**Αναφορά 2ης Άσκησης Τεχνικές Εξόρυξης Δεδομένων**

**1.Οπτικοποίηση Δεδομένων**

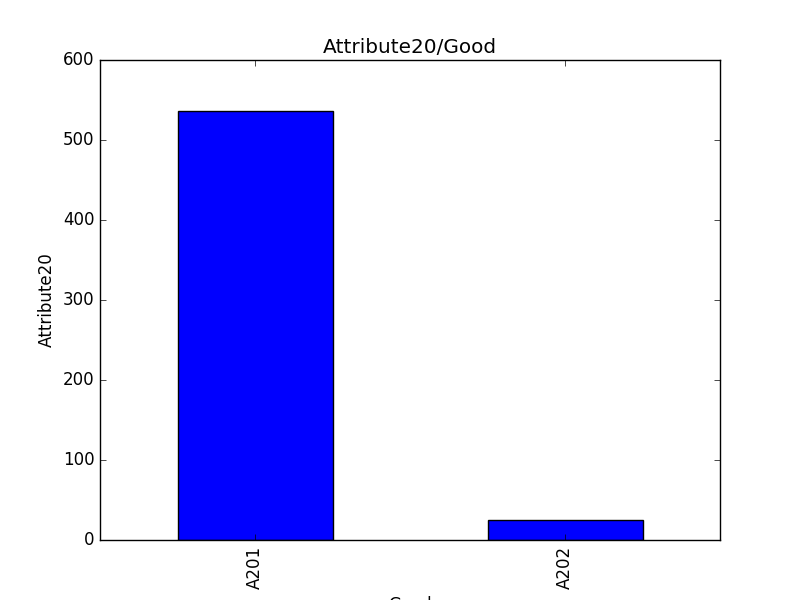
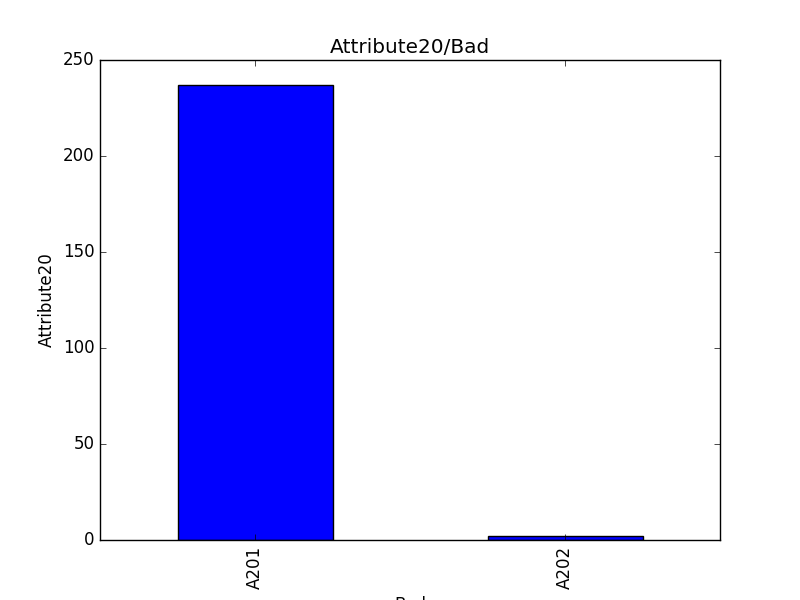
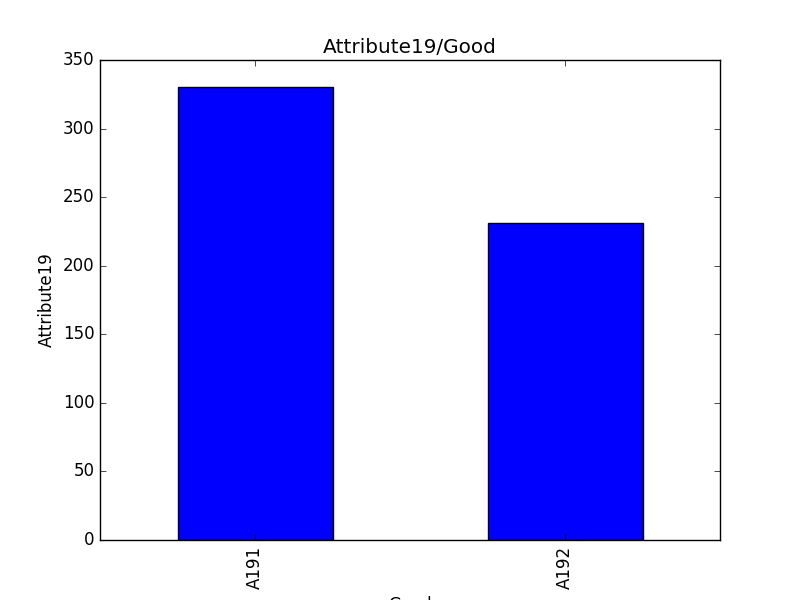
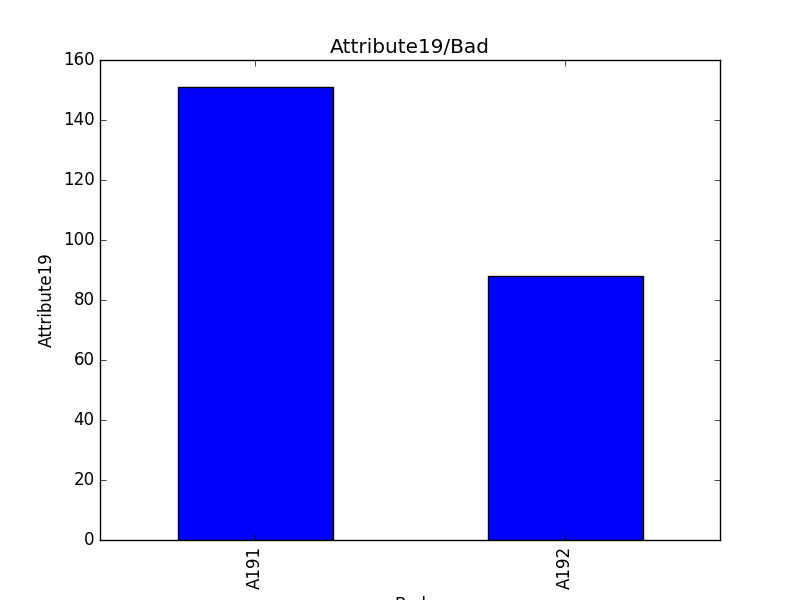
