

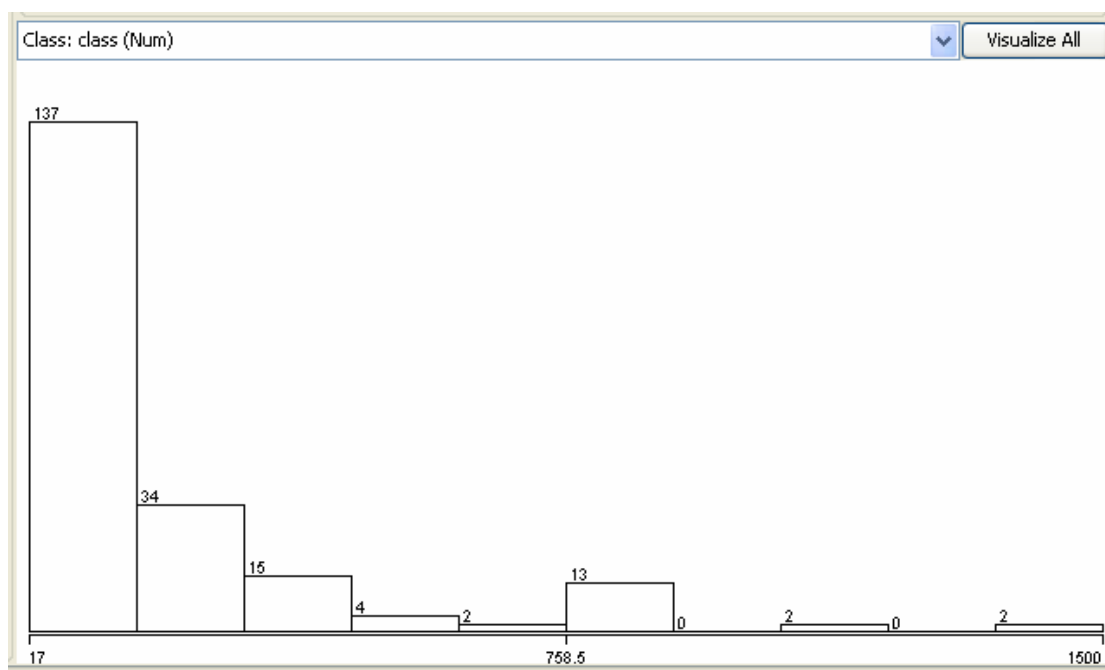
2ο μέρος εργασίας (Αρχείο cru)

Στο dataset cru, υπάρχουν 209 instances που αντιστοιχούν σε διαφορετικά configurations ενός υπολογιστή. Εξετάζεται το κατά πόσο επηρεάζεται η απόδοση του υπολογιστή από τις διαφορετικές τιμές των 6 διαφορετικών attributes-παραγόντων. Τα attributes αυτά, είναι: MYCT, MMIN, MMAX, CACH, CHMIN, CHMAX και όλα είναι numeric. Το ίδιο ισχύει και για το class-attribute Performance. Σκοπός της εργασίας είναι μέσω απλής οπτικής διερεύνησης των δεδομένων, να εξαχθούν κάποια συμπεράσματα όσον αφορά το συγκεκριμένο dataset. Έχουμε λοιπόν τα εξής:

A) MYCT

Selected attribute	
Name: MYCT	Type: Numeric
Missing: 0 (0%)	Distinct: 60
	Unique: 19 (9%)
Statistic	Value
Minimum	17
Maximum	1500
Mean	203.823
StdDev	260.263

Όσον αφορά το πρώτο attribute, βλέπουμε από τον πίνακα, ότι δεν υπάρχουν missing values, ότι έχει 60 distinct values (δηλαδή 60 διαφορετικές τιμές του attribute), και 19 unique values. Επίσης αναφέρονται και κάποια στατιστικά, συγκεκριμένα η μέση τιμή (mean), η τυπική απόκλιση (StdDev), η ελάχιστη και η μέγιστη τιμή (Minimum, Maximum).

