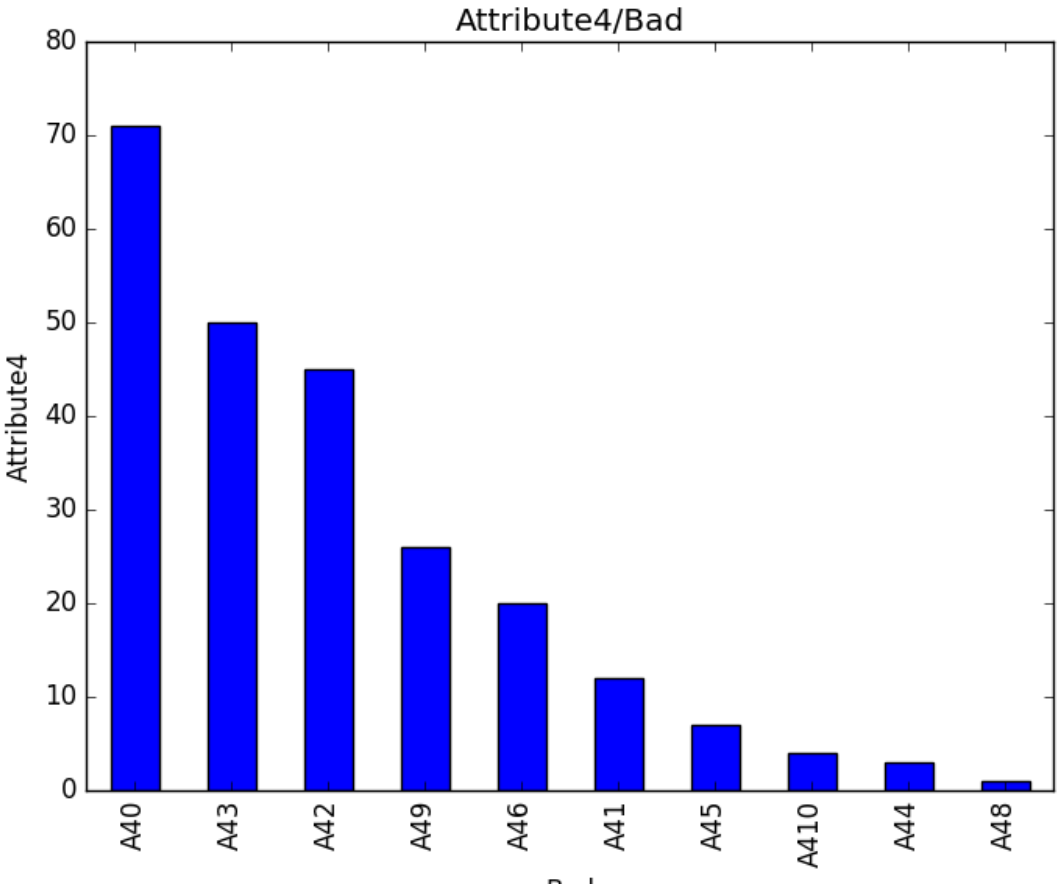
*Επειδή ζητήσατε όλα τα διαγράμματα είναι προφανές ότι το report ξεπερνάει τις 30 σελίδες.