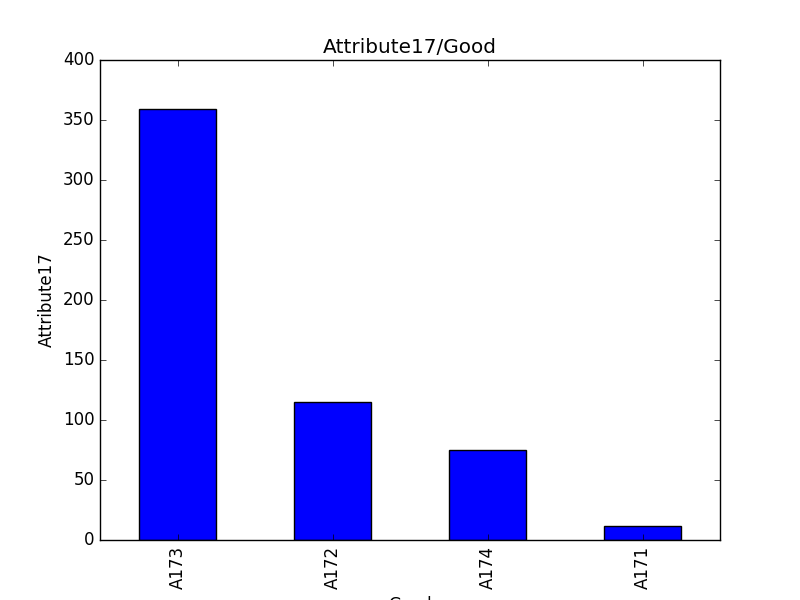
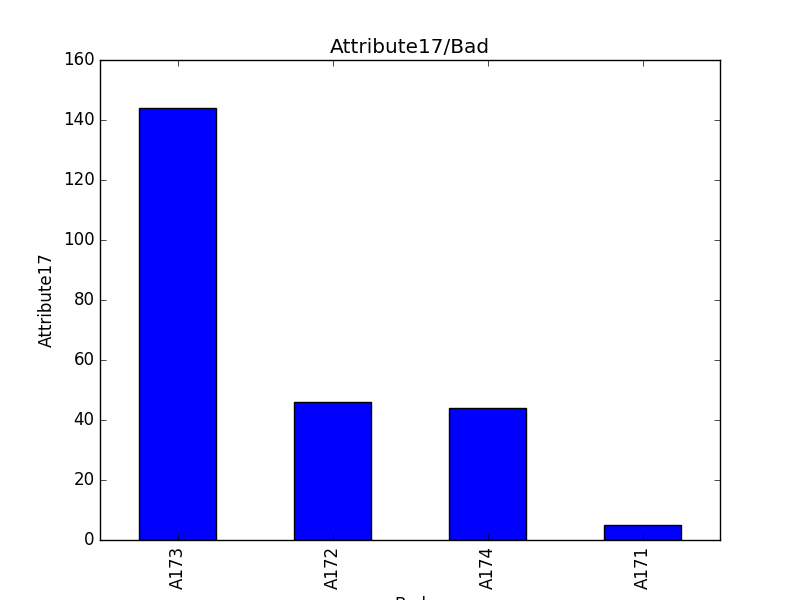
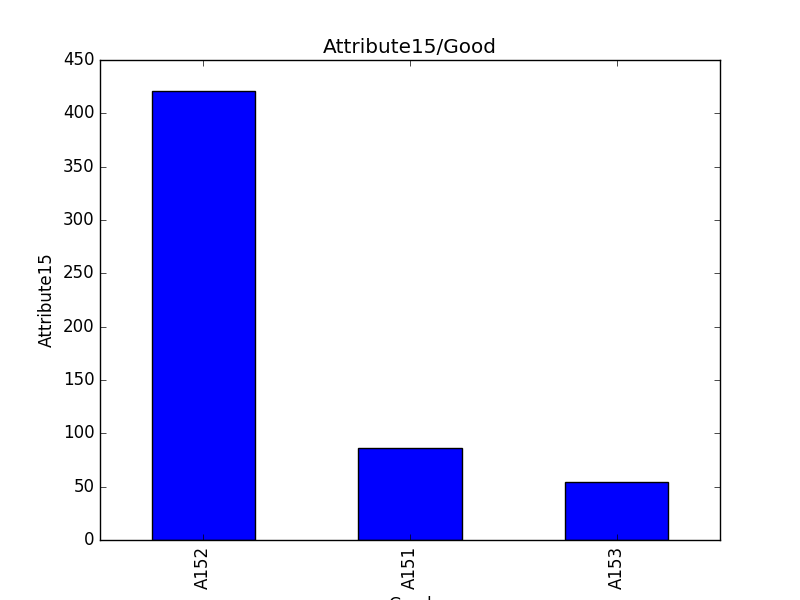
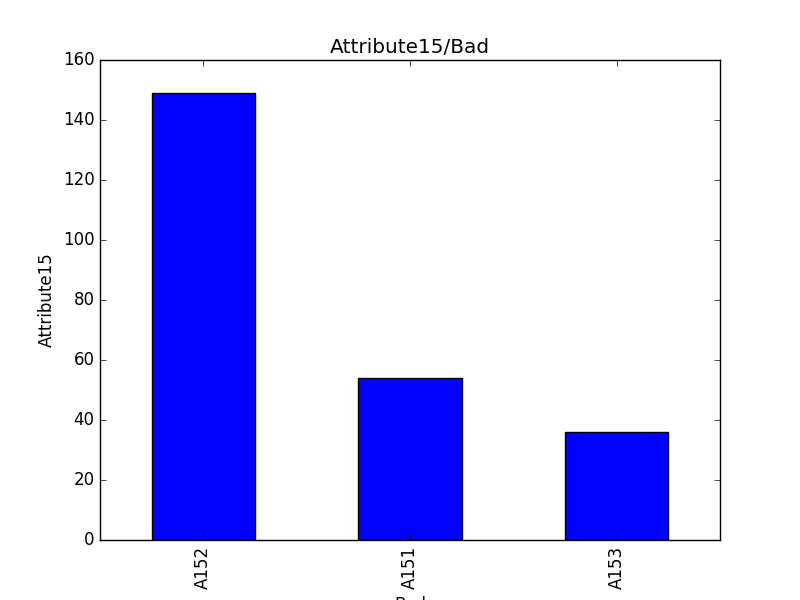
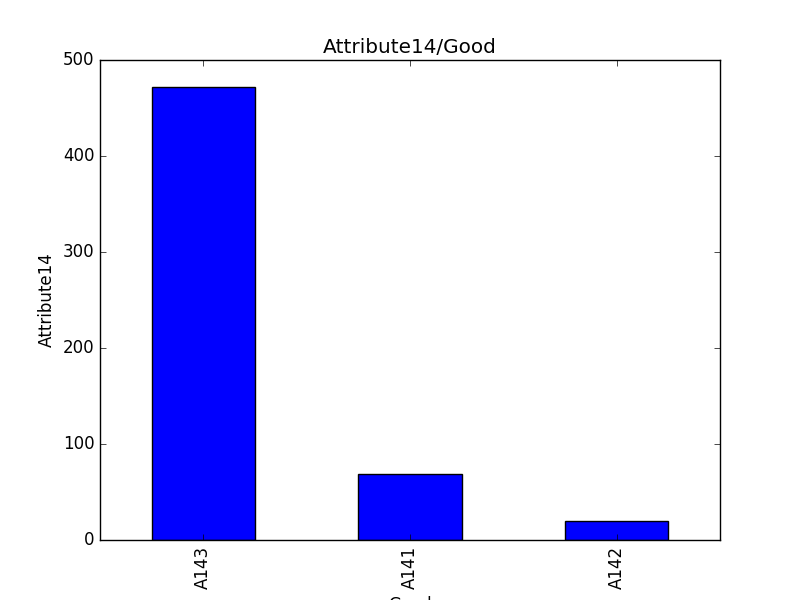
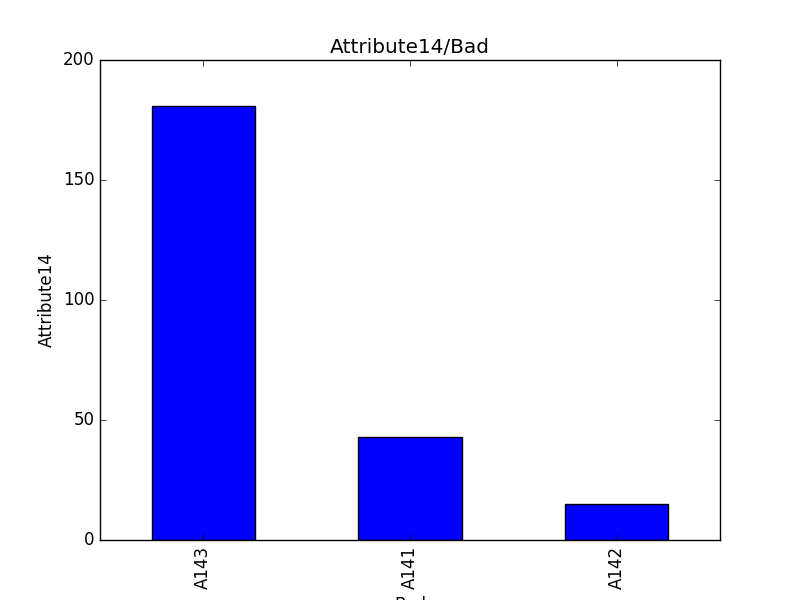