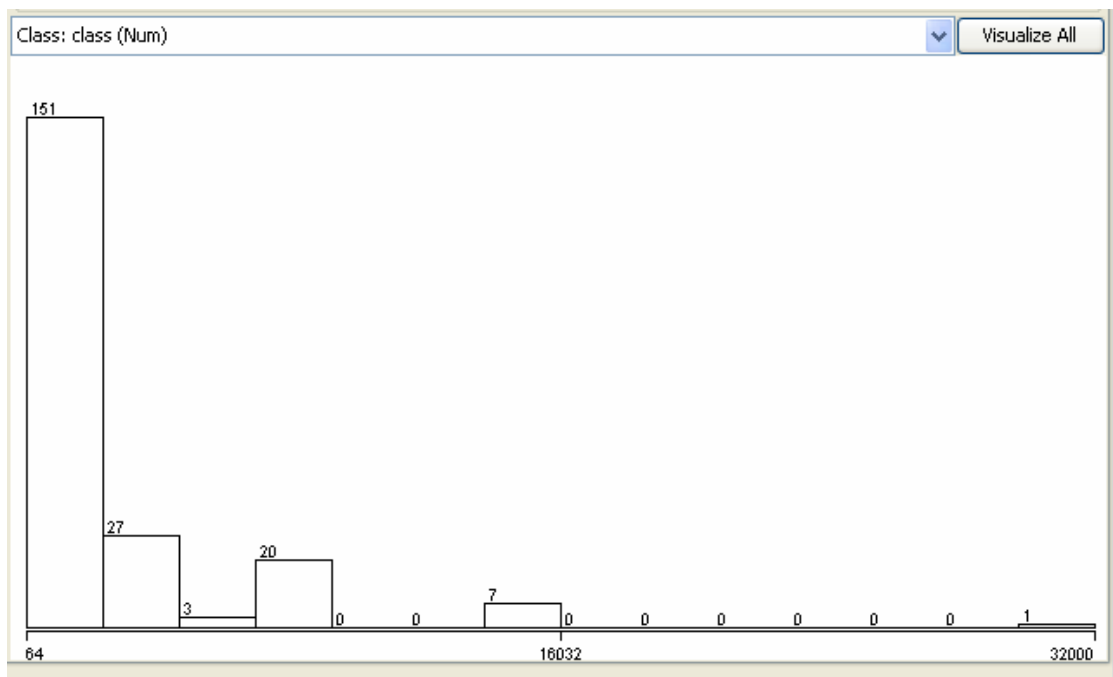


Επιπλέον, από το παραπάνω ιστόγραμμα, βλέπουμε πώς κατανέμεται το class-attribute Performance συναρτήσει των διαφορετικών τιμών του MYCT. Παρατηρούμε ότι οι περισσότερες τιμές (137) ανήκουν στο διάστημα 17 έως 165,3, ενώ 34 ανήκουν στο διάστημα 165,3 έως 313,6. Όμοια συμπεράσματα προκύπτουν για όλο το εύρος των τιμών, ενώ ιδιαίτερο ενδιαφέρον παρουσιάζουν τα διαστήματα όπου υπάρχουν μόνο δύο καταγραφές, όπως στο διάστημα 1351,7 μέχρι 1500, διότι οι τιμές αυτές μπορούν να θεωρηθούν ως outliers, σε περίπτωση που γίνει περαιτέρω διερεύνηση και ανάλυση του dataset.

B)MMIN

Selected attribute	
Name: MMIN	Type: Numeric
Missing: 0 (0%)	Distinct: 25
	Unique: 11 (5%)
Statistic	Value
Minimum	64
Maximum	32000
Mean	2867.981
StdDev	3878.743

Όσον αφορά το δεύτερο attribute, βλέπουμε από τον αντίστοιχο πίνακα, ότι δεν υπάρχουν missing values, ότι έχει 25 distinct values και 11 unique values. Επίσης αναφέρονται τα στατιστικά: μέση τιμή (mean), τυπική απόκλιση (StdDev), ελάχιστη και μέγιστη τιμή (Minimum, Maximum).