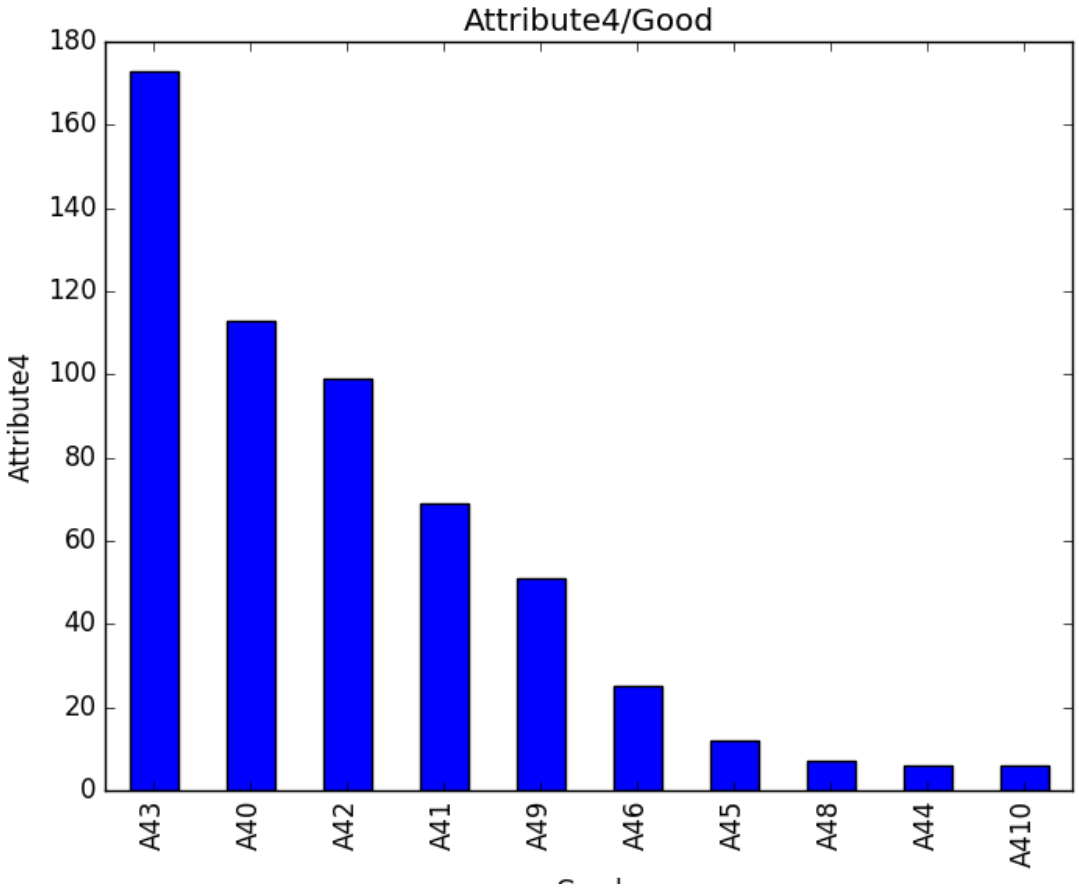

Attribute1/Bad
100
80
60
40
20
0
Attribute1
A11
A12
A14
A13
Bad

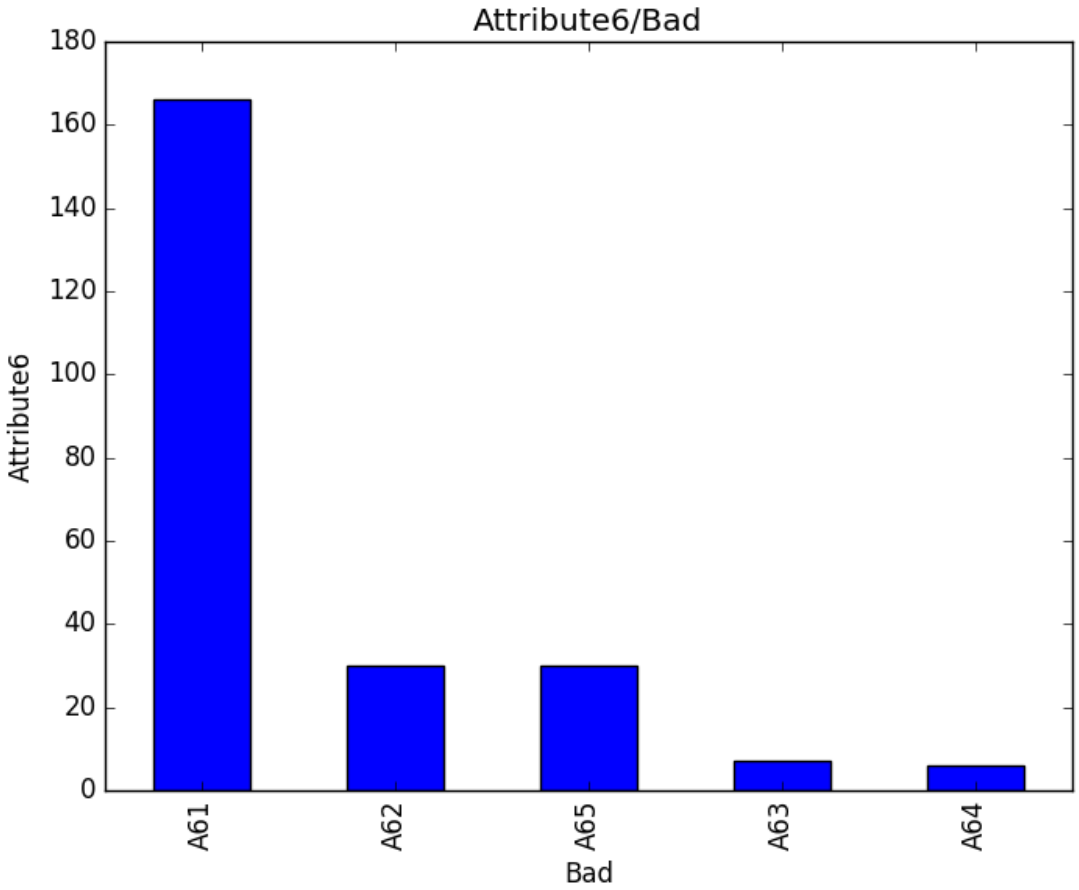

Attribute1/Good
300
250
200
150
100
50
0
Attribute1
A14
A12
A11
A13
Good

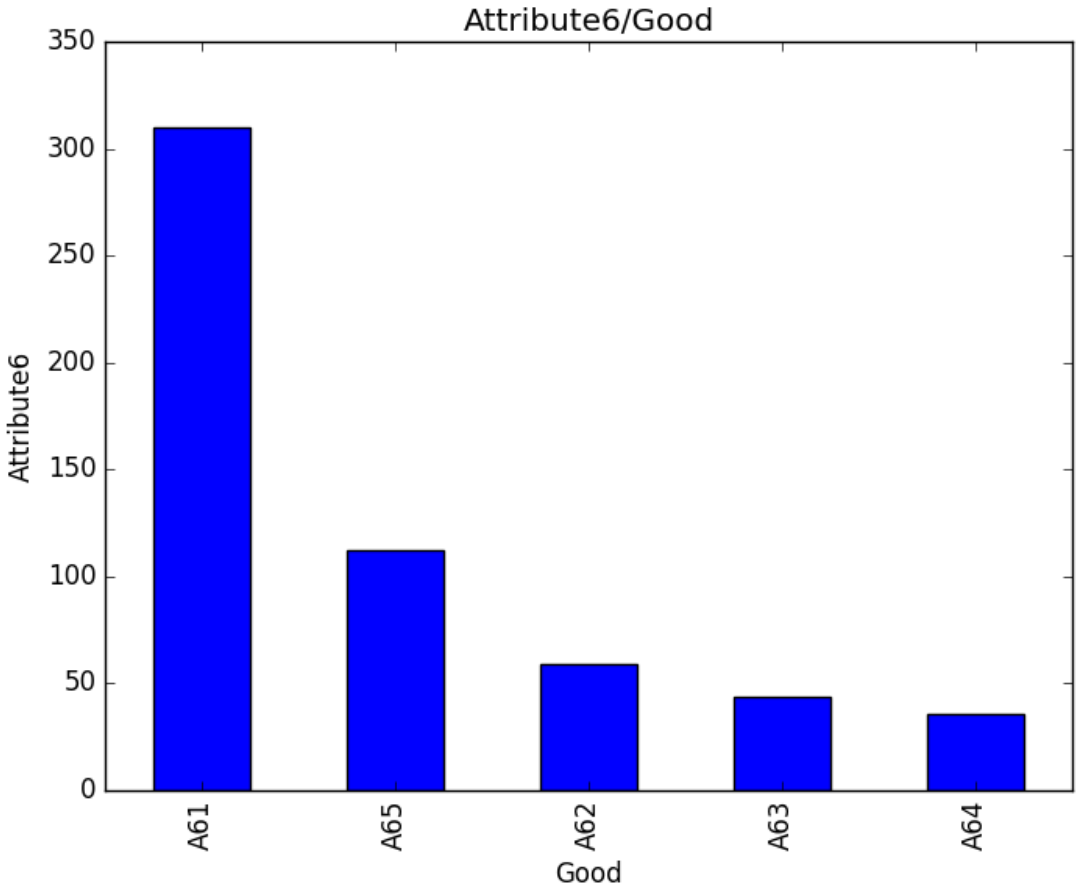

Attribute3/Bad
140
120
100
80
60
40
20
0
Attribute3
A32
A34
A33
A31
A30
Bad

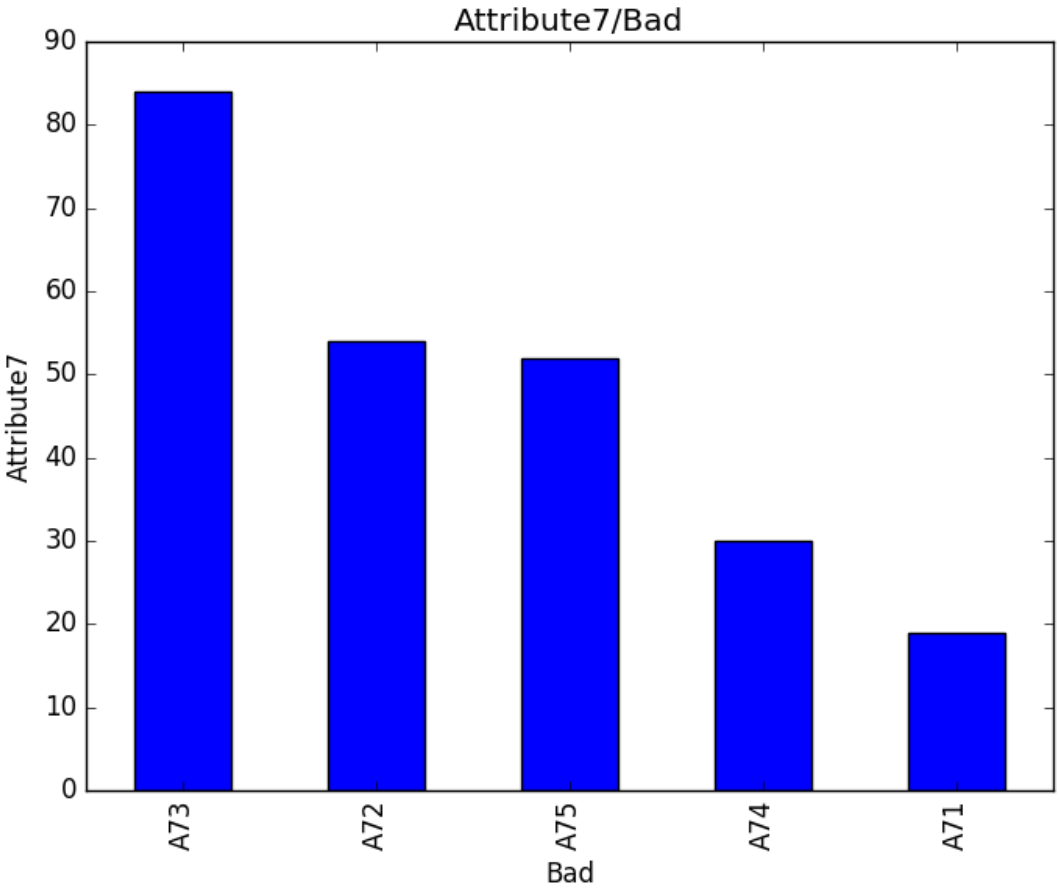

Attribute3/Good
300
250
200
150
100
50
0
Attribute3
A32
A34
A33
A31
A30
Good

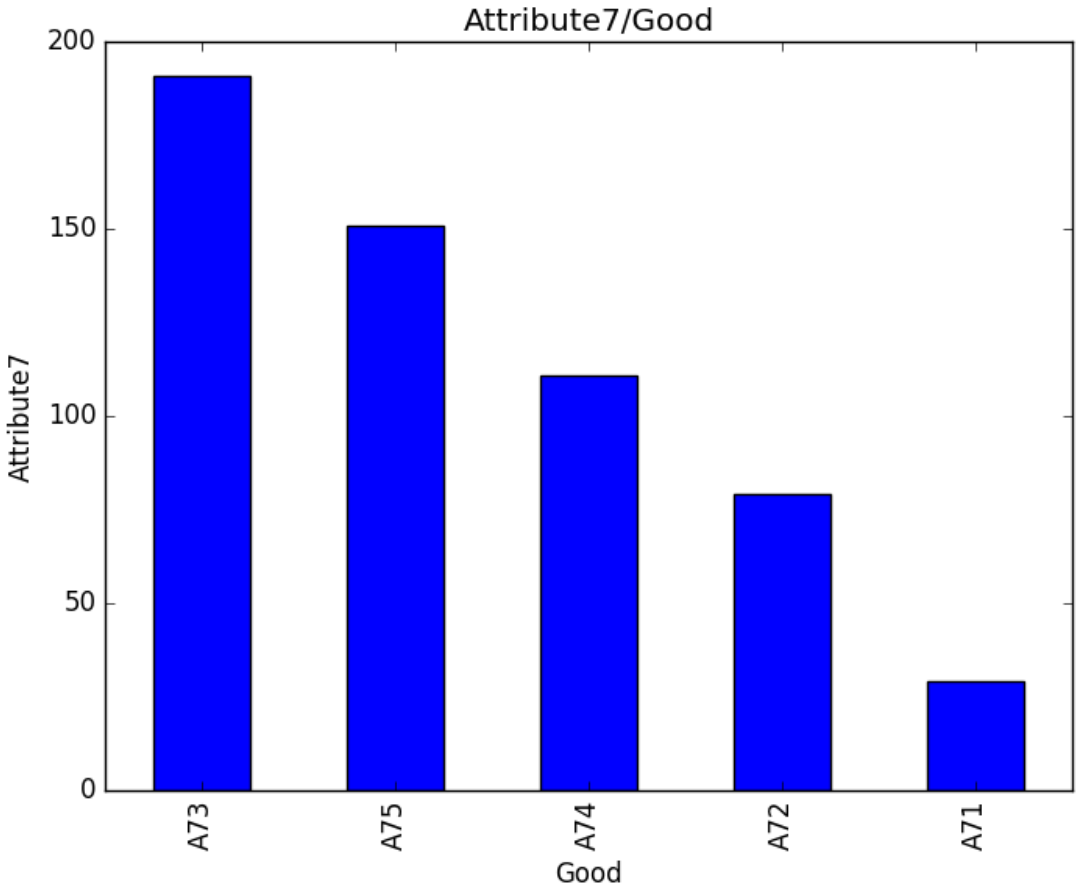

Attribute4/Bad
Attribute4
80
70
60
50
40
30
20
10
0
A40
A43
A42
A49
A46
A41
A45
A410
A44
A48

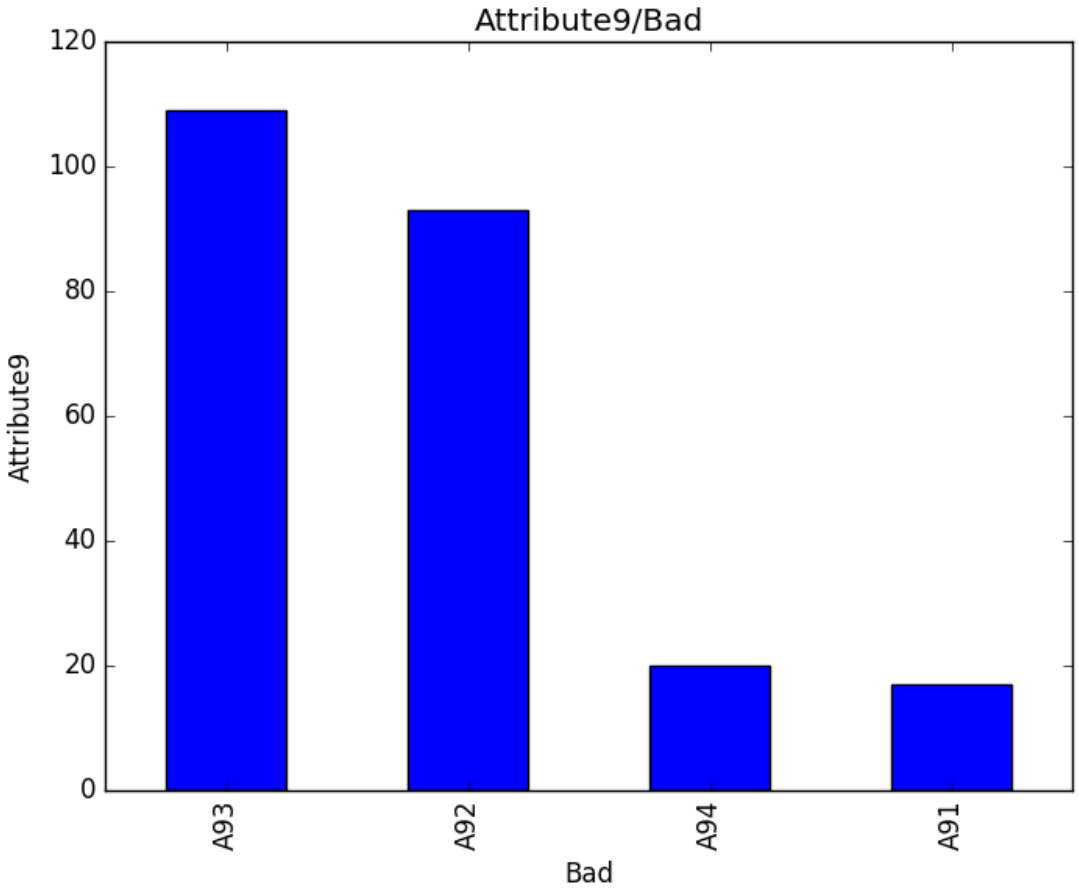

Attribute4/Good
Attribute4
180
160
140
120
100
80
60
40
20
0
A43
A40
A42
A41
A49
A46
A45
A48
A44
A410

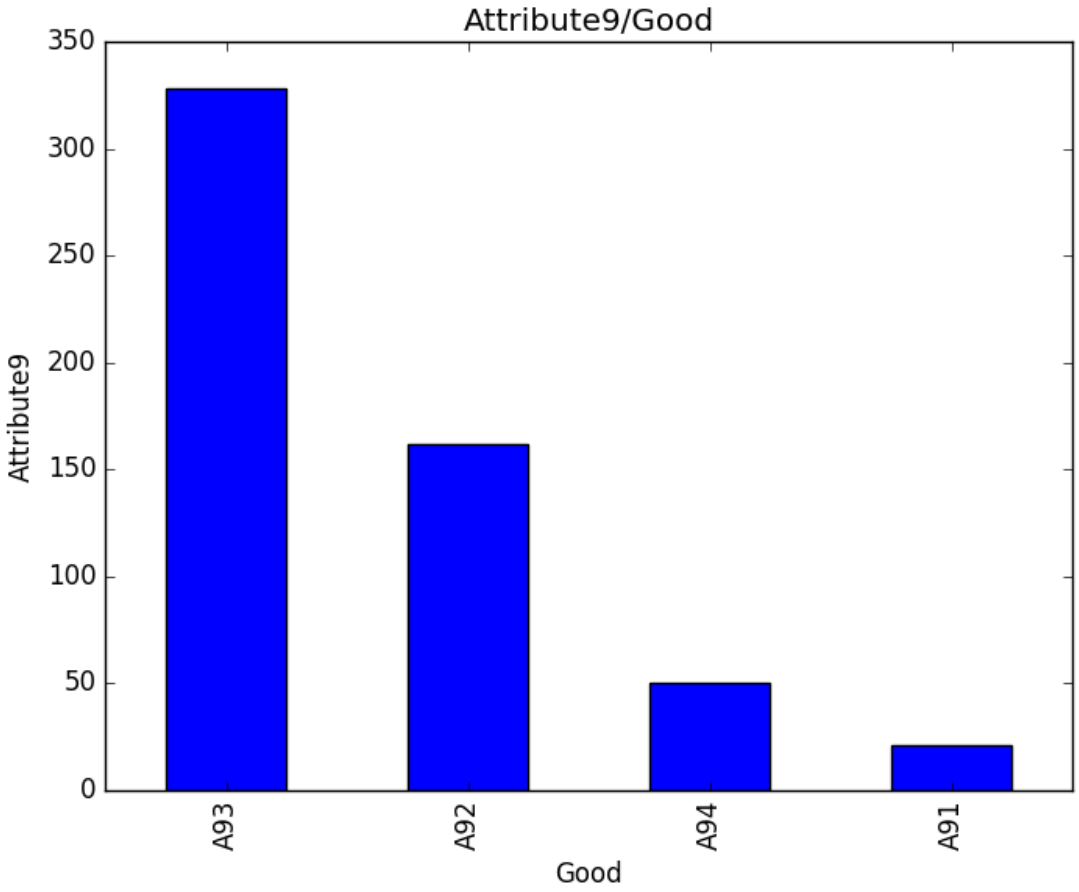

Attribute6/Bad
180
160
140
120
100
80
60
40
20
0
Attribute6
A61
A62
A65
A63
A64
Bad