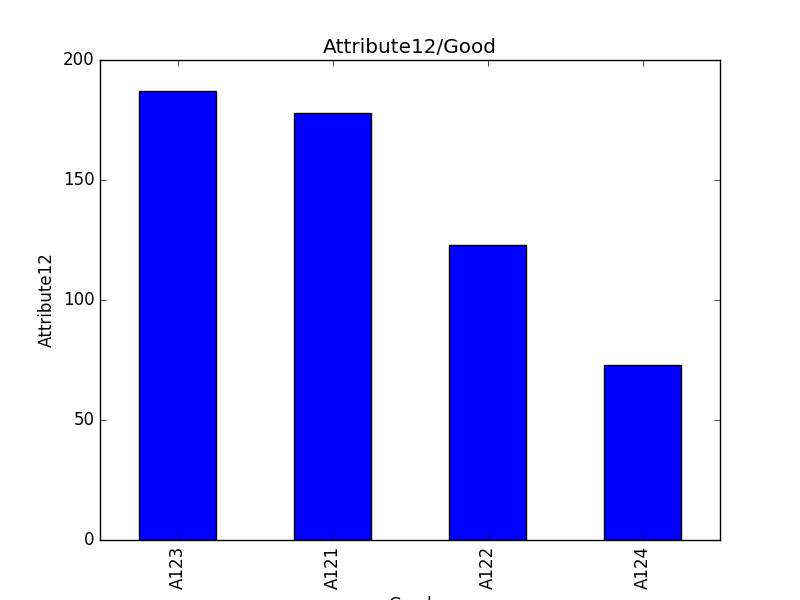
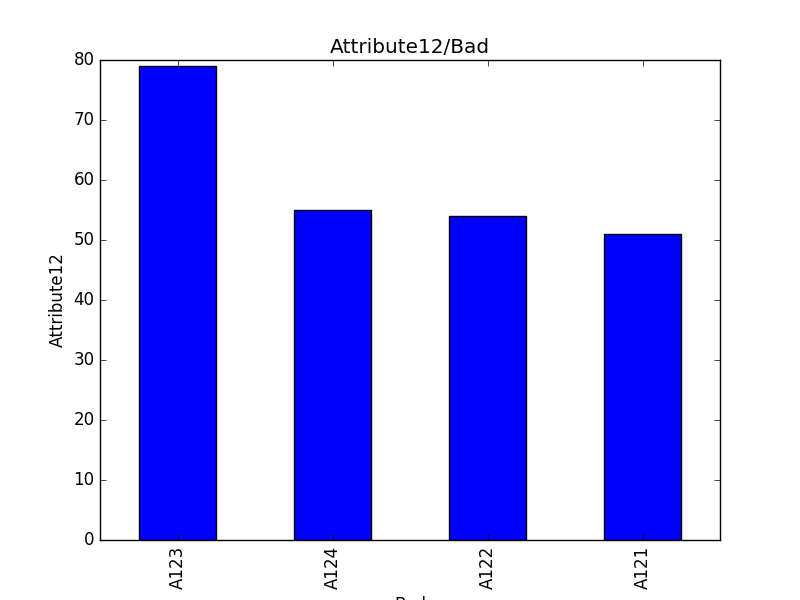
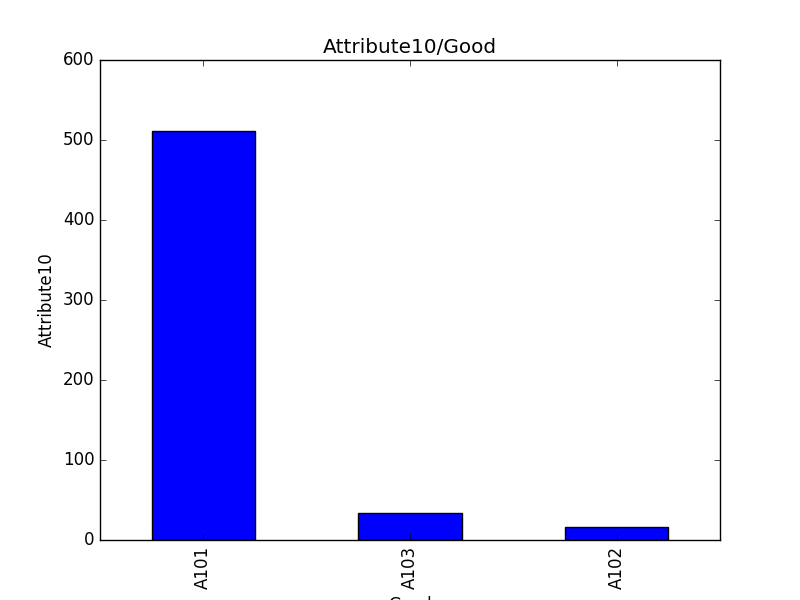
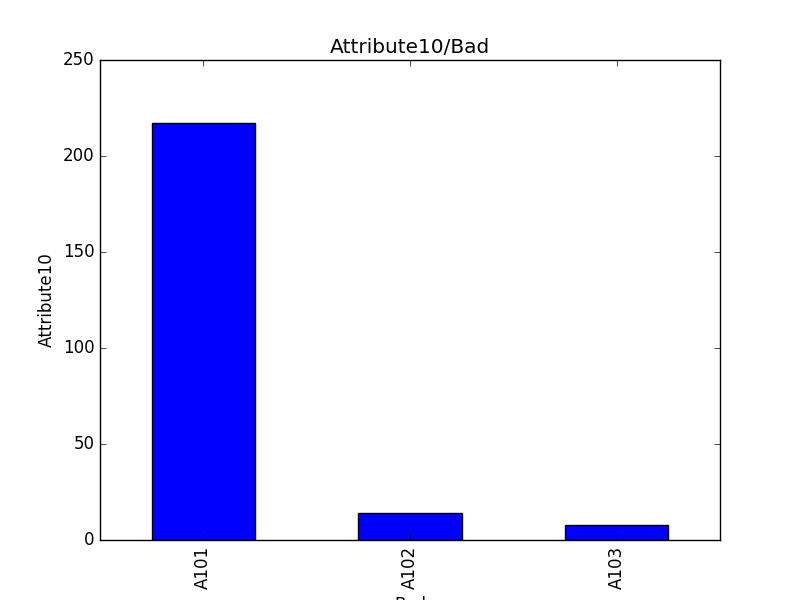
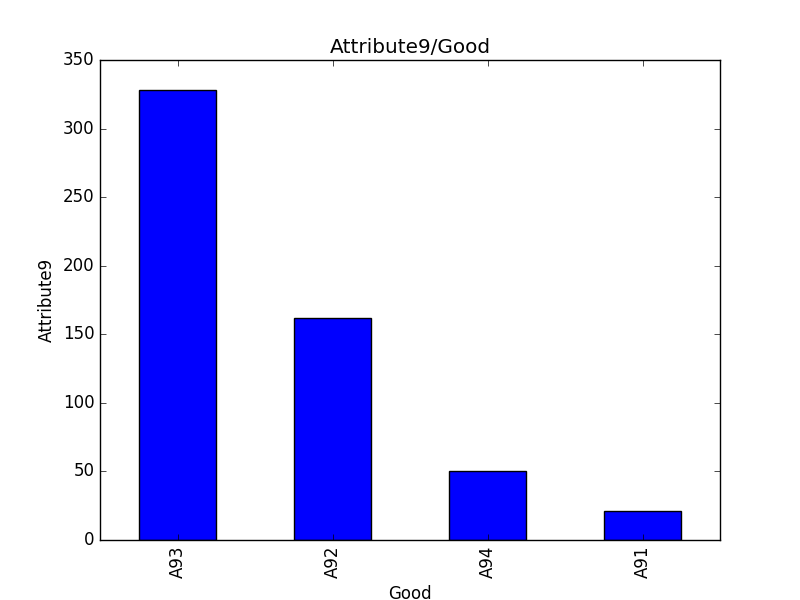
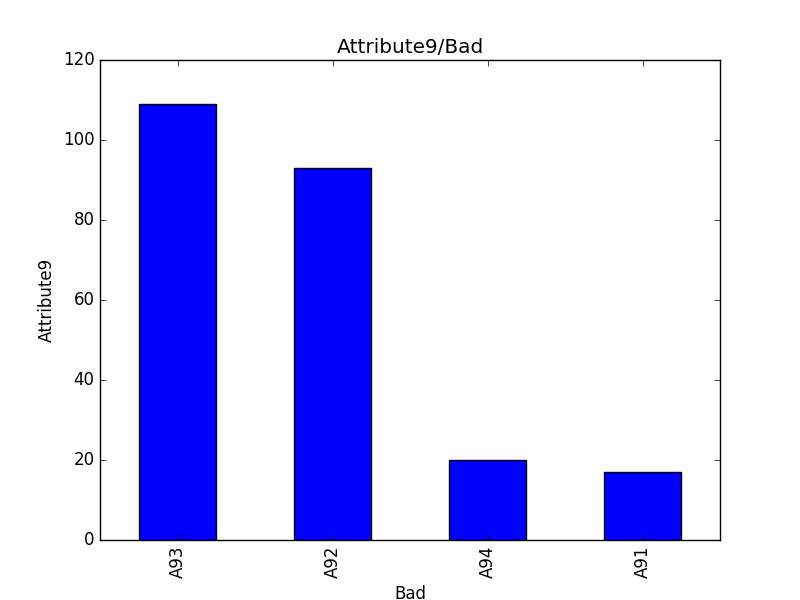