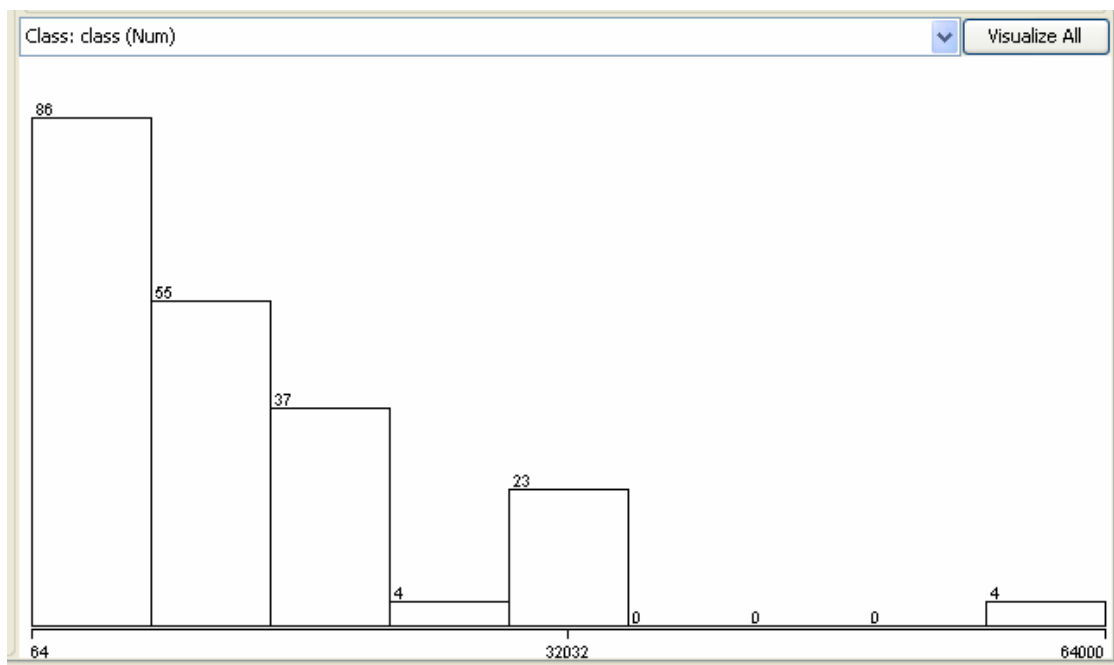
Από το αντίστοιχο ιστόγραμμα, βλέπουμε πώς κατανέμεται το class-attribute Performance συναρτήσει των διαφορετικών τιμών του MMIN. Βλέπουμε ότι η

πλειοψηφία των τιμών(151) ανήκει στο διάστημα 64 έως 2345,143, ενώ 27 τιμές ανήκουν στο διάστημα 2345,143 μέχρι 4626,286. Στο διάστημα 16032 έως 29718,857 δεν παρατηρούνται καταγραφές, ενώ στο διάστημα 29718,857 έως 32000 παρατηρείται μόνο μια καταγραφή η οποία πιθανόν να μπορεί να θεωρηθεί ως outlier σε μια δεύτερη ανάλυση του προβλήματος.

C) MMAX

Selected attribute	
Name: MMAX	Type: Numeric
Missing: 0 (0%)	Distinct: 23
	Unique: 6 (3%)
Statistic	Value
Minimum	64
Maximum	64000
Mean	11796.153
StdDev	11726.564

Όσον αφορά το attribute MMAX, βλέπουμε από τον αντίστοιχο πίνακα, ότι δεν υπάρχουν missing values, ότι έχει 23 distinct values και 6 unique values. Επίσης αναφέρονται τα στατιστικά: μέση τιμή (mean), τυπική απόκλιση (StdDev), ελάχιστη και μέγιστη τιμή (Minimum, Maximum).