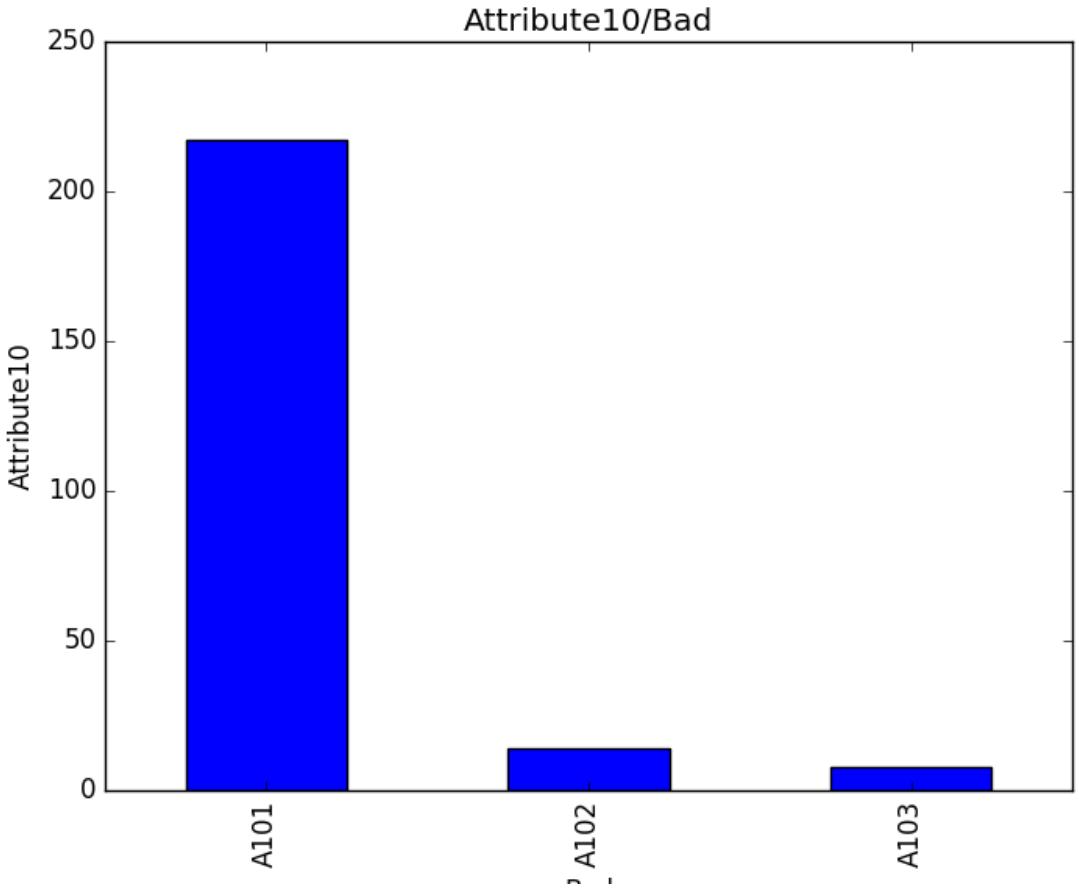

Attribute6/Good
350
300
250
200
150
100
50
0
Attribute6
A61
A65
A62
A63
A64
Good

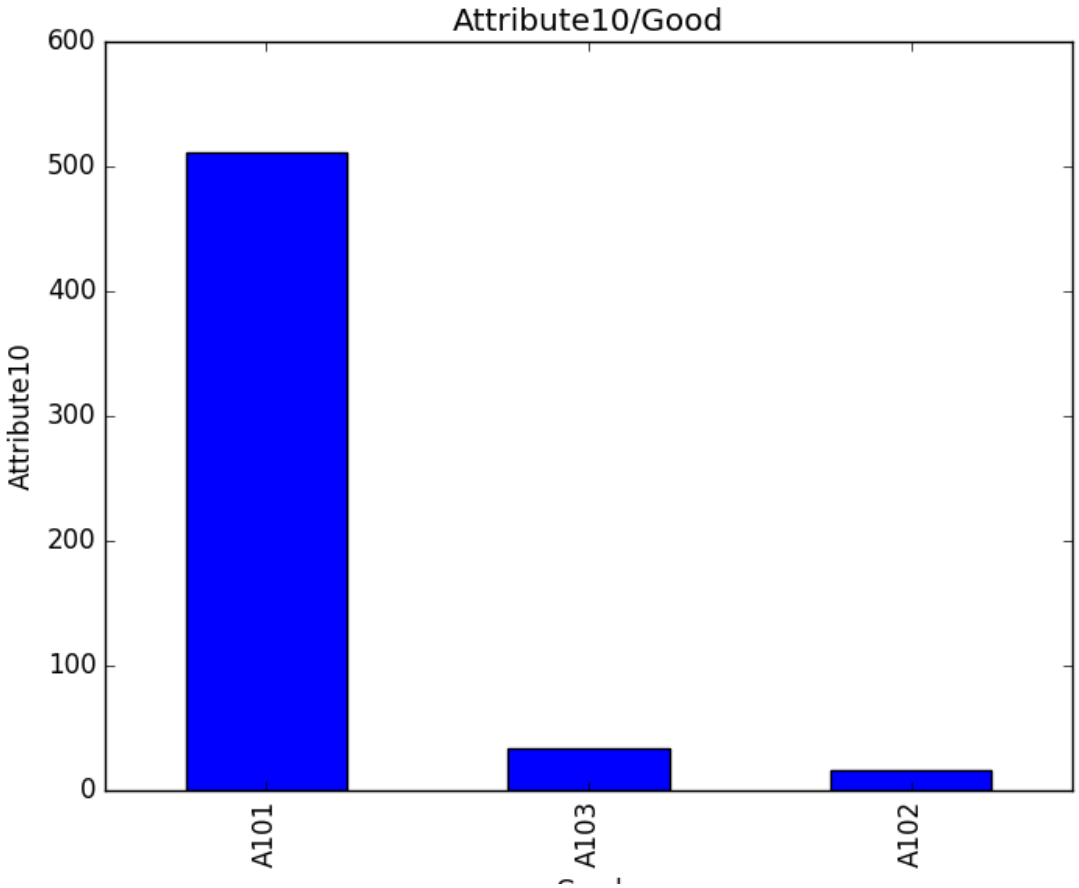

Attribute7/Bad
Attribute7
Bad
0
10
20
30
40
50
60
70
80
90
A73
A72
A75
A74
A71

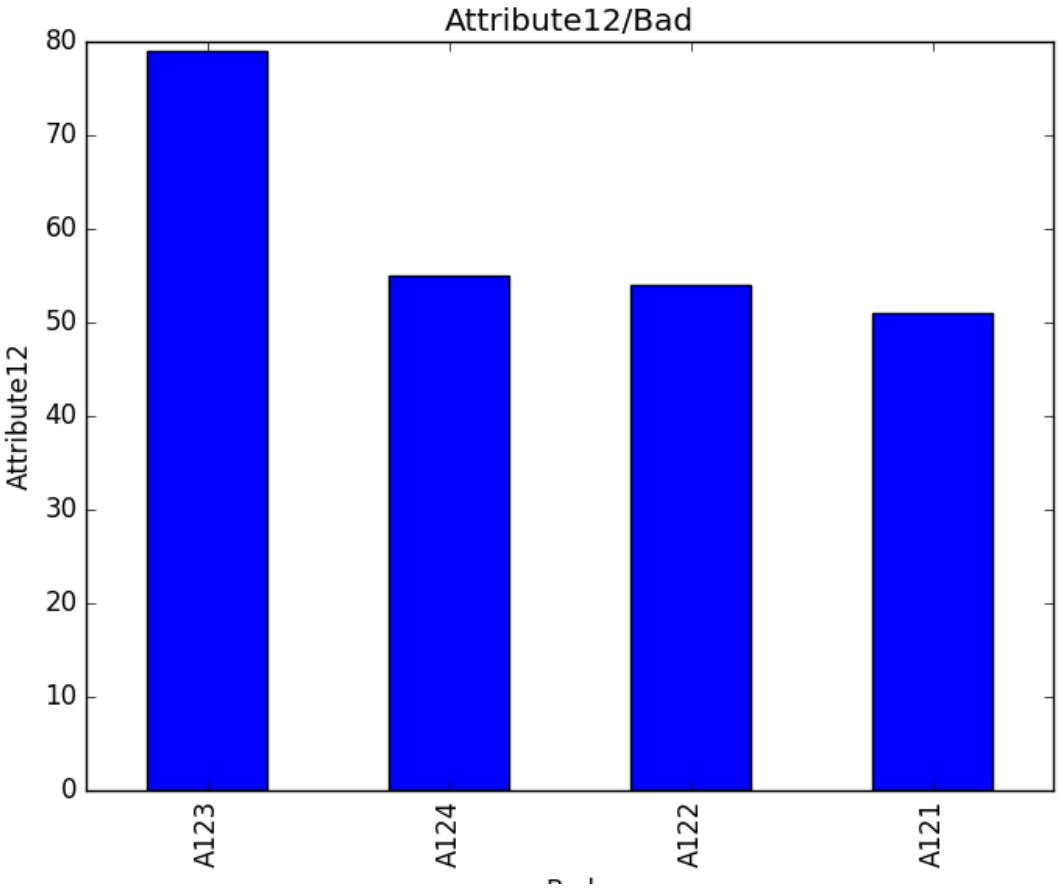

Attribute7/Good
200
150
100
50
0
Attribute7
A73
A75
A74
A72
A71
Good

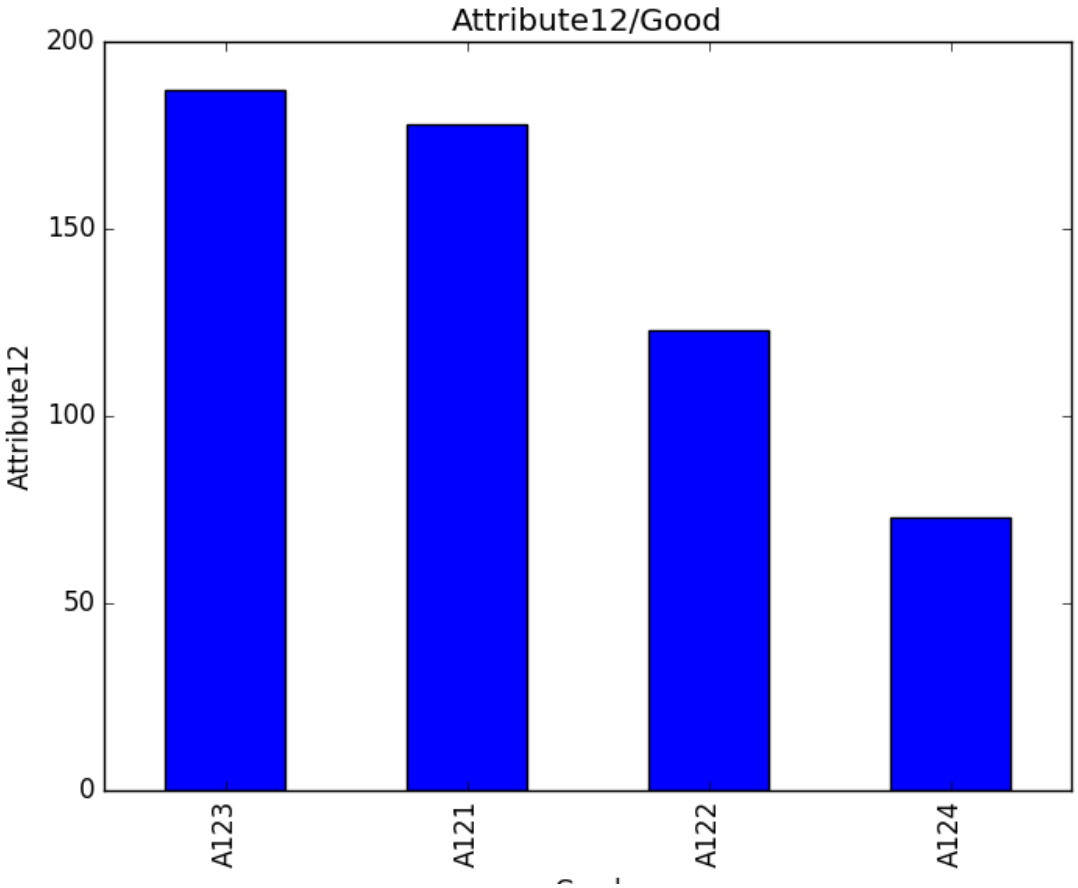

Attribute9/Bad
120
100
80
60
40
20
0
Attribute9
A93
A92
A94
A91
Bad

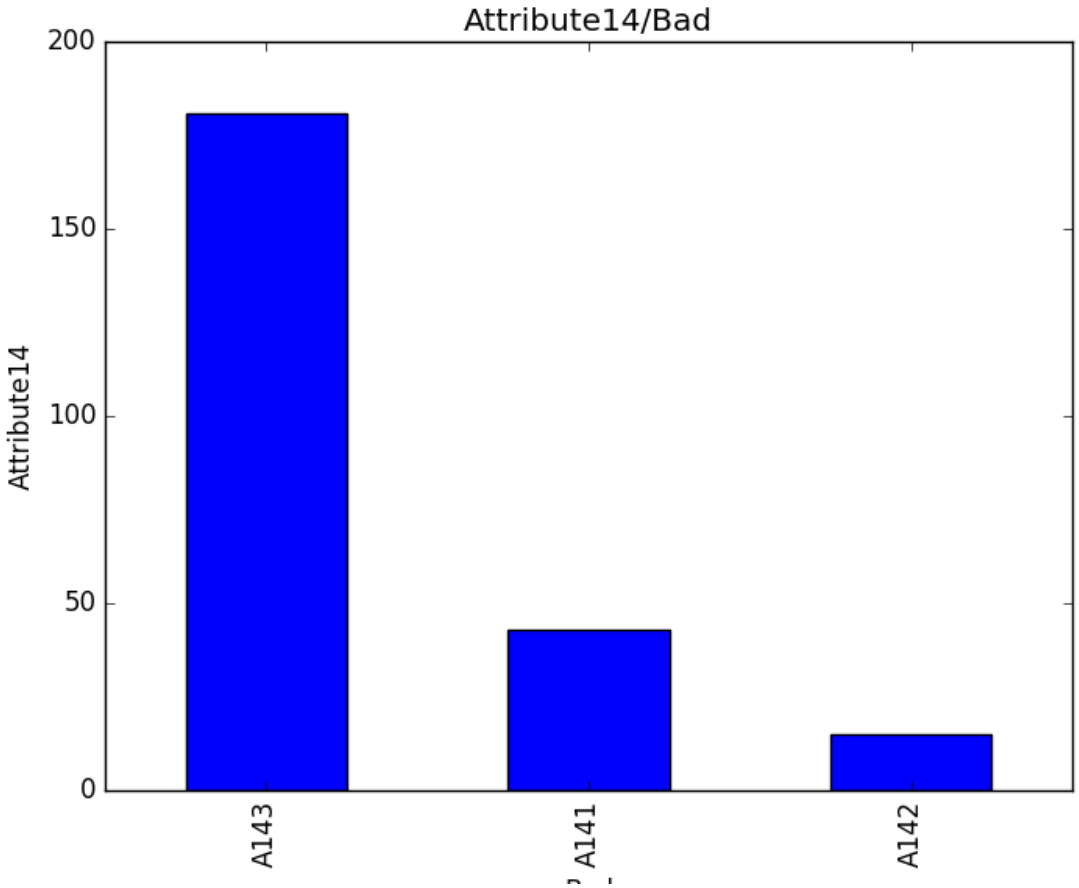

Attribute9/Good
350
300
250
200
150
100
50
0
Attribute9
A93
A92
A94