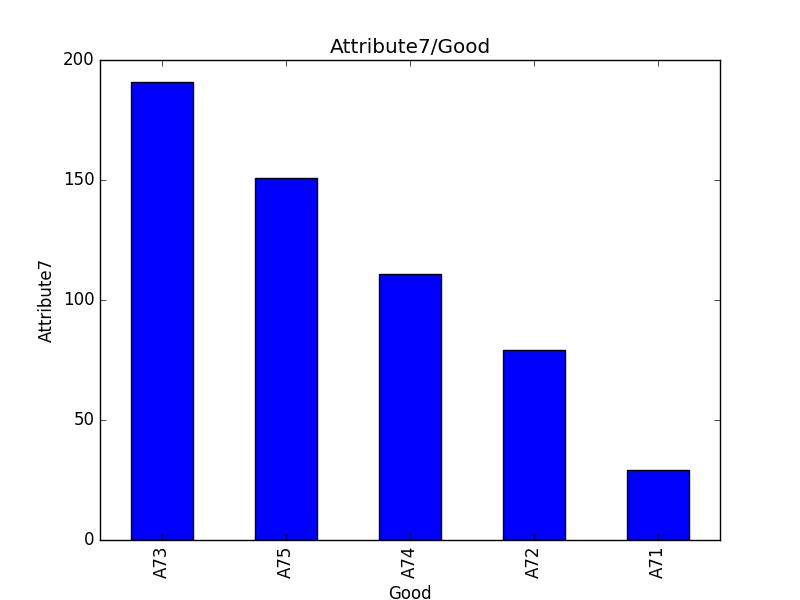
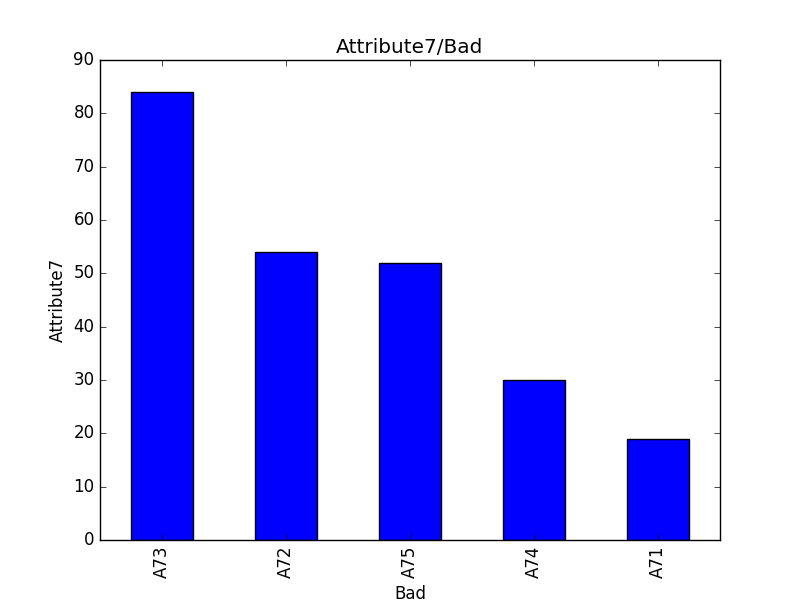
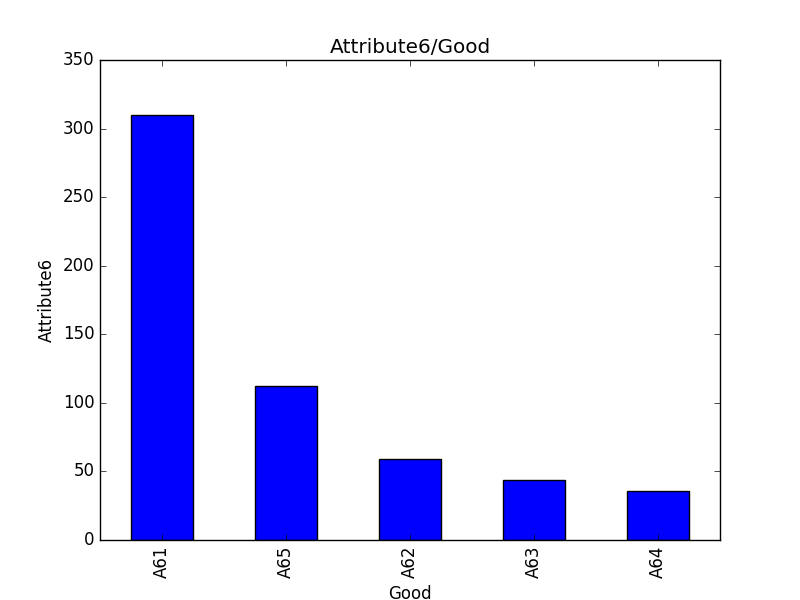
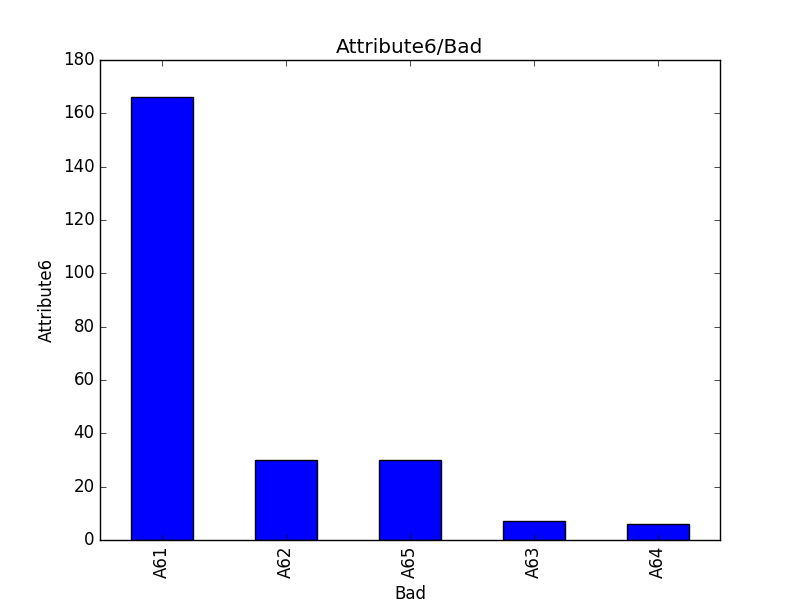
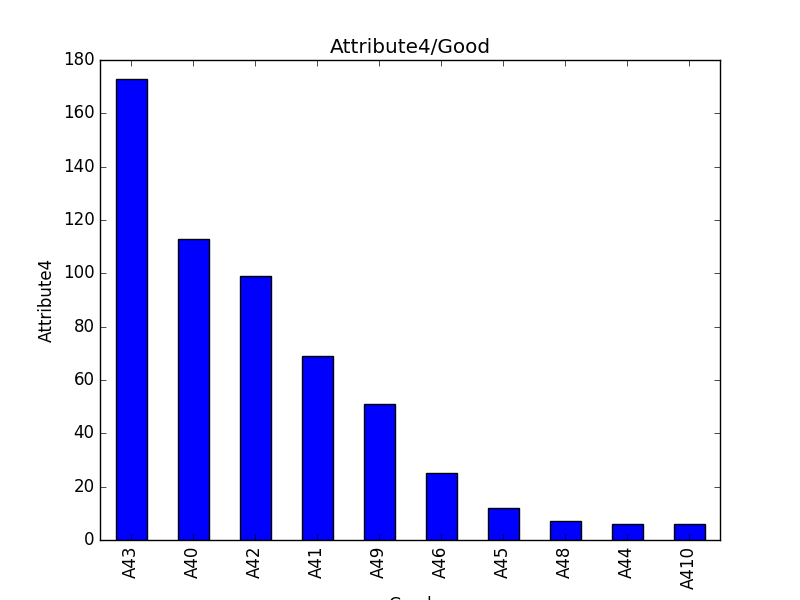
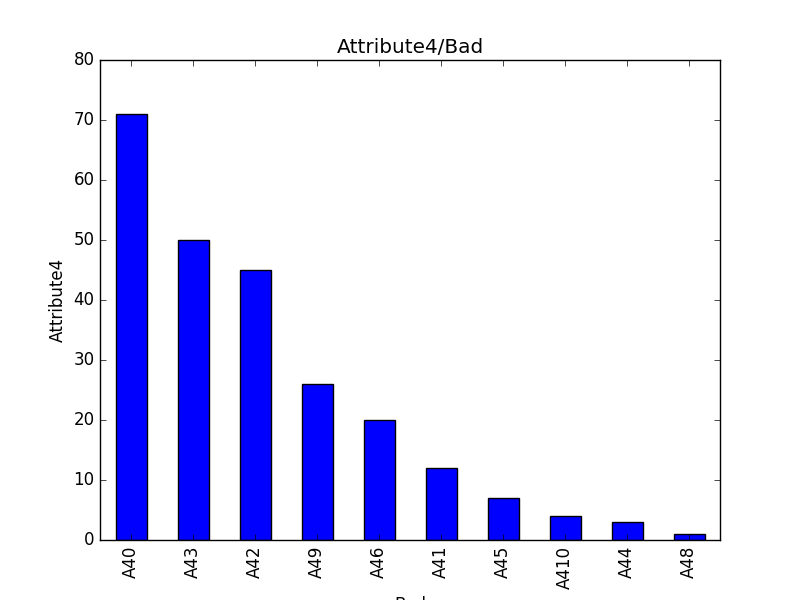
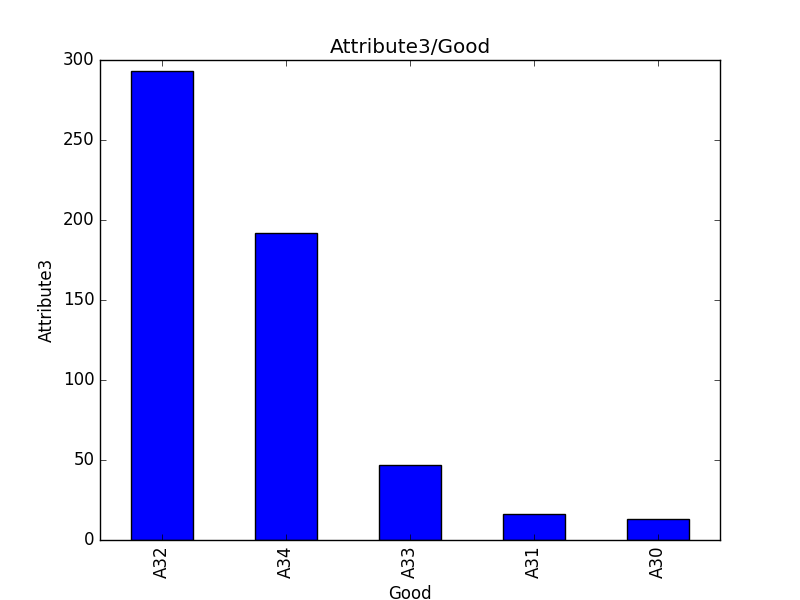