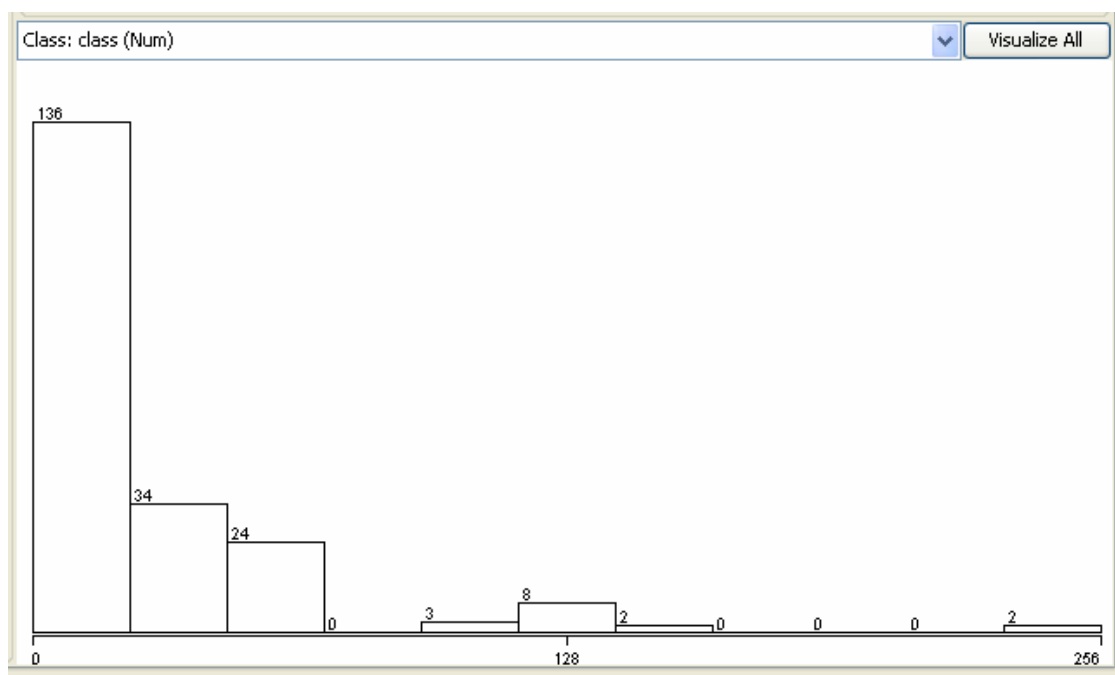
Επίσης από το παραπάνω ιστόγραμμα, βλέπουμε πώς κατανέμεται το class-attribute Performance συναρτήσει των διαφορετικών τιμών του MMAX. Στο διάστημα 64 έως 7168 παρατηρούνται 86 καταγραφές, ενώ στο 7168 έως 14272

παρατηρούνται 55 καταγραφές. Στο διάστημα 35584 έως 56896 παρατηρούνται μηδενικές καταγραφές.

D) CACH

Selected attribute	
Name: CACH	Type: Numeric
Missing: 0 (0%)	Distinct: 22
	Unique: 4 (2%)
Statistic	Value
Minimum	0
Maximum	256
Mean	25.206
StdDev	40.629

Σχετικά με το attribute CACH, βλέπουμε από τον αντίστοιχο πίνακα, ότι δεν υπάρχουν missing values, ότι έχει 22 distinct values και 4 unique values. Επίσης αναφέρονται τα στατιστικά: μέση τιμή (mean), τυπική απόκλιση (StdDev), ελάχιστη και μέγιστη τιμή (Minimum, Maximum).