A91
Good

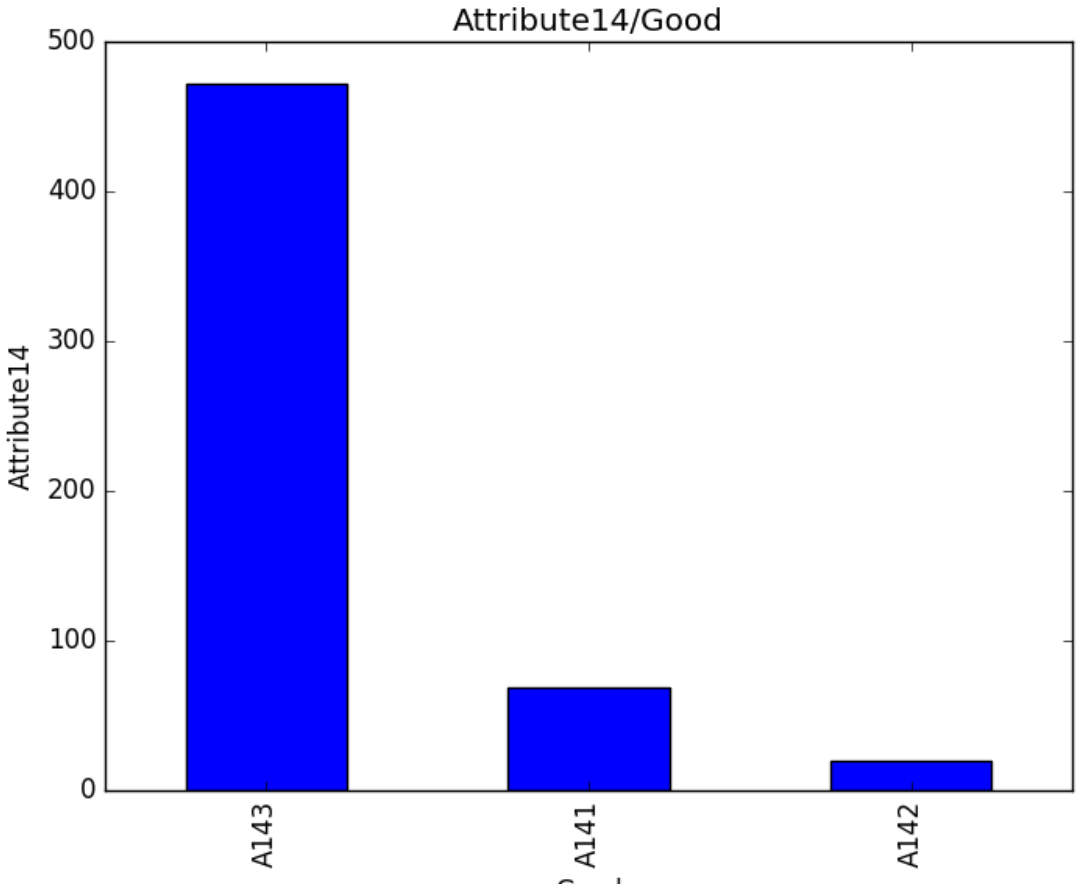

Attribute10/Bad
250
200
150
100
50
0
Attribute10
A101
A102
A103

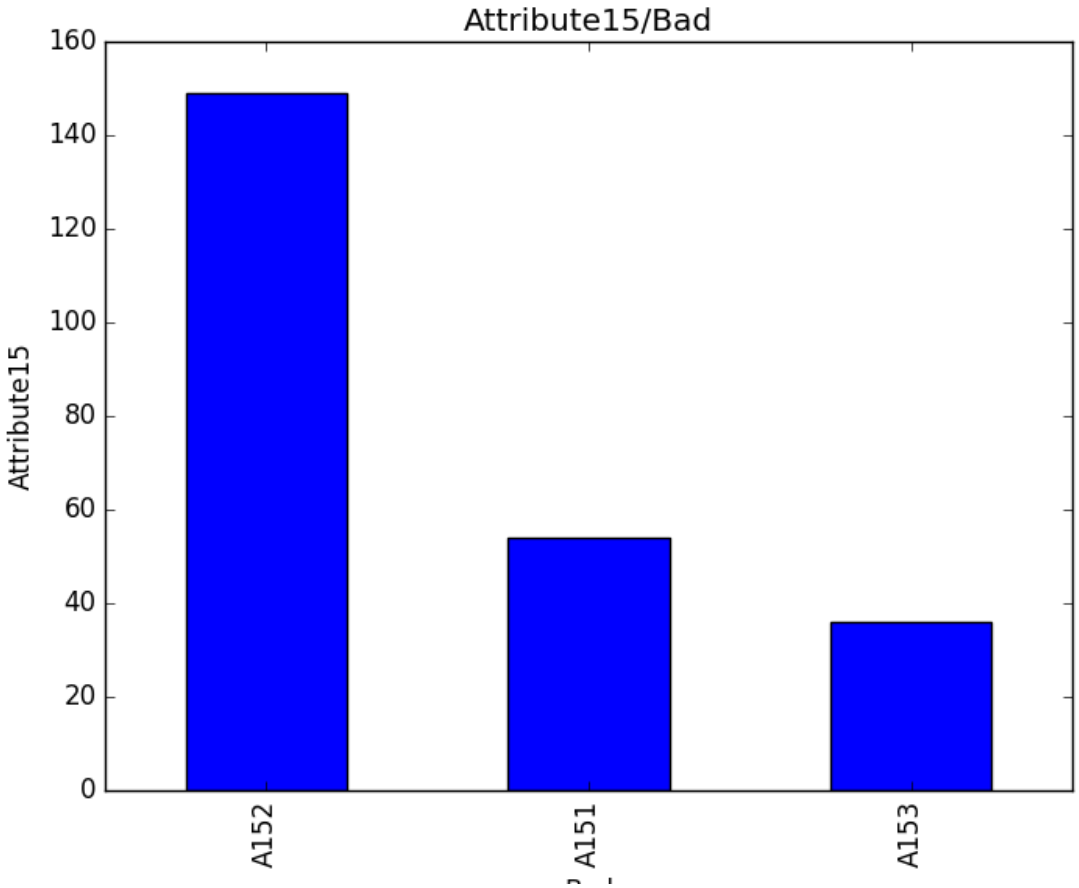

Attribute10/Good
600
500
400
300
200
100
0
Attribute10
A101
A103
A102

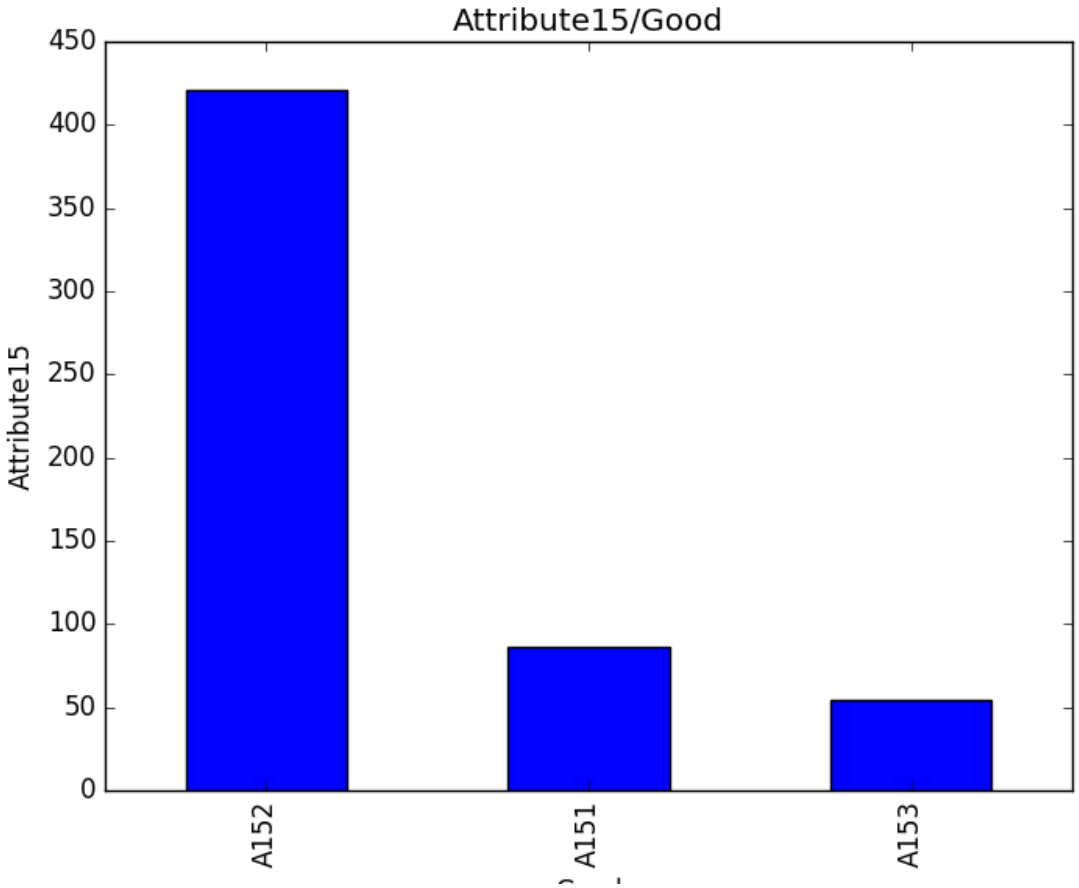

Attribute12/Bad
Attribute12
80
70
60
50
40
30
20
10
0
A123
A124
A122
A121

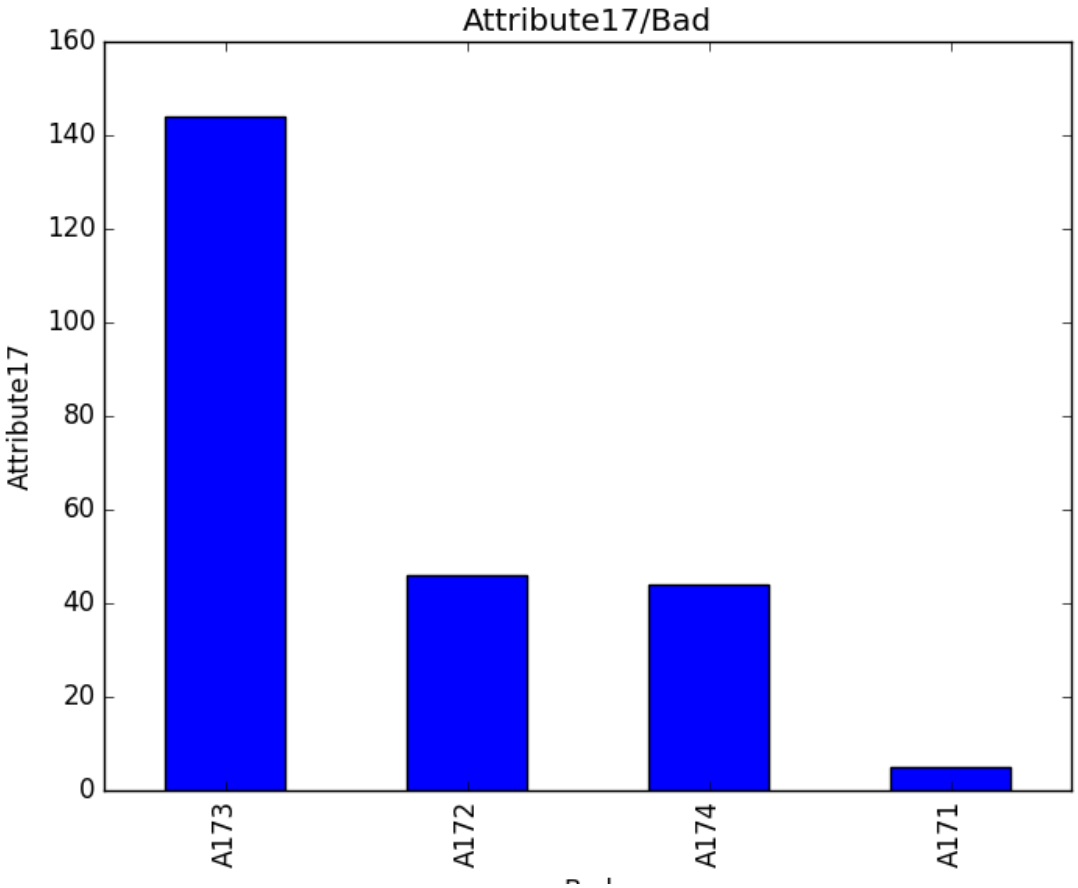

Attribute12/Good
200
150
100
50
0
Attribute12
A123
A121
A122
A124

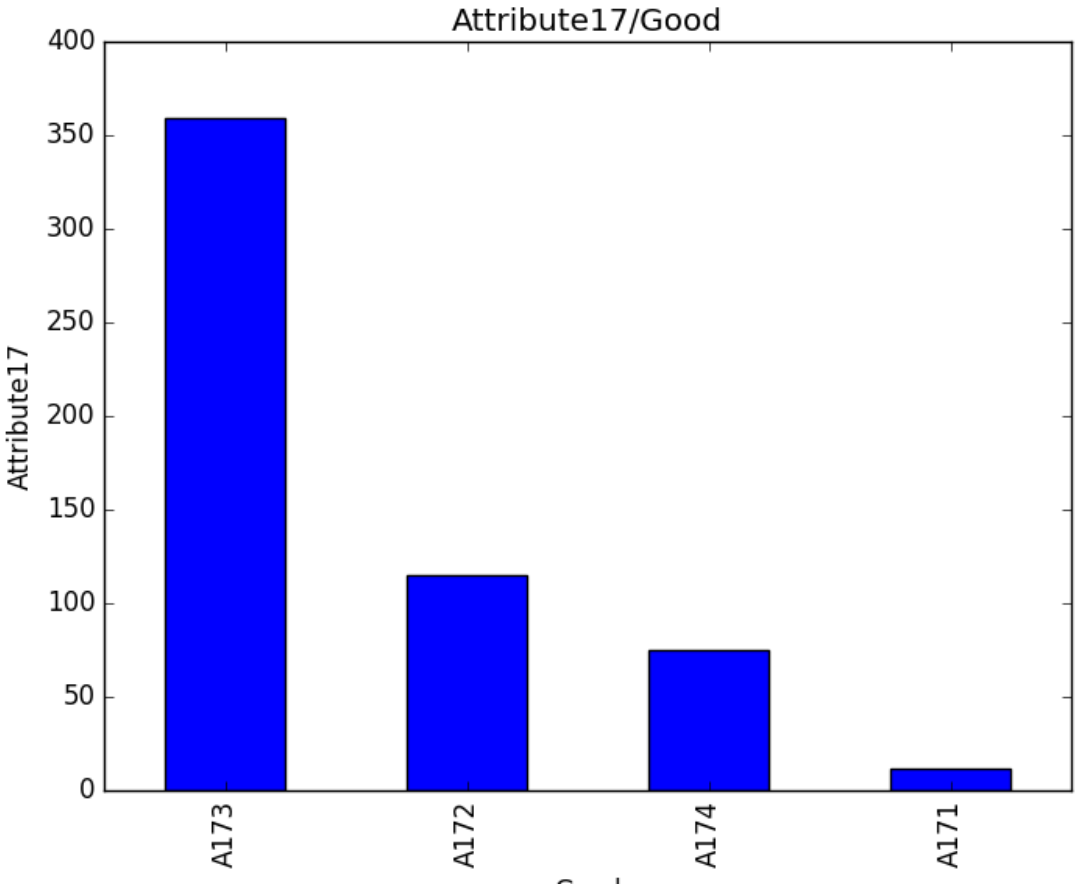

Attribute14/Bad
200
150
100
50
0
Attribute14
A143
A141
A142