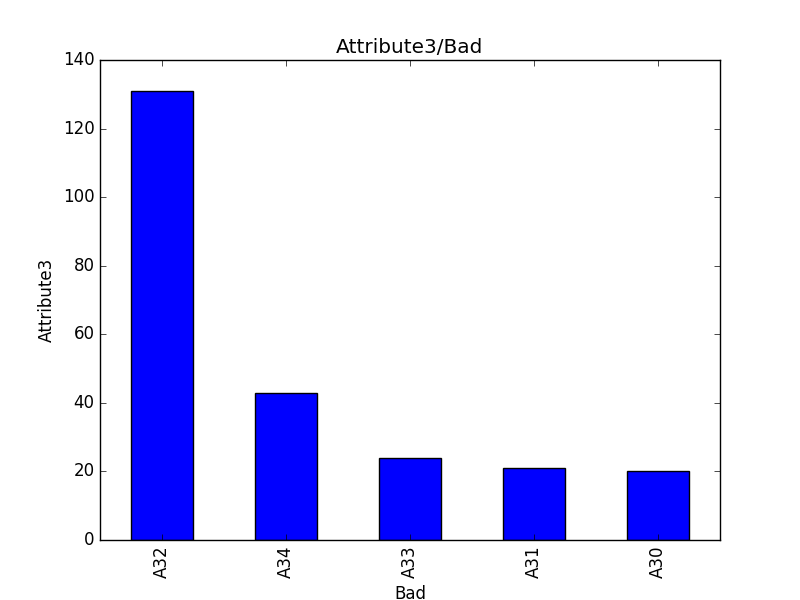
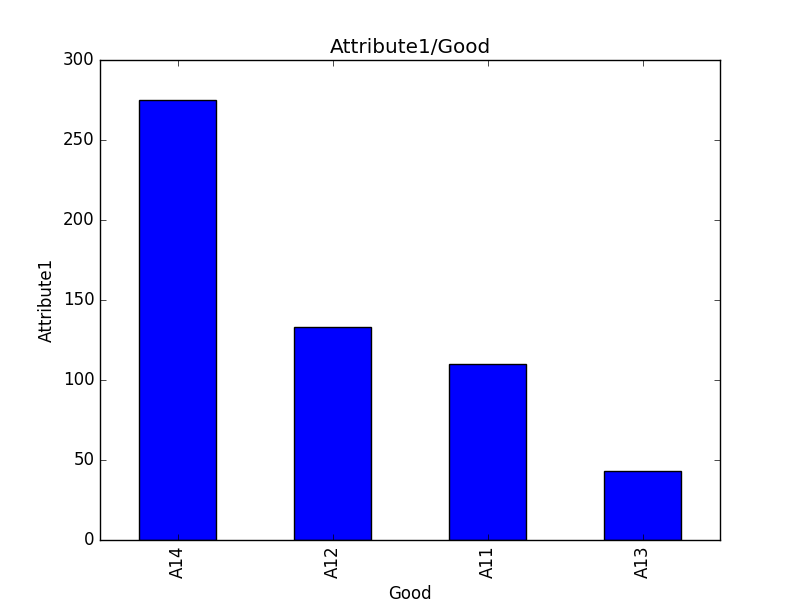
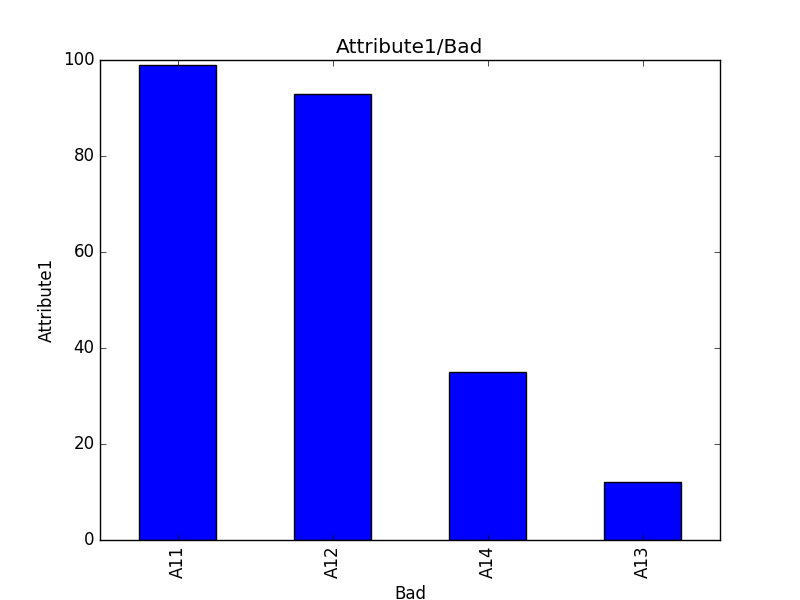
Σε αυτό το ερώτημα υλοποιήσαμε πλήρως τα histograms και box plots για τα categorical και numerical data αντίστοιχα.