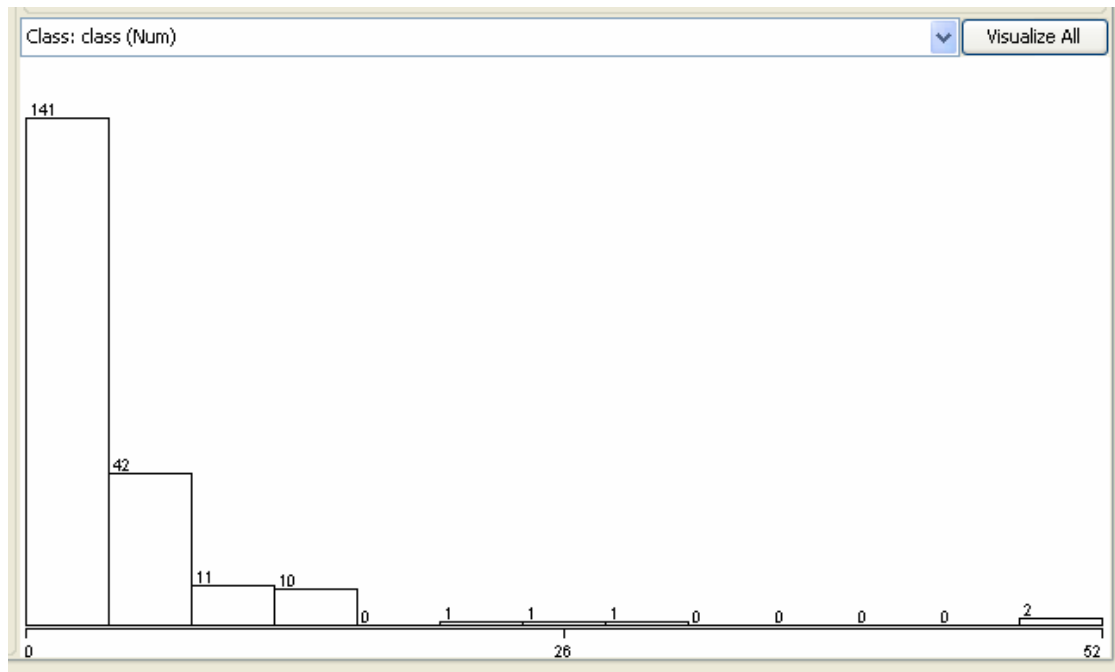


Από το παραπάνω ιστόγραμμα, βλέπουμε πώς κατανέμεται το class-attribute Performance συναρτήσει των διαφορετικών τιμών του CACH. Οι περισσότερες καταγραφές εντοπίζονται στο διάστημα 0 μέχρι 23,73, ενώ μόλις δύο καταγραφές παρουσιάζονται στο διάστημα 232,725 έως 256.

E) CHMIN

Selected attribute	
Name: CHMIN	Type: Numeric
Missing: 0 (0%)	Distinct: 15
	Unique: 4 (2%)
Statistic	Value
Minimum	0
Maximum	52
Mean	4.699
StdDev	6.816

Σχετικά με το attribute CHMIN, βλέπουμε από τον αντίστοιχο πίνακα, ότι δεν υπάρχουν missing values, ότι έχει 15 distinct values και 4 unique values. Επίσης αναφέρονται και εδώ τα στατιστικά: μέση τιμή (mean), τυπική απόκλιση (StdDev), ελάχιστη και μέγιστη τιμή (Minimum, Maximum). Παρατηρούμε ότι η ελάχιστη τιμή εδώ είναι το μηδεν.