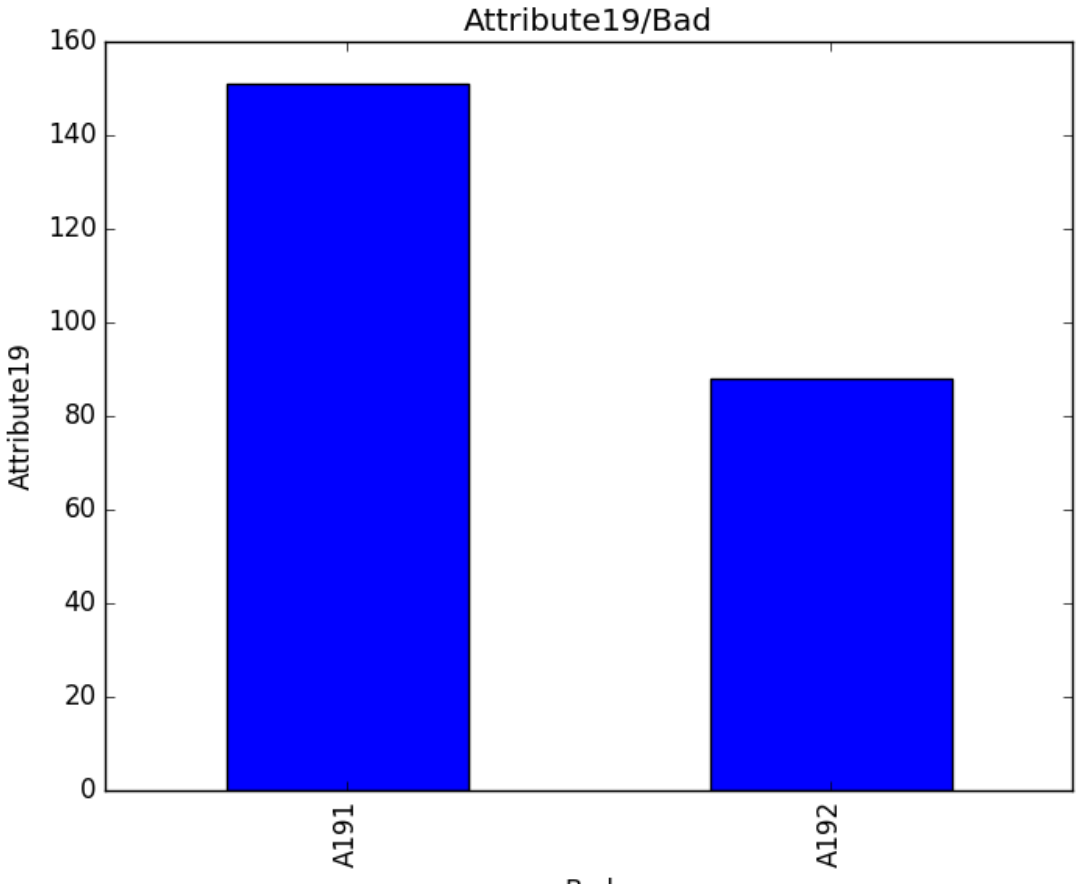

Attribute14/Good
500
400
300
200
100
0
Attribute14
A143
A141
A142

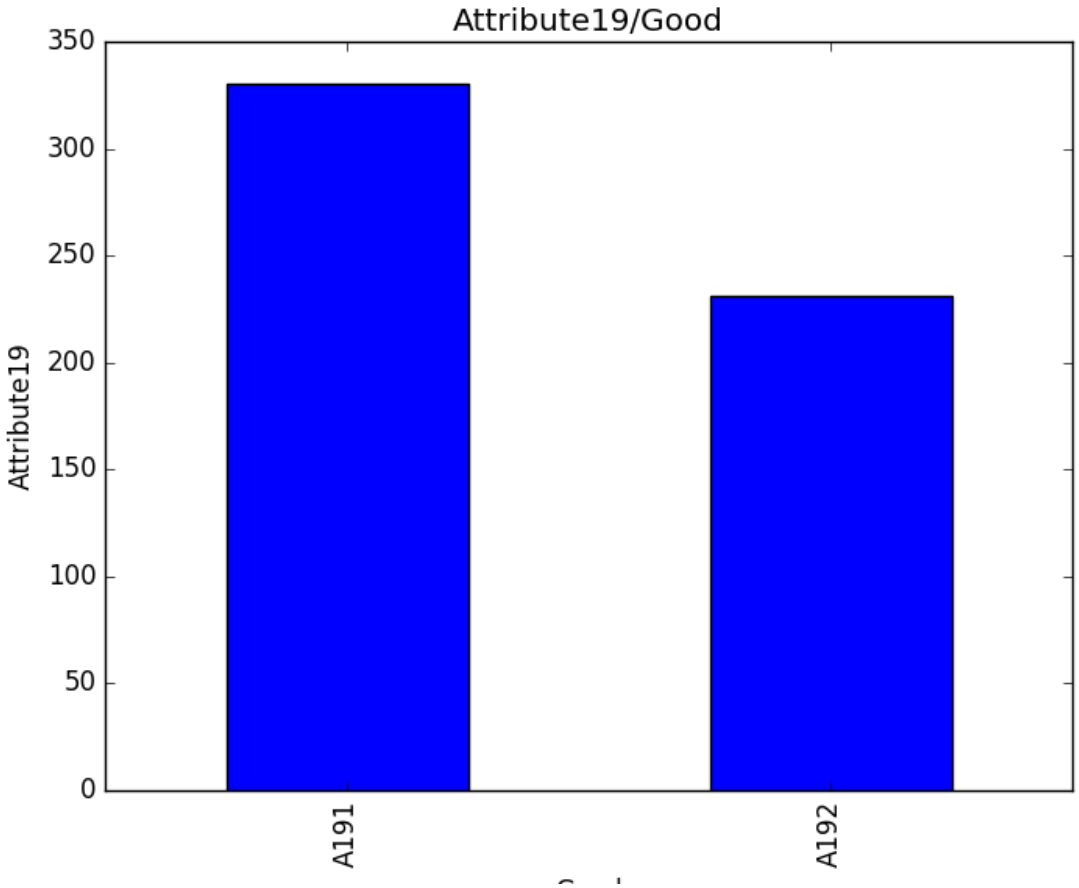

Attribute15/Bad
Attribute15
160
140
120
100
80
60
40
20
0
A152
A151
A153

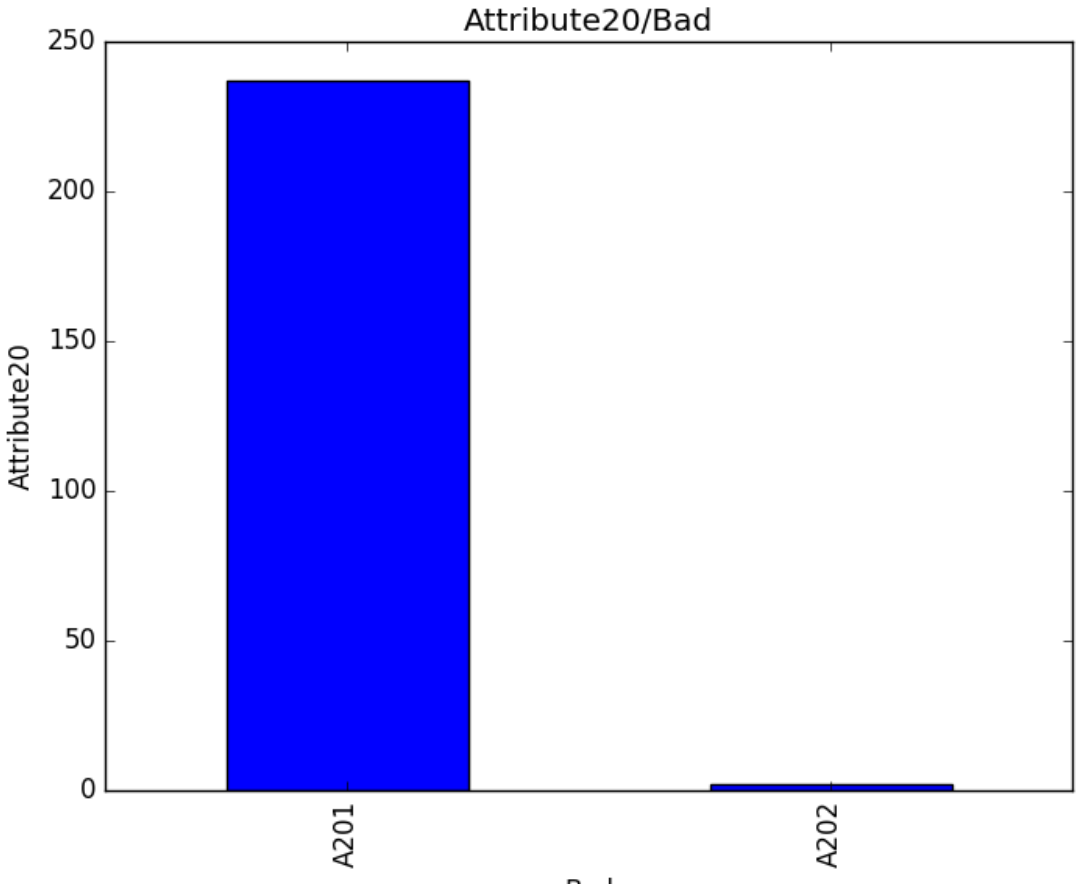

Attribute15/Good
Attribute15
450
400
350
300
250
200
150
100
50
0
A152
A151
A153

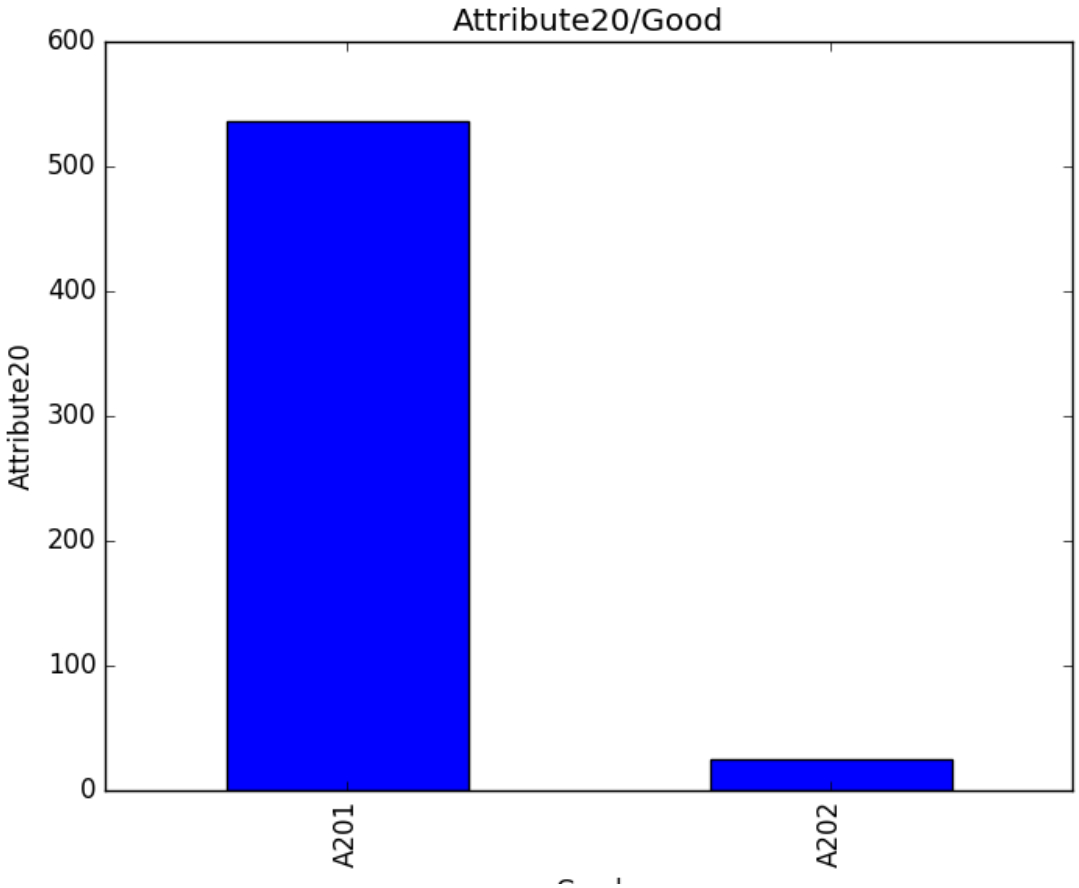

Attribute17/Bad
160
140
120
100
80
60
40
20
0
Attribute17
A173
A172
A174
A171

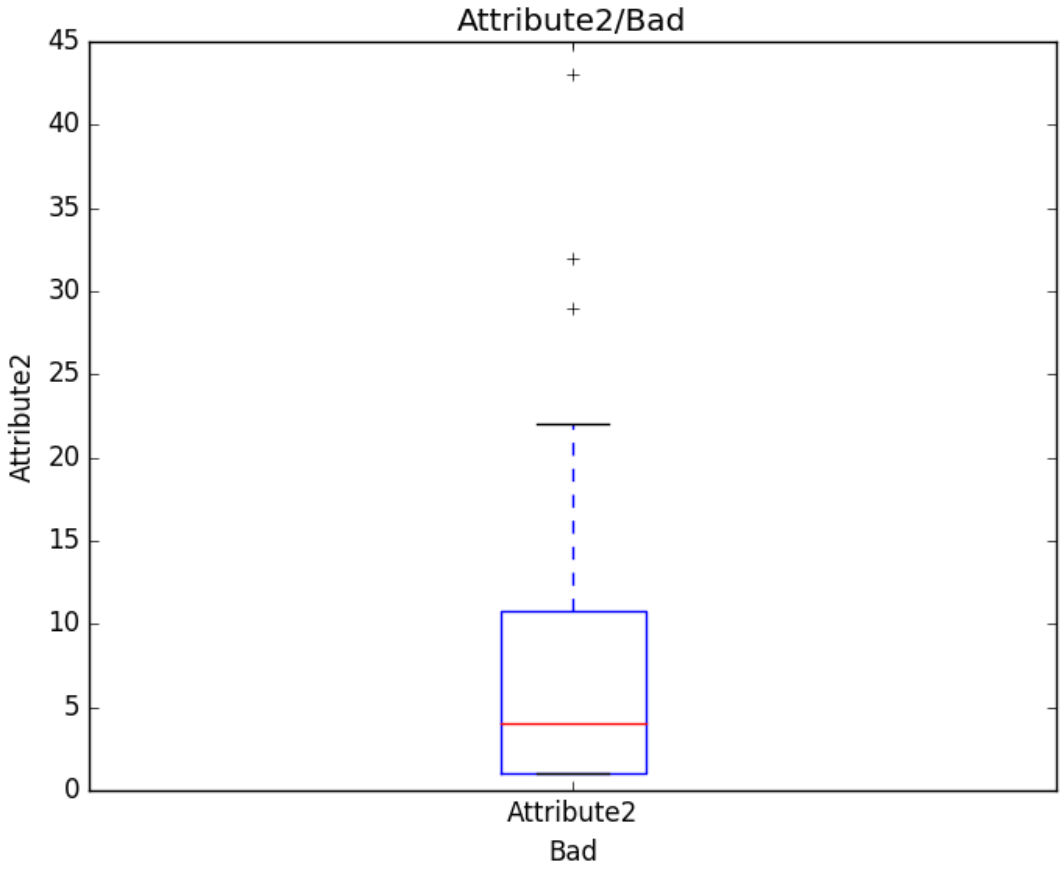

Attribute17/Good
Attribute17
400
350
300
250
200
150
100
50
0
A173
A172
A174
A171