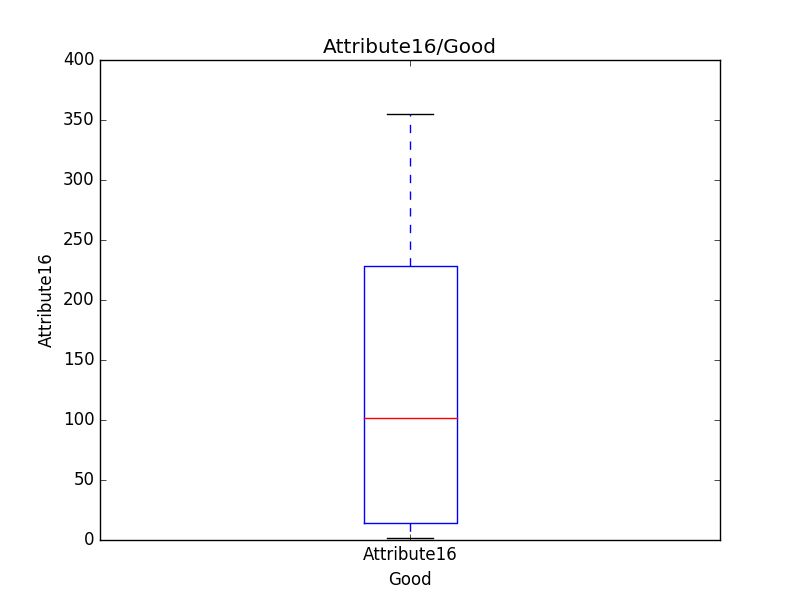
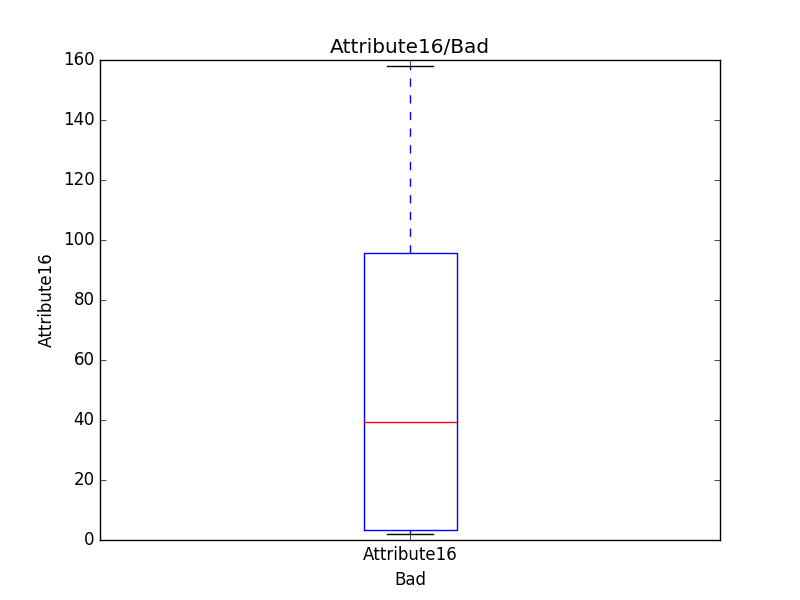
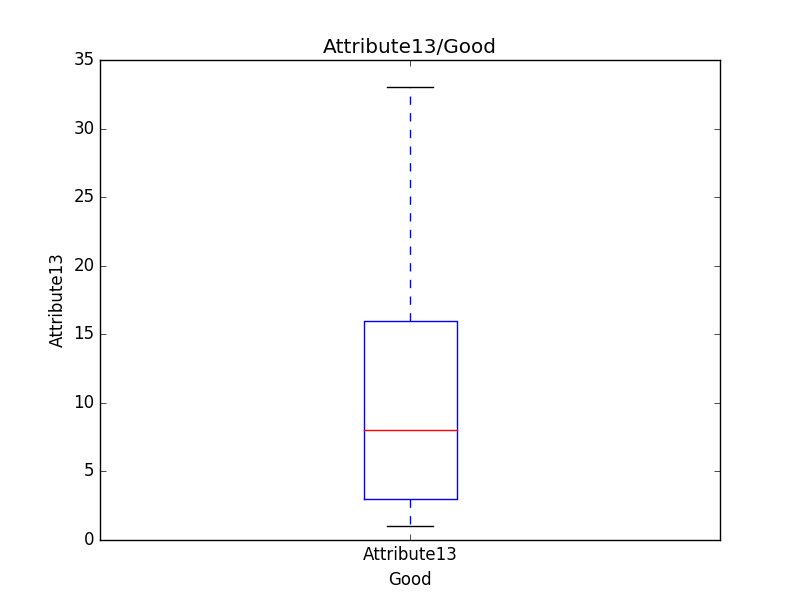
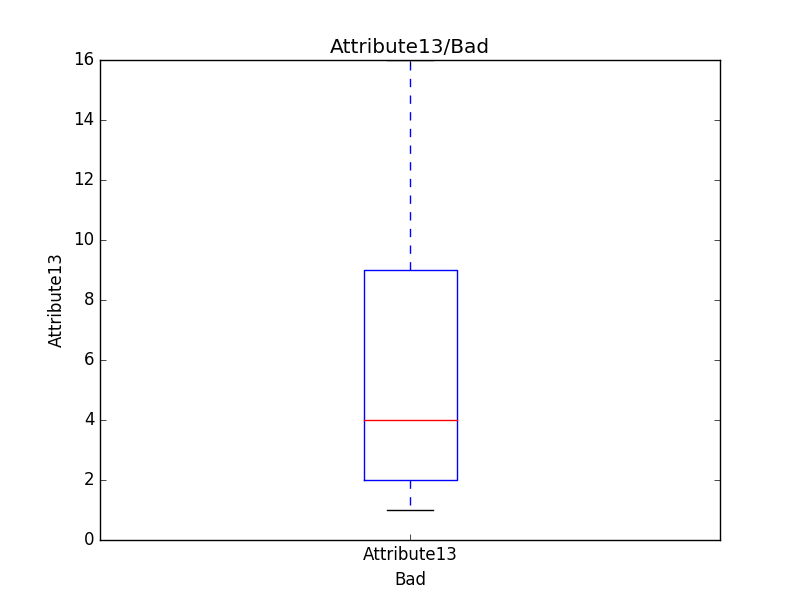
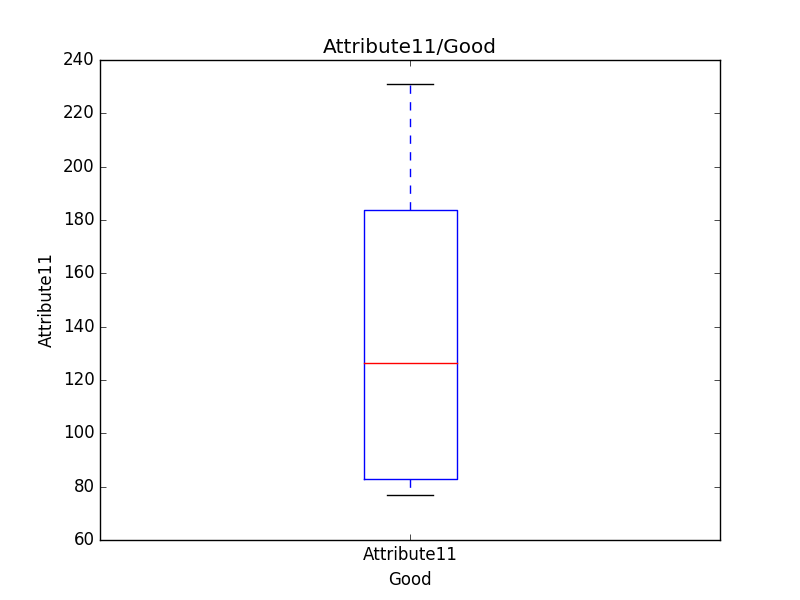
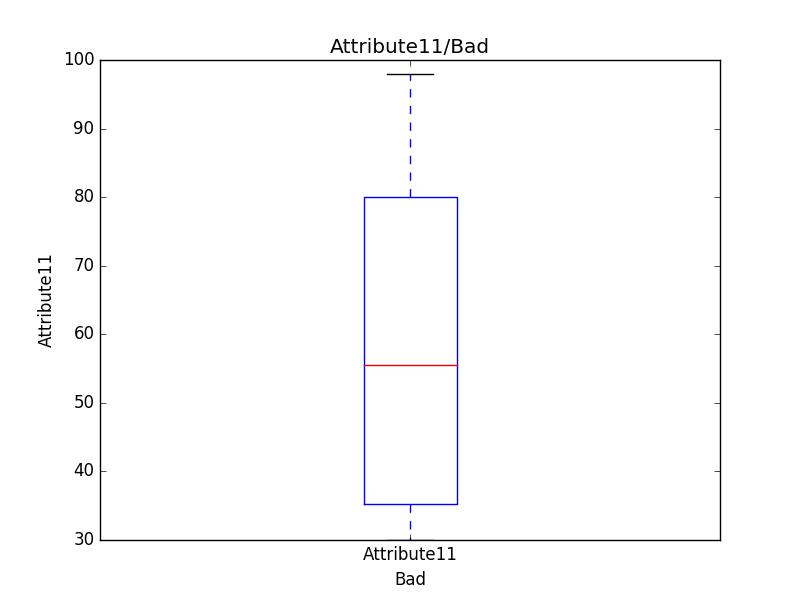
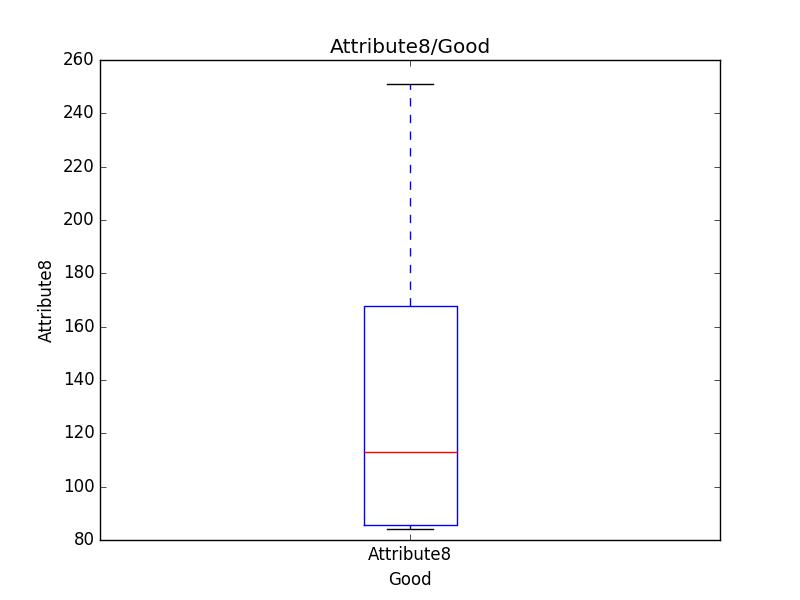