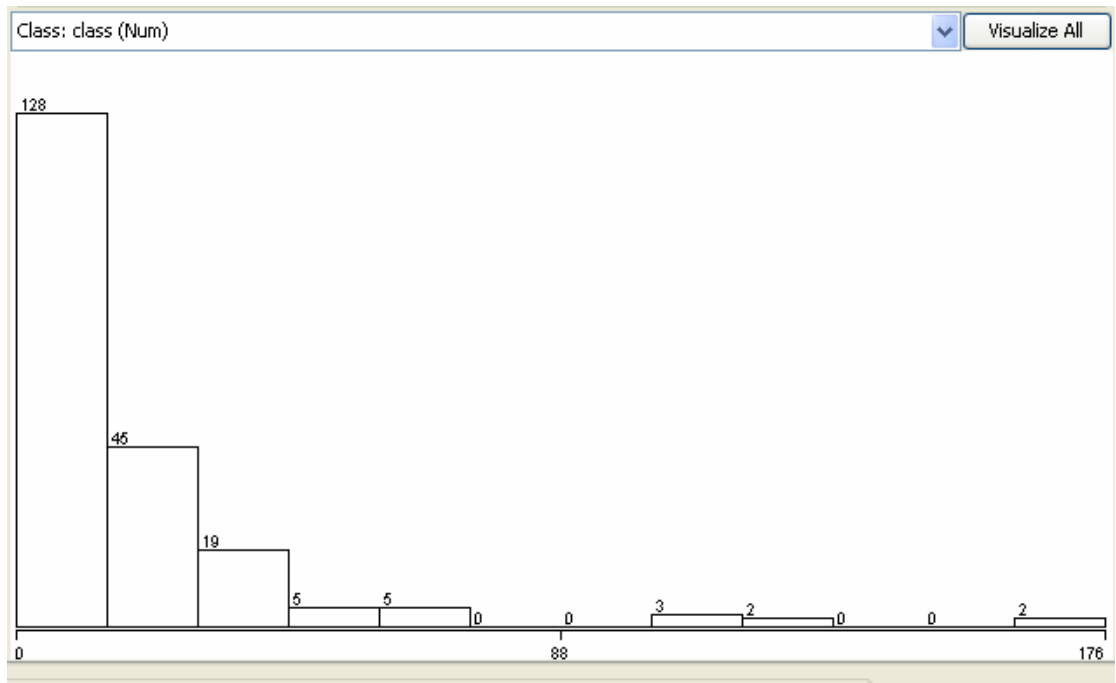


Από το παραπάνω ιστόγραμμα, βλέπουμε πώς κατανέμεται το class-attribute Performance συναρτήσει των διαφορετικών τιμών του CHMIN. Στο ιστόγραμμα αυτού του attribute υπάρχουν πάρα πολλές μηδενικές τιμές, καθώς και τιμές αρκετά χαμηλές, π.χ. η τιμή 1 που ίσως να μπορεί να θεωρηθεί ως outlier. Η πλειοψηφία των καταγραφών (141) ανήκει στο διάστημα 0 έως 4 και ανάλογα συμπεράσματα μπορούμε να βγάλουμε για όλο το εύρος των τιμών.

F) CHMAX

Selected attribute	
Name: CHMAX	Type: Numeric
Missing: 0 (0%)	Distinct: 31
	Unique: 9 (4%)
Statistic	Value
Minimum	0
Maximum	176
Mean	18.268
StdDev	25.997

Σχετικά με το attribute CHMAX, βλέπουμε από τον αντίστοιχο πίνακα, ότι δεν υπάρχουν missing values, ότι έχει 31 distinct values και 9 unique values. Επίσης αναφέρονται και εδώ τα στατιστικά: μέση τιμή (mean), τυπική απόκλιση (StdDev), ελάχιστη και μέγιστη τιμή (Minimum, Maximum), με ελάχιστη τιμή το μηδέν.