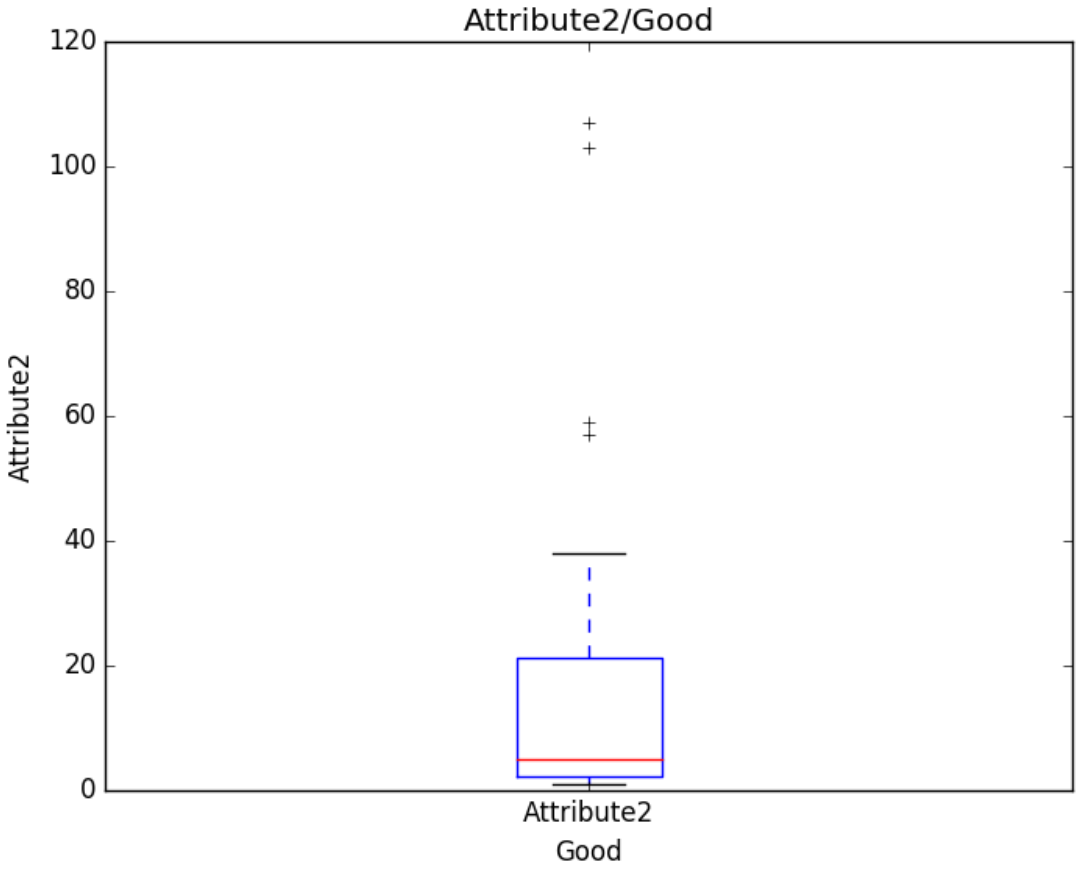

Attribute19/Bad
160
140
120
100
80
60
40
20
0
Attribute19
A191
A192

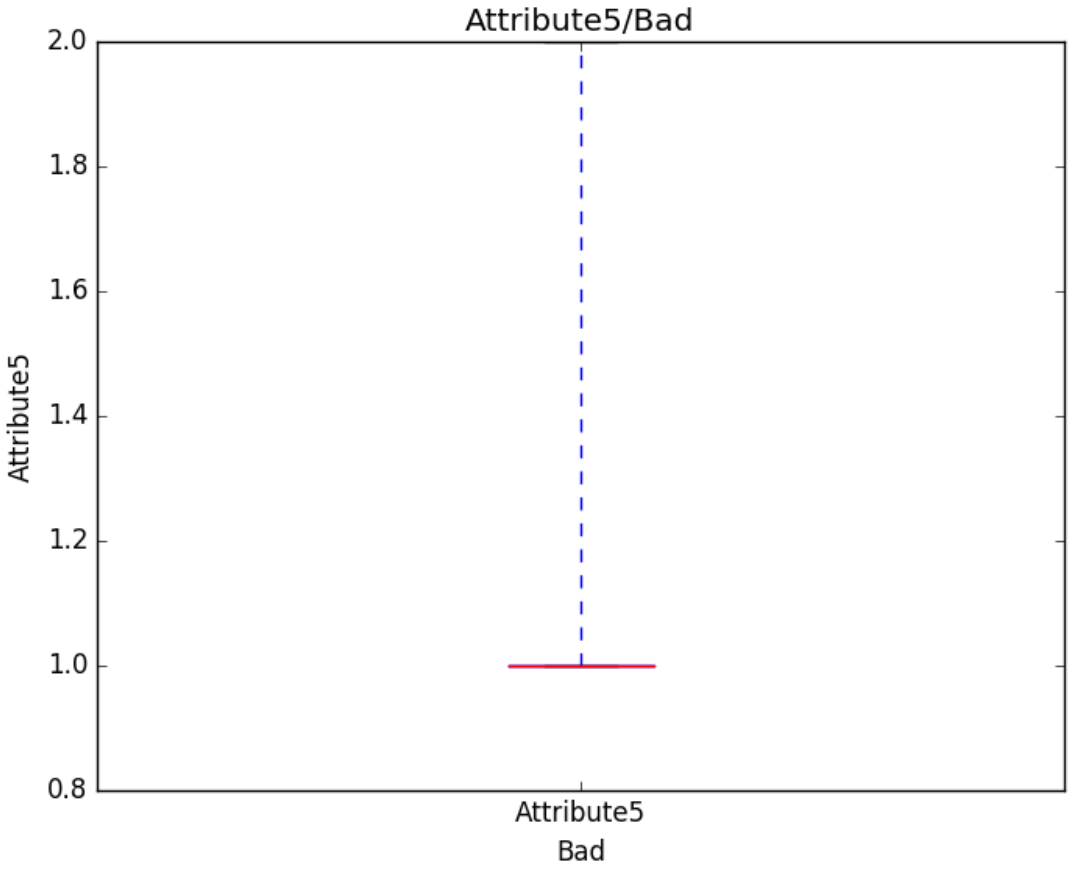

Attribute19/Good
350
300
250
200
150
100
50
0
Attribute19
A191
A192

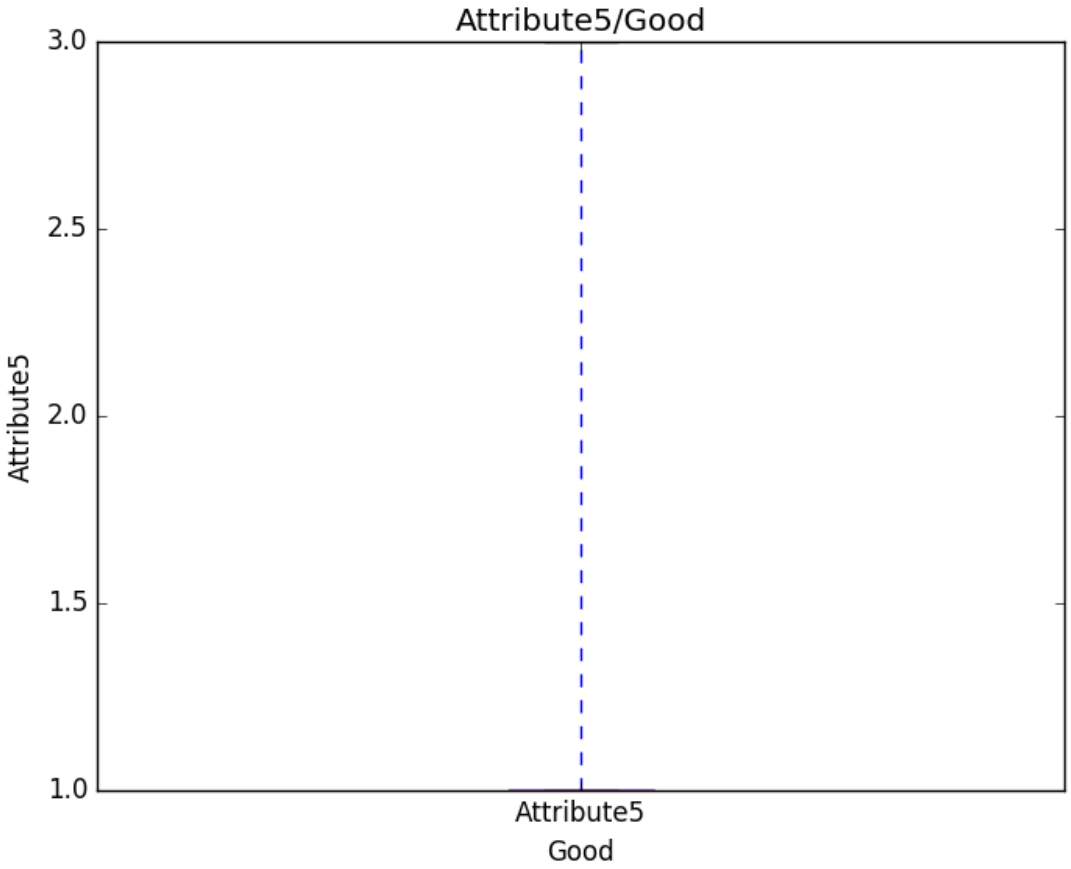

Attribute20/Bad
250
200
150
100
50
0
Attribute20
A201
A202

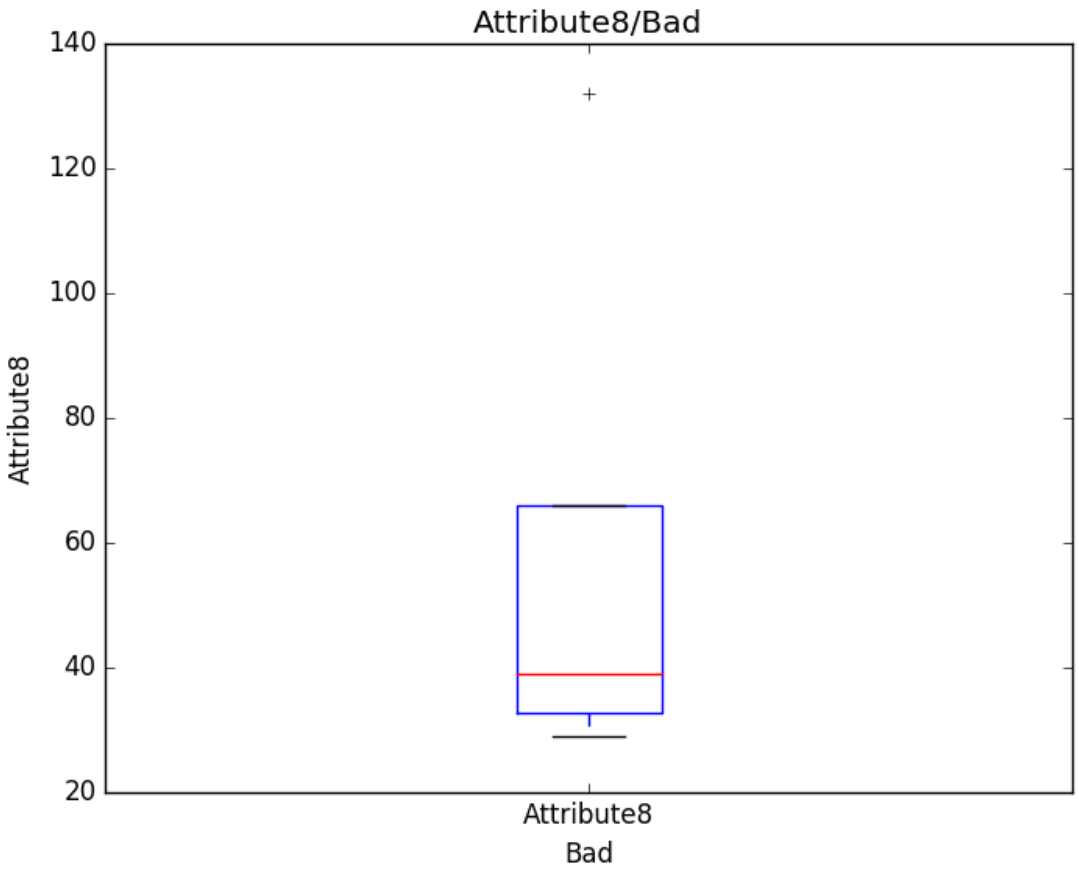

Attribute20/Good
600
500
400
300
200
100
0
Attribute20
A201
A202

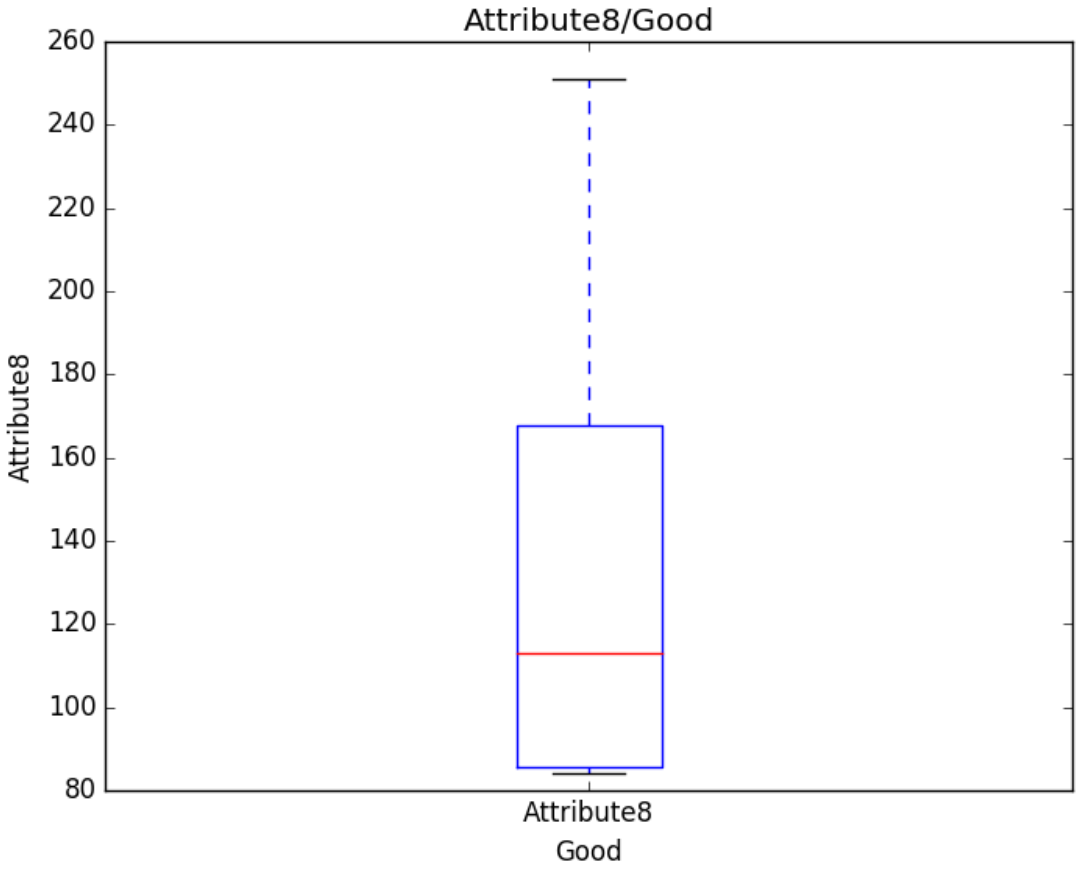

Attribute2/Bad
45
40
35
30
25
20
15
10
5
0
Attribute2
Attribute2
Bad

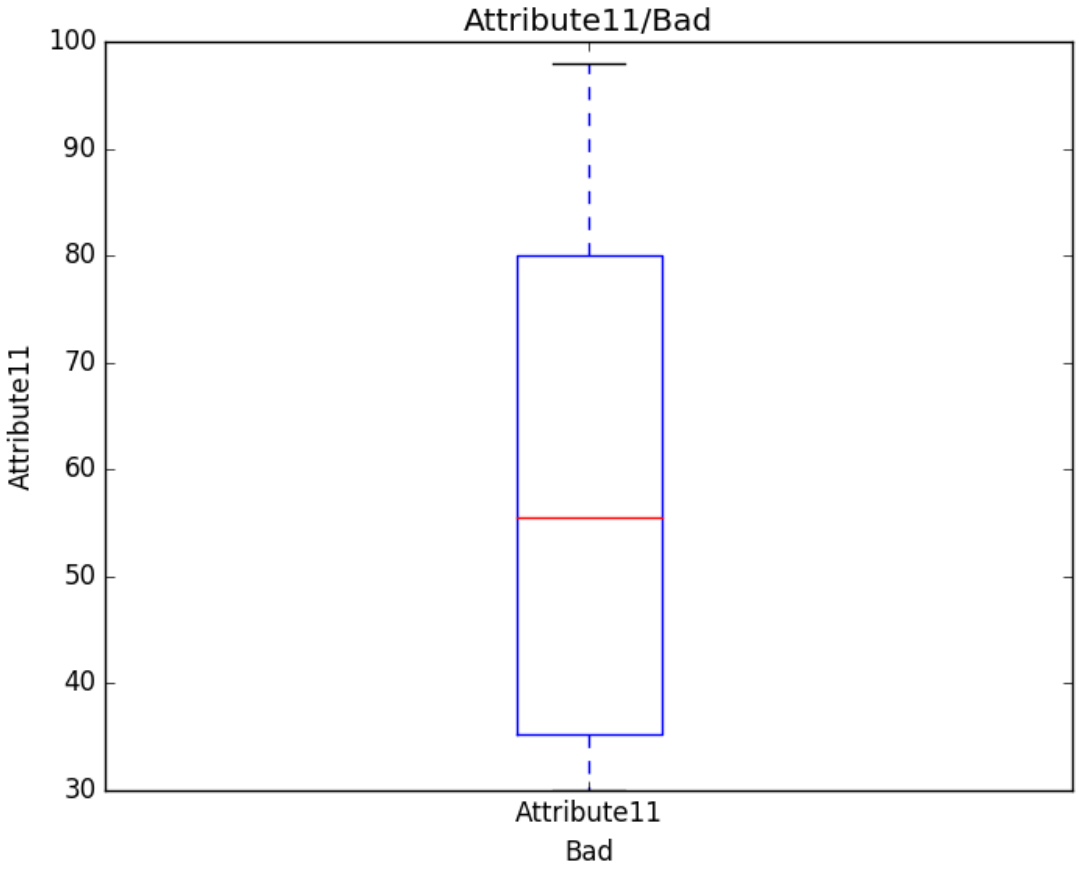

Attribute2/Good
120
100
80
60
40
20
0
Attribute2
Attribute2
Good