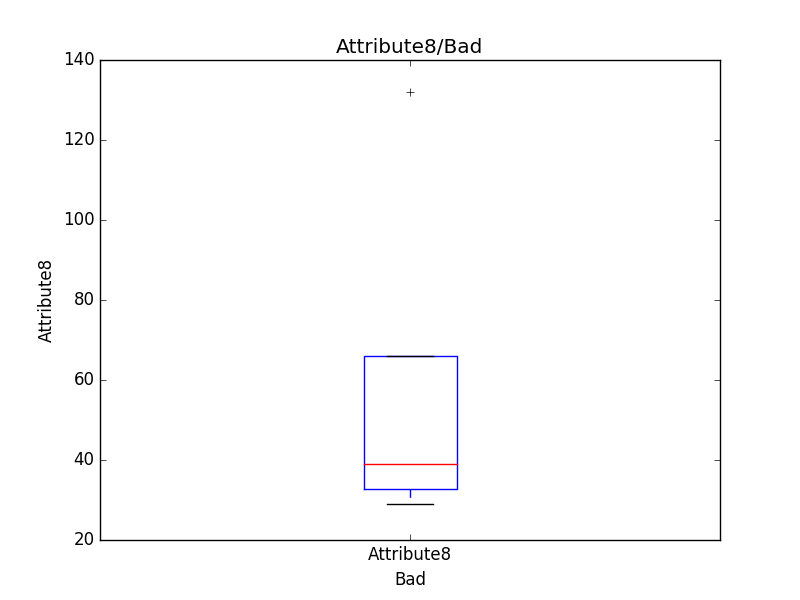
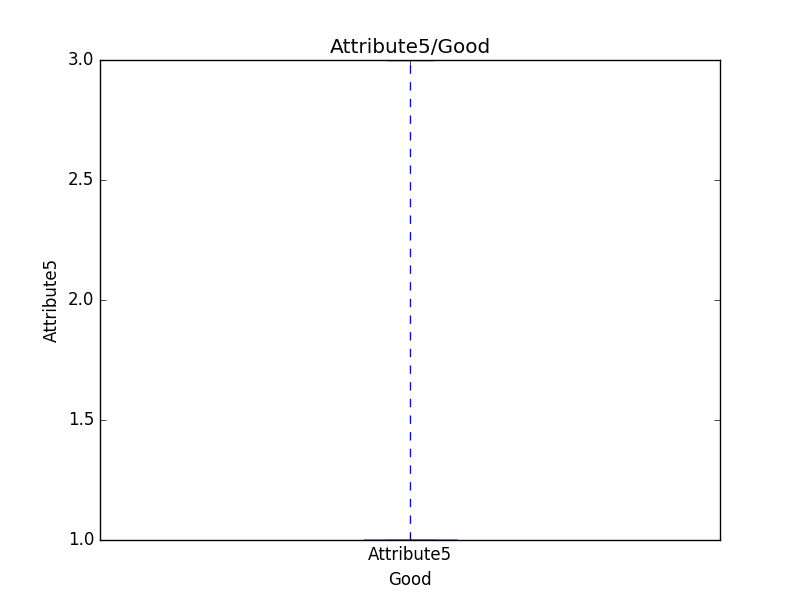
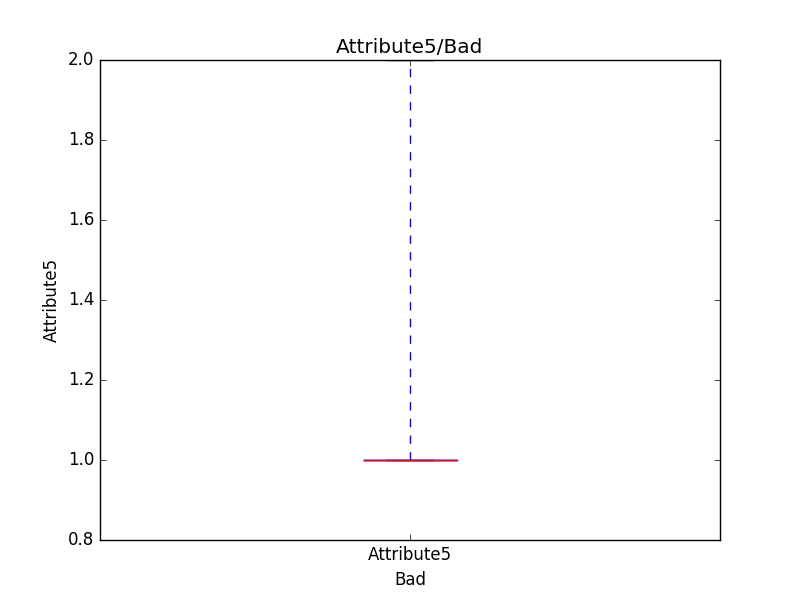
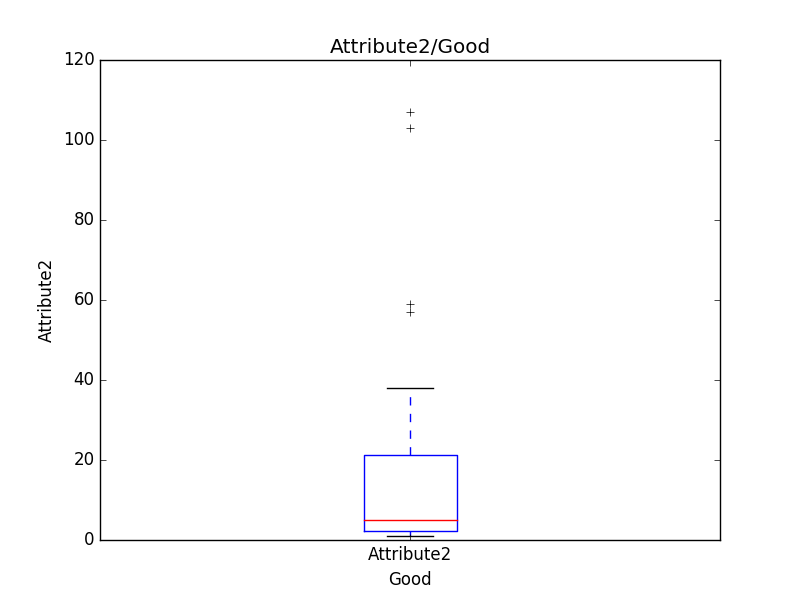
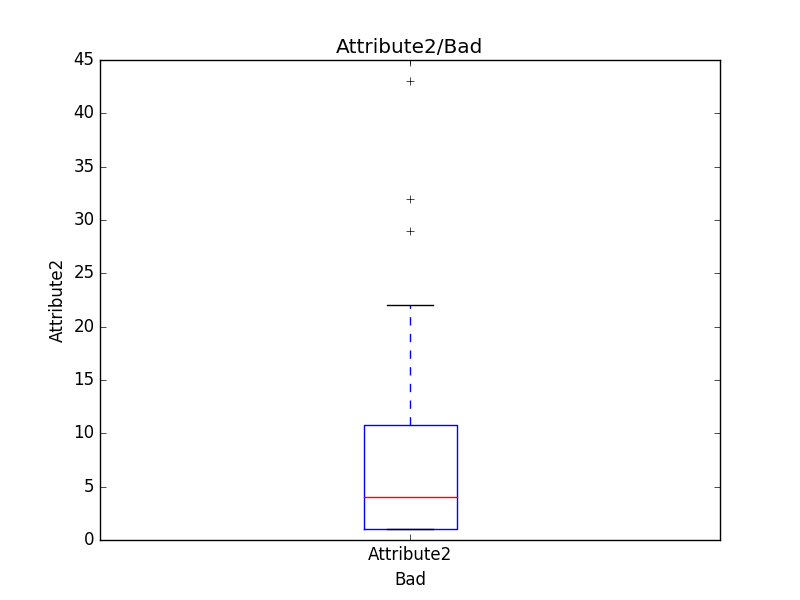
**Συμπέρασμα:**

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.