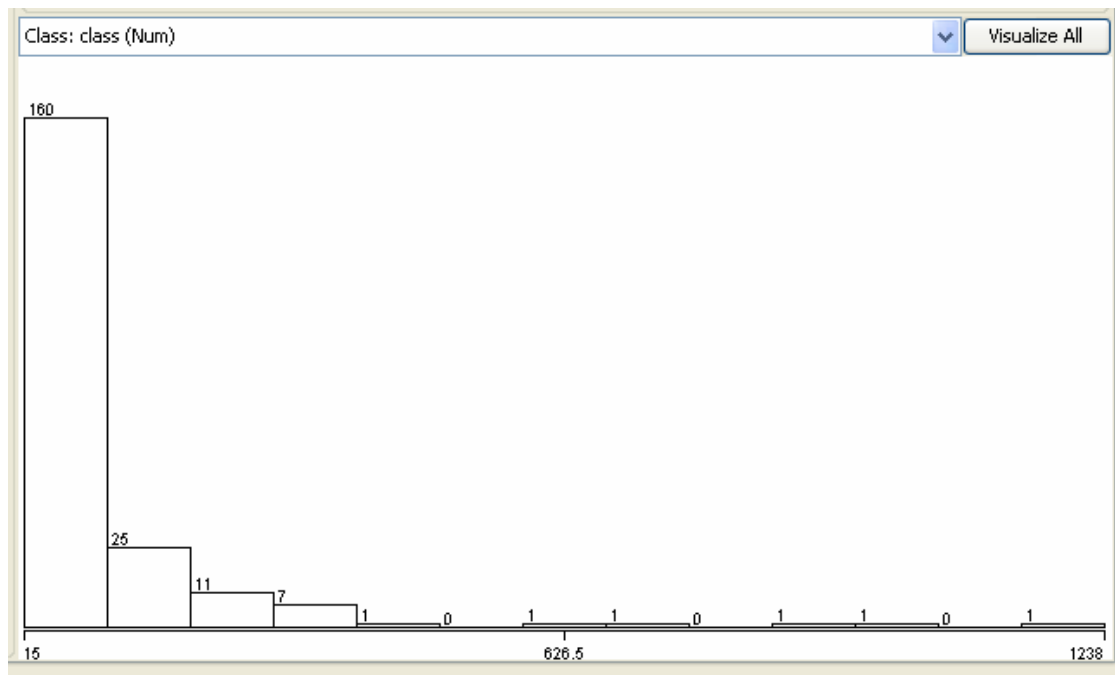


Από το παραπάνω ιστόγραμμα, βλέπουμε πώς κατανέμεται το class-attribute Performance συναρτήσει των διαφορετικών τιμών του CHMAX. Στο ιστόγραμμα αυτού του attribute η πλειοψηφία των τιμών (128) ανήκει στο διάστημα 0 έως 14667, ενώ υπάρχουν κάποιες μηδενικές τιμές, καθώς και τιμές αρκετά χαμηλές, π.χ. οι ράβδοι με τιμή 2.

G) Class (cpu performance)

Selected attribute	
Name: class	Type: Numeric
Missing: 0 (0%)	Distinct: 104
	Unique: 60 (29%)
Statistic	Value
Minimum	15
Maximum	1238
Mean	99.33
StdDev	154.757

Στον παραπάνω πίνακα βλέπουμε ότι δεν υπάρχουν missing values, ότι υπάρχουν 104 distinct values και 60 unique values. Επίσης αναφέρονται και εδώ τα στατιστικά: μέση τιμή (mean), τυπική απόκλιση (StdDev), ελάχιστη και μέγιστη τιμή (Minimum, Maximum).



Στο παραπάνω γράφημα, βλέπουμε πως είναι κατανομημένο το class-attribute cpu performance στο σύνολο του πληθυσμού μας. Παρατηρούμε ότι οι επικρατέστερες τιμές είναι οι χαμηλές, με την πλειοψηφία (160) να ανήκει στο διάστημα 15 έως 109,077, ενώ στο υπόλοιπο εύρος τιμών παρατηρούμε πολύ λιγότερες τιμές.

Λόγω του μεγάλου όγκου των δεδομένων για μια ασφαλή εξαγωγή association rules προτείνεται η χρήση αλγορίθμων, καθώς οπτικά μπορεί να γίνει μόνο μια απλή περιγραφή του dataset.