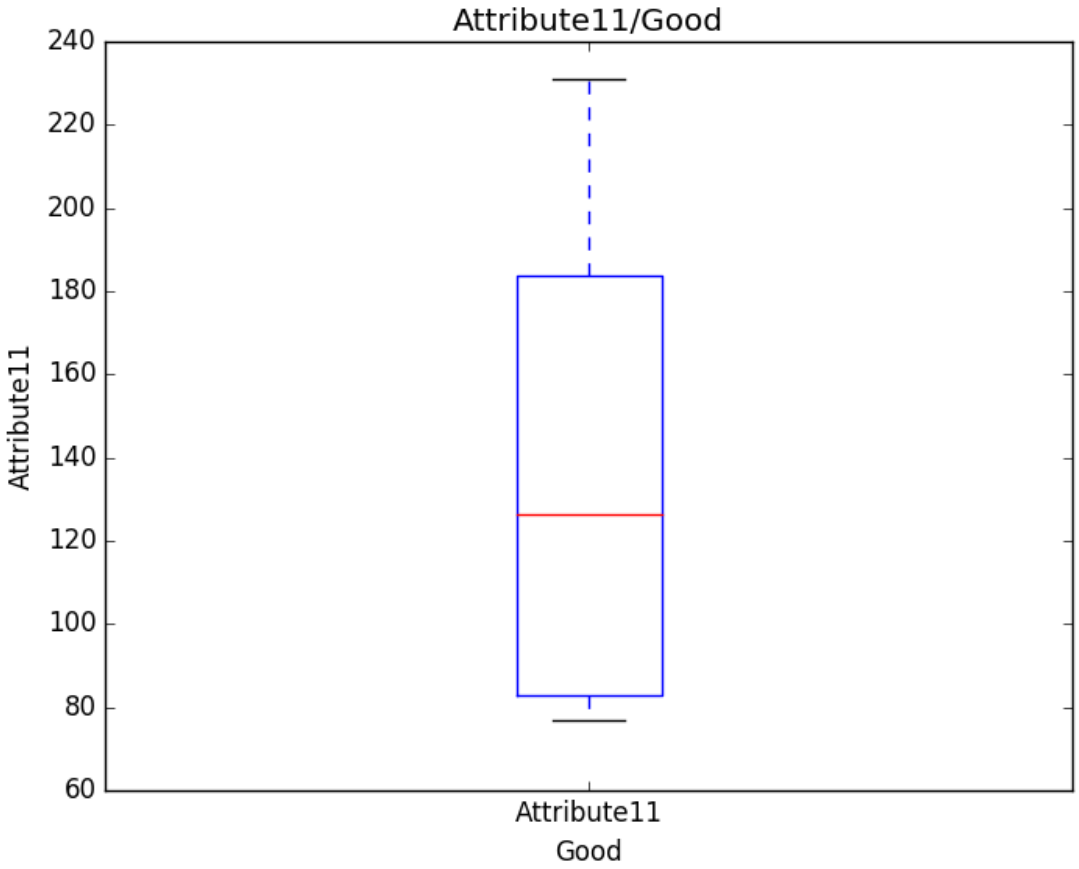

Attribute5/Bad
2.0
1.8
1.6
1.4
1.2
1.0
0.8
Attribute5
Attribute5
Bad

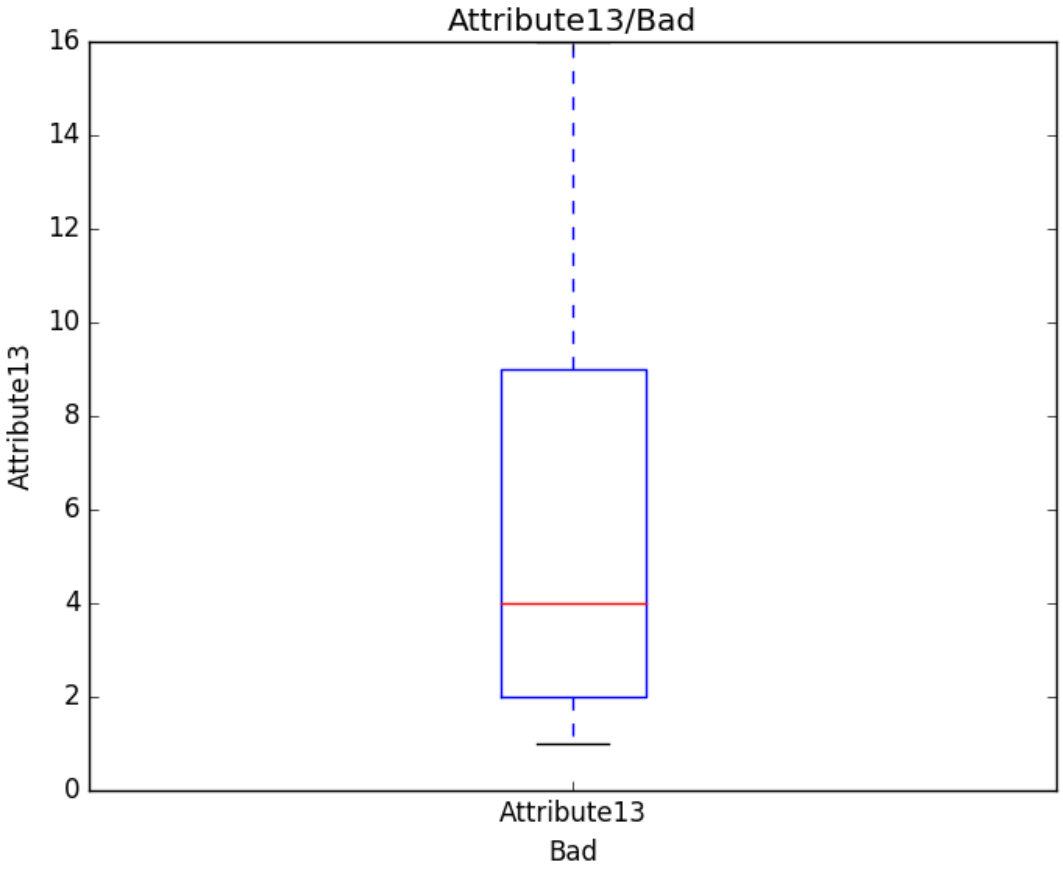

Attribute5/Good
3.0
2.5
2.0
1.5
1.0
Attribute5
Attribute5
Good

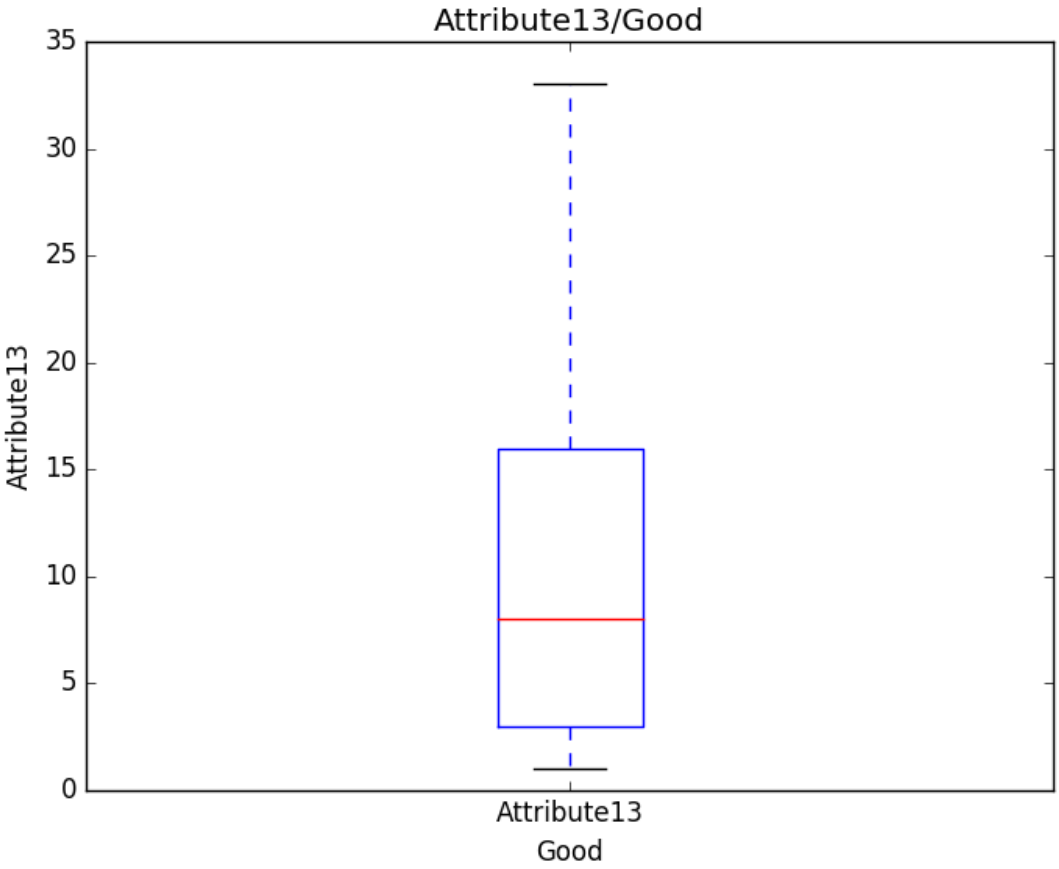

Attribute8/Bad
140
120
100
80
60
40
20
Attribute8
Attribute8
Bad

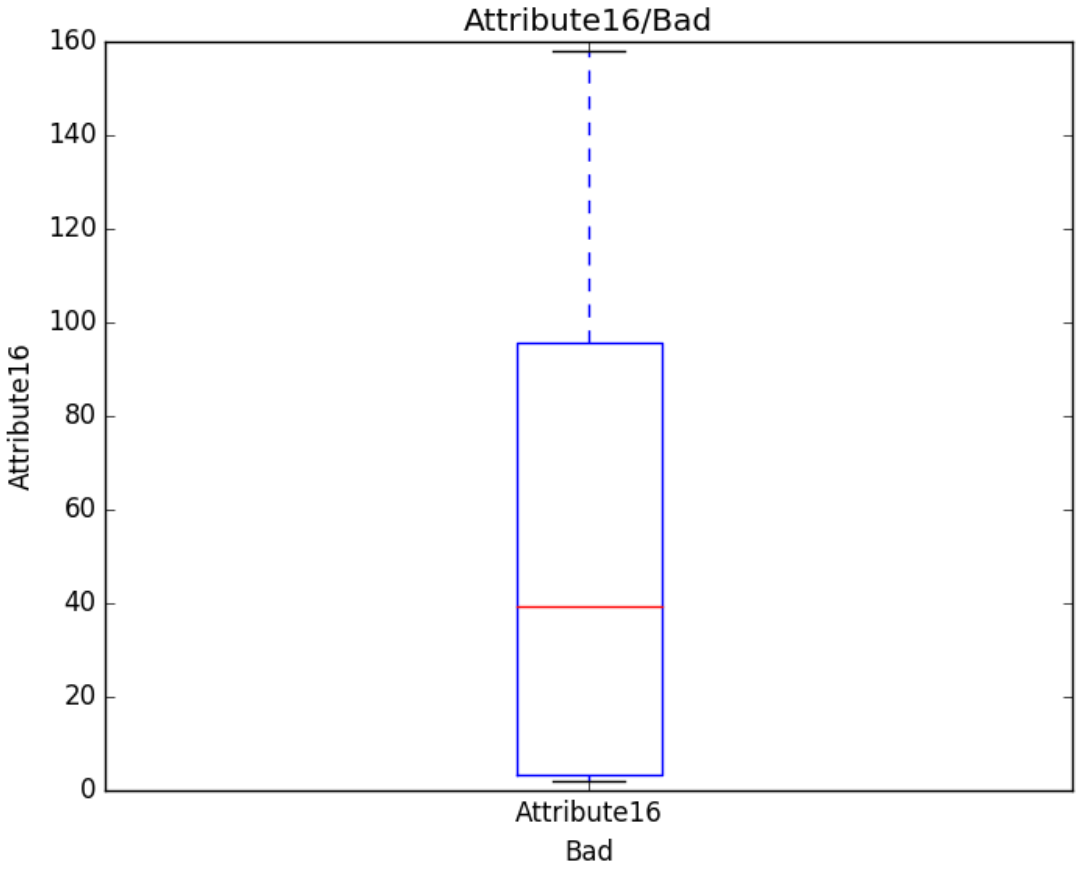

Attribute8/Good
260
240
220
200
180
160
140
120
100
80
Attribute8
Attribute8
Good

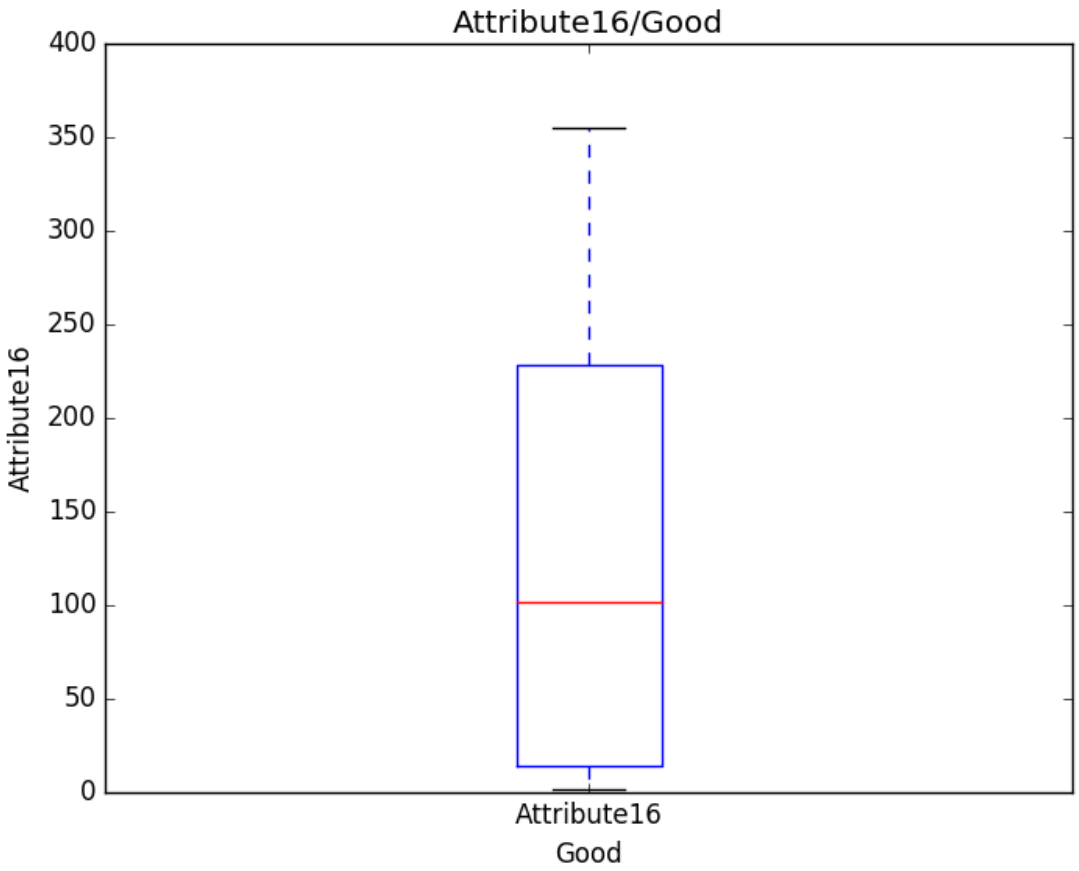

Attribute11/Bad
100
90
80
70
60
50
40
30
Attribute11
Attribute11
Bad


Attribute11/Good
240
220
200
180
160
140
120
100
80
60
Attribute11
Attribute11
Good