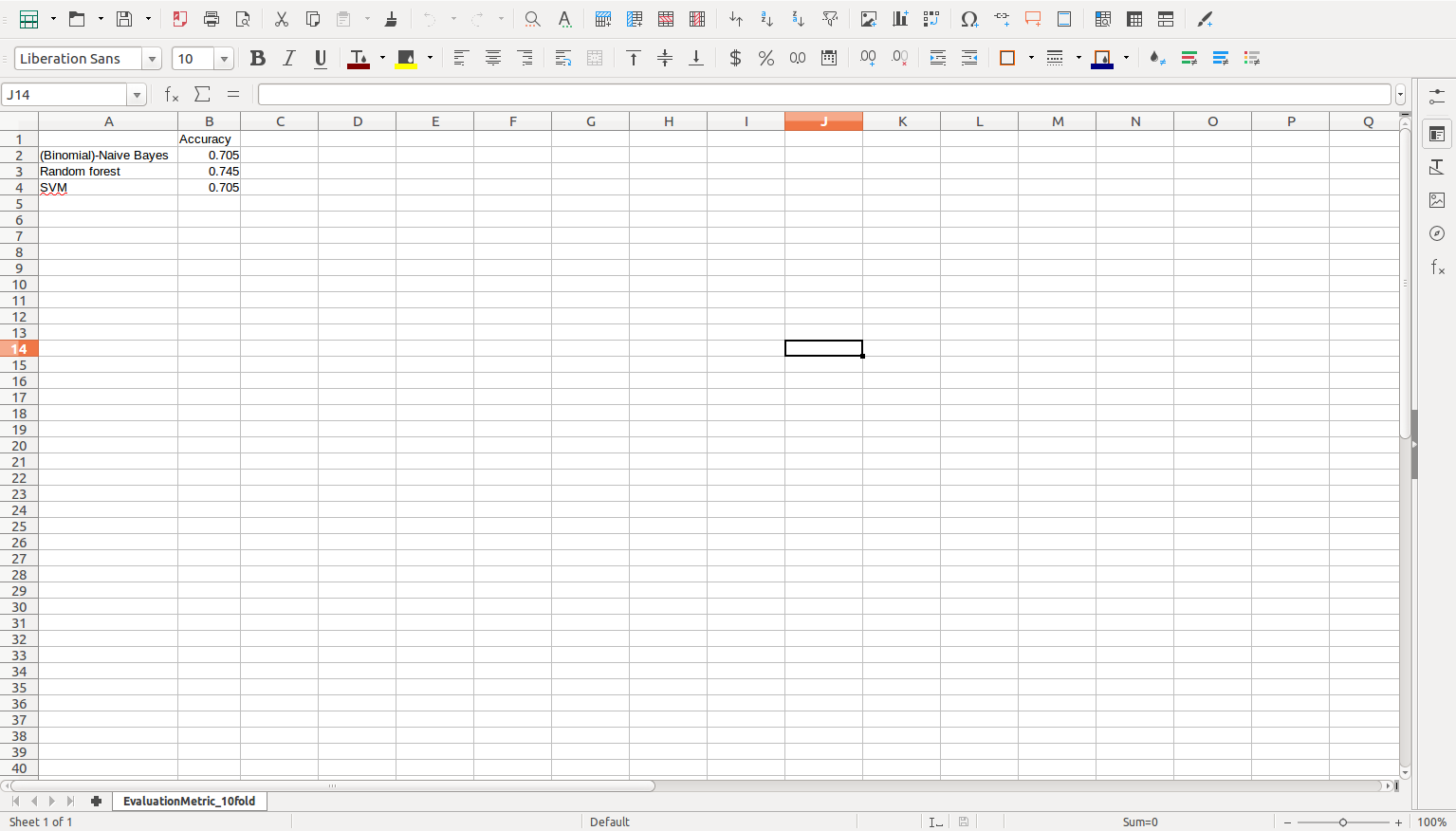


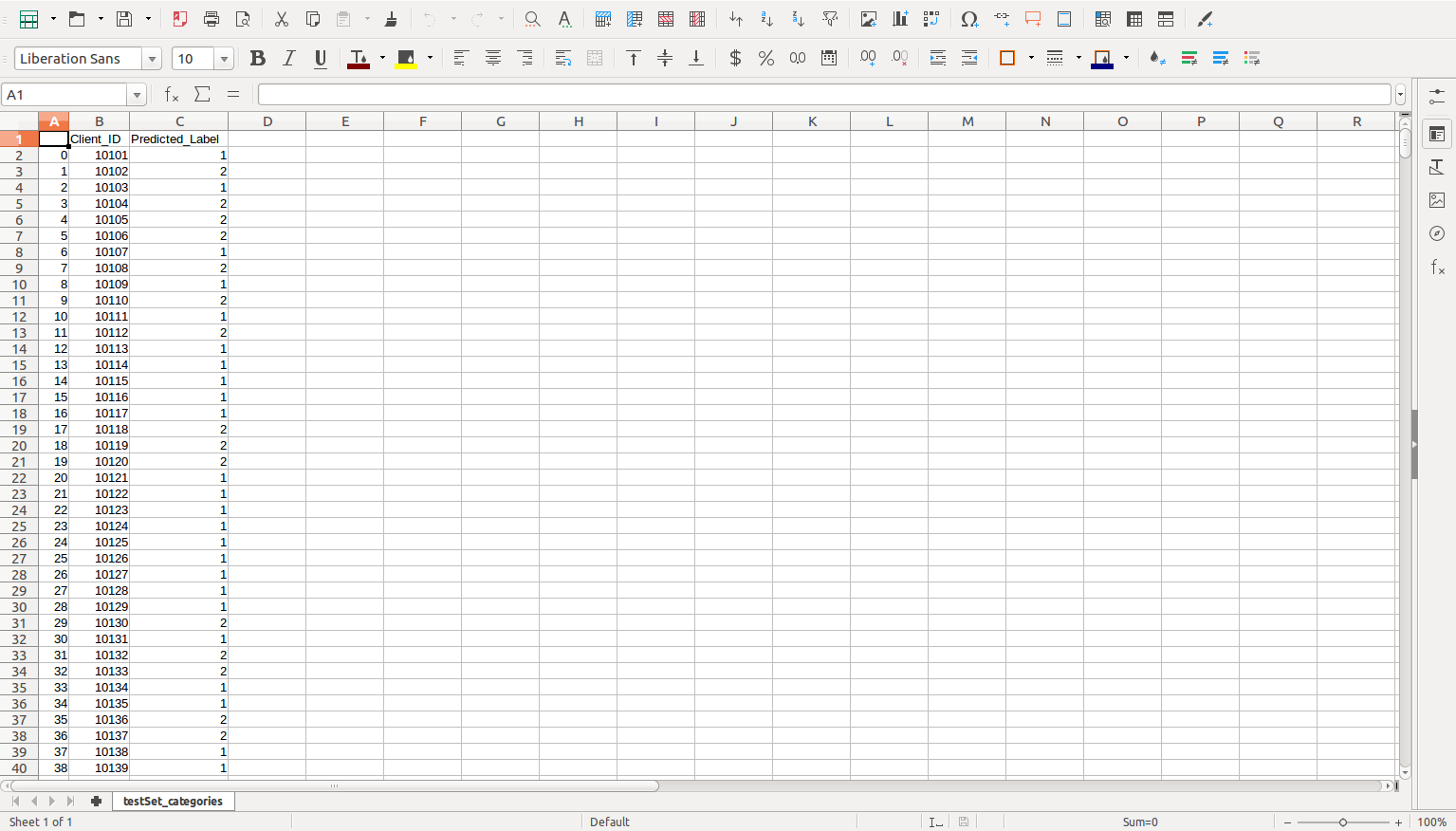


**2.Υλοποίηση Κατηγοριοποίησης (Classification)**

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):





**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 10ο 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)