Attribute13/Bad
16
14
12
10
8
6
4
2
0
Attribute13
Attribute13
Bad


Attribute13/Good
35
30
25
20
15
10
5
0
Attribute13
Attribute13
Good

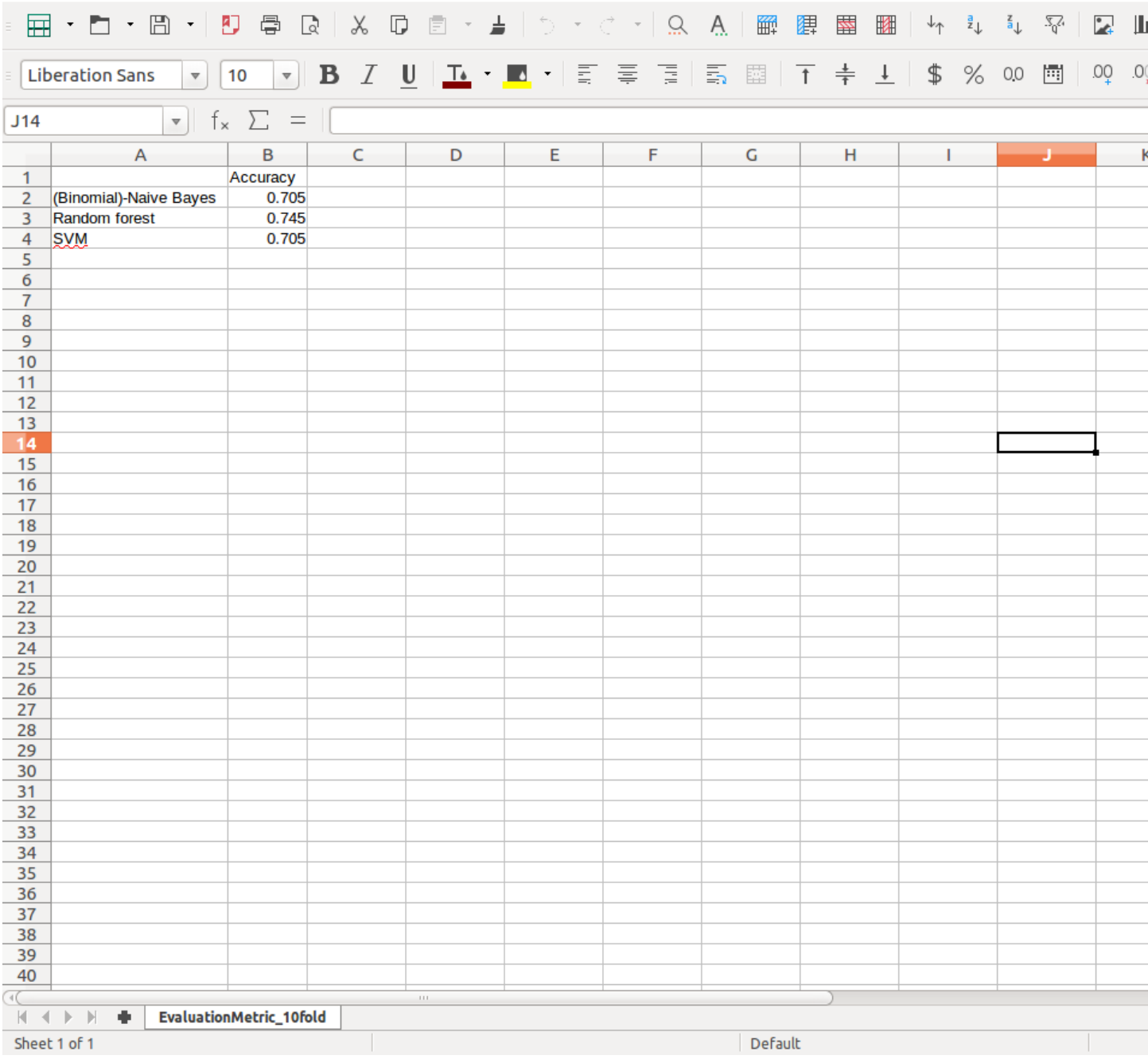

Attribute16/Bad
160
140
120
100
80
60
40
20
0
Attribute16
Attribute16
Bad



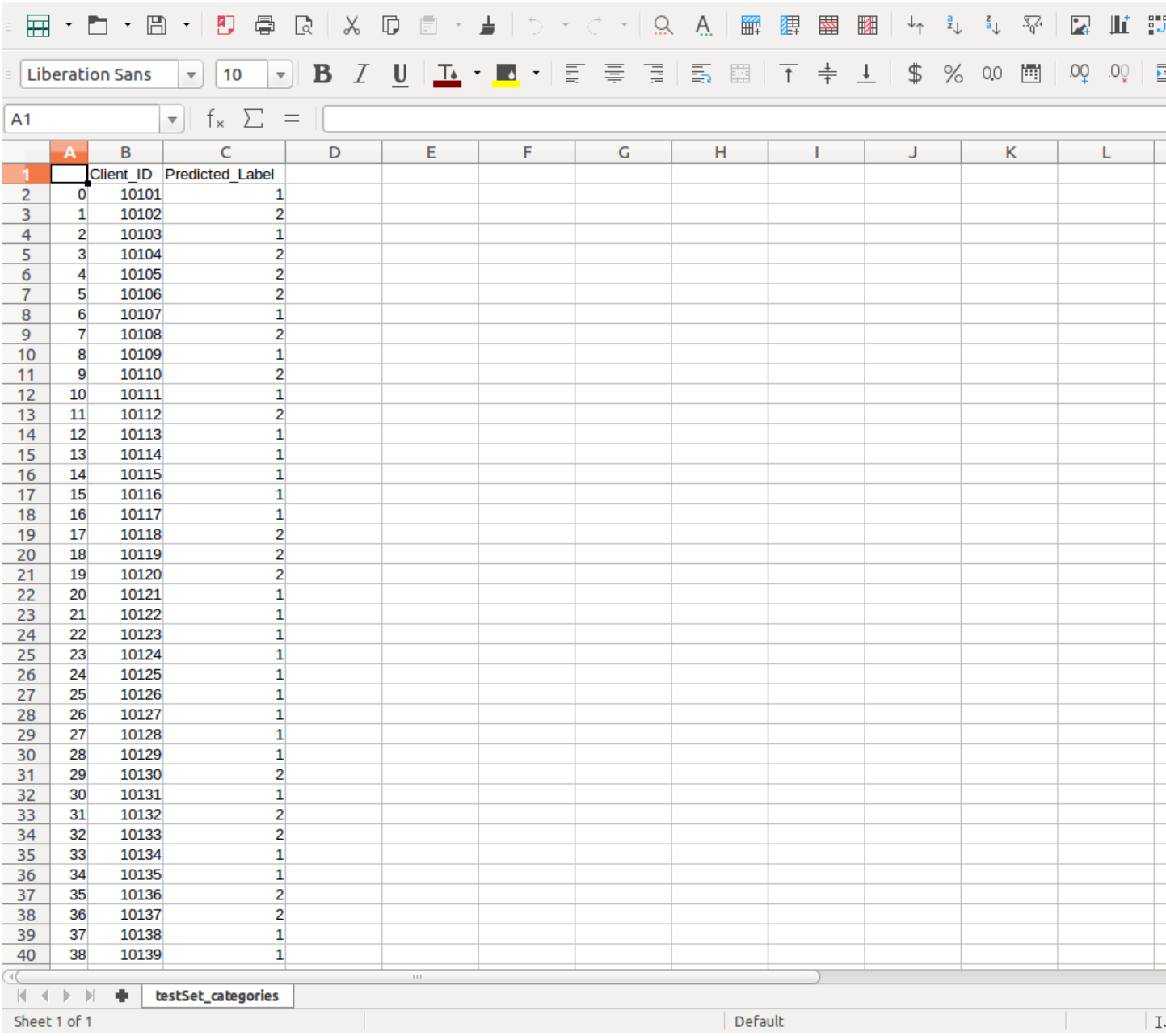
2.Υλοποίηση Κατηγοριοποίησης (Classification)

Σε αυτό το ερώτημα υλοποιήσαμε επίσης πλήρως τους αλγόριθμους κατηγοριοποίησης παρατηρώντας accuracy μετά από αρκετές εκτελέσεις περίπου 70-75% και για τους τρεις, με μεγαλύτερη αστάθεια στον random forest περίπου 5%.

Παραθέτεται screenshot του EvaluationMetric10fold(accuracy) και screenshot του test_setPredicitons(δεν αλλαξαμε το 1 σε good και το 2 σε bad):



	A	B
1		Accuracy
2	(Binomial)-Naive Bayes	0.705
3	Random forest	0.745
4	SVM	0.705



	A	B	C
1		Client_ID	Predicted_Label
2	0	10101	1
3	1	10102	2
4	2	10103	1
5	3	10104	2
6	4	10105	2
7	5	10106	2
8	6	10107	1
9	7	10108	2
10	8	10109	1
11	9	10110	2
12	10	10111	1
13	11	10112	2
14	12	10113	1
15	13	10114	1
16	14	10115	1
17	15	10116	1
18	16	10117	1
19	17	10118	2
20	18	10119	2
21	19	10120	2
22	20	10121	1
23	21	10122	1
24	22	10123	1
25	23	10124	1
26	24	10125	1
27	25	10126	1
28	26	10127	1
29	27	10128	1
30	28	10129	1
31	29	10130	2
32	30	10131	1
33	31	10132	2
34	32	10133	2
35	33	10134	1
36	34	10135	1
37	35	10136	2
38	36	10137	2
39	37	10138	1
40	38	10139	1

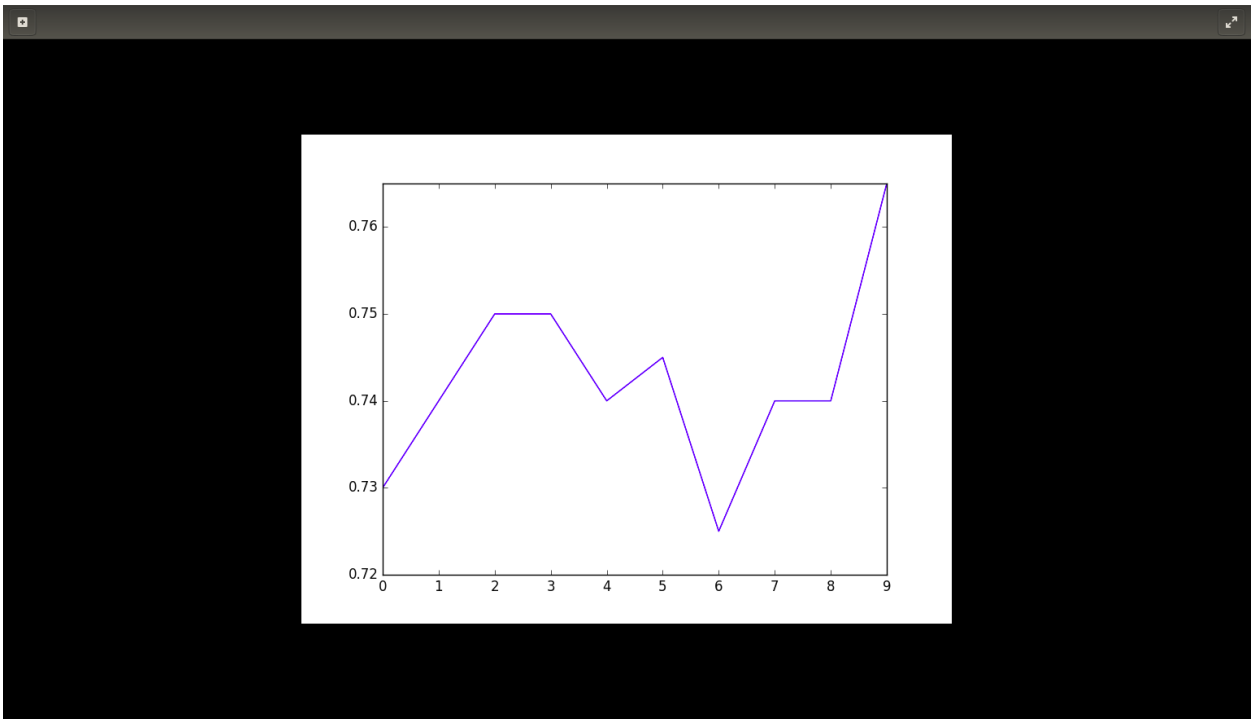
3.Επιλογή Features

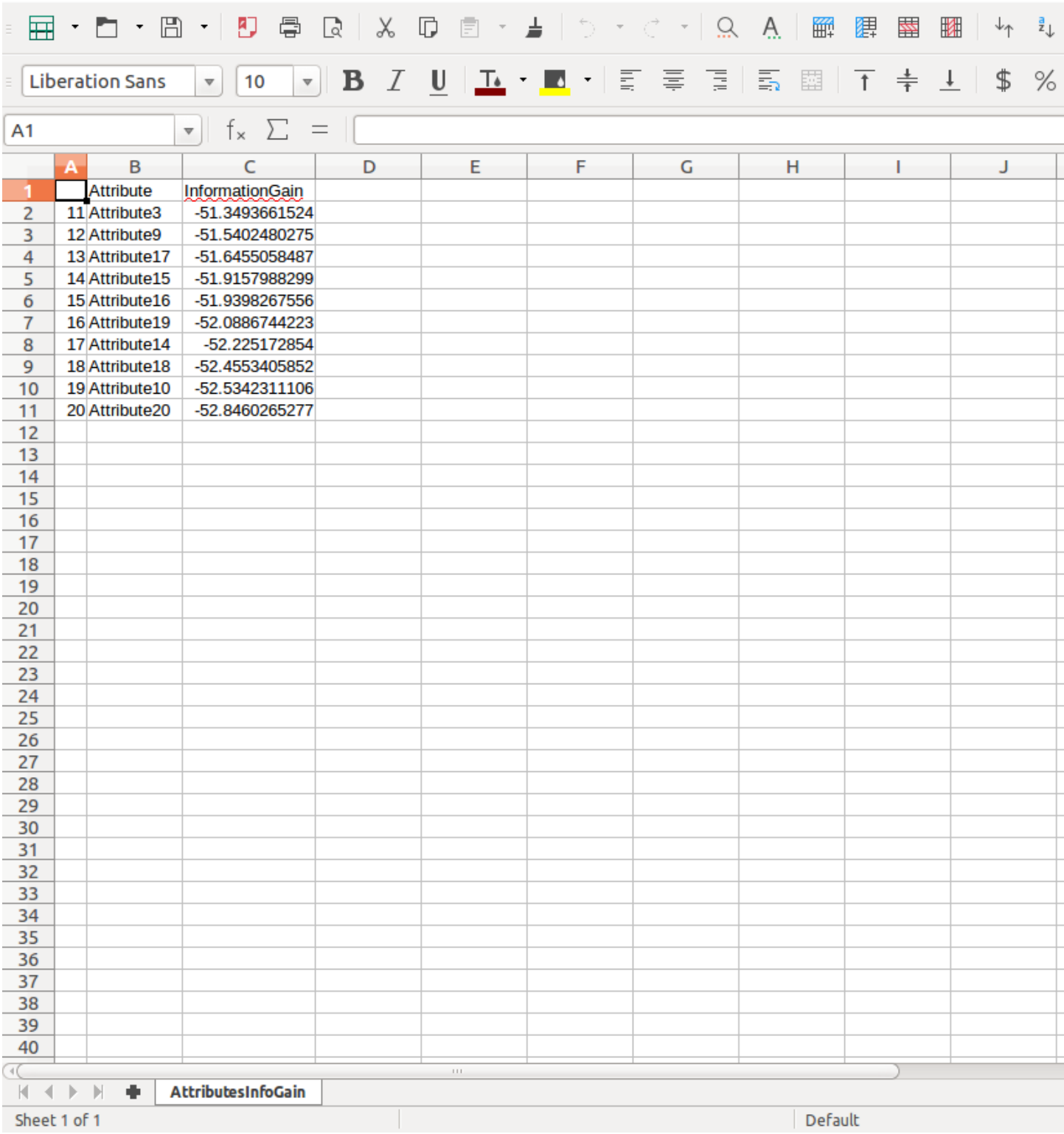
Σε αυτό το ερώτημα είχαμε το μεγαλύτερο πρόβλημα καθώς πολλά ζητούμενα ήταν ασαφή. Δημιουργήσαμε δικές μας συναρτήσεις για υπολογισμό του entropy ολόκληρου του dataset, υπολογισμό του entropy ενός μόνο Attribute και υπολογισμό του InformationGain ενός Attribute ως προς το dataset.Στη συνέχεια

αφαιρούσαμε το Attribute με το μικρότερο InfoGain και τρέχαμε ξανά τον RandomForest Classifier. Εδώ παρουσιάστηκε το πρόβλημα ότι μετά το $10^{ο}$ Attribute τερμάτιζε, οπότε έχουμε αποτελέσματα για τα 10 χειρότερα ως προς το InfoGain Attributes. Τα αποτελέσματά μας ωστόσο επιβεβαιώνουν τις εκτιμήσεις μας καθώς το accuracy εν τέλει ανεβαίνει.Κανονκά,αν βγάζαμε και τα υπόλοιπα 10 το accuracy θα έπρεπε να πέφτει καθώς μειώνονται τα Attributes και ουσιαστικά ο classifier γινεται τυχαίος. Σε κάποια σημεία κάνει καμπή, την οποία περιμέναμε, λόγω της αστάθειας του accuracy του αλγορίθμου.

Παραθέτονται το accuracy plot και τα Attributes/InfoGain που αφαιρέσαμε:

(το csv έβαζε το καινούργιο Attribute στην αρχή του αρχείου, οπότε τα στοιχεία είναι ανάποδα. Έχουμε και μια βοηθητική λίστα στην οποία τα στοιχεία έχουν μπει με την σωστή σειρά)





	A	B	C	D	E	F	G	H	I	J
1		Attribute	InformationGain							
2	11	Attribute3	-51.3493661524							
3	12	Attribute9	-51.5402480275							
4	13	Attribute17	-51.6455058487							
5	14	Attribute15	-51.9157988299							
6	15	Attribute16	-51.9398267556							
7	16	Attribute19	-52.0886744223							
8	17	Attribute14	-52.225172854							
9	18	Attribute18	-52.4553405852							
10	19	Attribute10	-52.5342311106							
11	20	Attribute20	-52.8460265277							