Blue text on a black background

Description automatically generated

***Σχολή Μηχανικών***

***Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών***

*Διαχείριση Δεδομένων Μεγάλης Κλίμακας*

***Αναφορά Ανάλυσης Δεδομένων***

*Εργασία: Best-Value Smartphone Analysis*

*ΕΥΦΡΟΣΥΝΗ ΒΑΡΣΟΥ 21390021*

*ΑΓΓΕΛΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΕΝΤΖΕΛΟΣ 21390132*

# 1. Εισαγωγή

Αυτό το έγγραφο αποτελεί την συγγραφή της αναφοράς για την ανάλυση δεδομένων που πραγματοποιήσαμε σε ένα dataset που αφορά για τα διάφορα κινητά που υπάρχουν στην αγορά και τα αντίστοιχα χαρακτηριστικά τους. Θα επικεντρωθούμε στην ανάλυση σχετικά με την τιμή του κάθε κινητού σε αντιπαράθεση με τον επεξεργαστή που έχουν και την RAM.

Αρχικά, θα περιγράψουμε το πρόβλημα που επιχειρούμε να λύσουμε με την ανάλυση αυτή, έπειτα θα περιγράψουμε εν συντομία το dataset που χρησιμοποιήσαμε και τέλος θα επικεντρωθούμε στην περιγραφή του κώδικα που υλοποιήσαμε καθώς και την περιγραφή/αποτίμηση των αποτελεσμάτων που συμπεραίνουμε από αυτά.

Σε αυτήν την αναφορά δεν θα συμπεριλάβουμε όλα τα περιεχόμενα των πινάκων καθώς είναι πολλά για τις δύο αναλύσεις που υλοποιήσαμε (Processor, RAM) ωστόσο υπάρχουν όλοι αναλυτικά στον κώδικα .ipynb. Ο κώδικας .py περιέχει και αυτός όλη την ανάλυση αλλά συμπεριλήφθηκε για πρακτικούς λόγους, συνεπώς η ανάλυση μαζί με τις πηγές που χρησιμοποιήσαμε βρίσκεται αναλυτικά στον κώδικα .ipynb (θα ήταν προτιμότερο δηλαδή να προβληθεί από εκεί αφού είναι ενσωματωμένοι όλοι οι πίνακες και τα σχήματα).

# 2. Ορισμός Προβλήματος

Επιλέξαμε αυτό το πρόβλημα γιατί θέλαμε να προσεγγίσουμε το θέμα ανάλυσης των πιο value for money smartphones. Γενικά, στις μέρες μας ειδικότερα στο tech world του διαδικτύου όλοι επιχειρούν να απαντήσουν στο ερώτημα ποιο κινητό ή λειτουργικό είναι καλύτερο. Όπως για παράδειγμα, συχνό debate είναι η επιλογή μεταξύ Apple ή Android κινητού. Με τους υποστηρικτές της Apple να υποστηρίζουν ότι αν και πιο ακριβά είναι πιο ποιοτικά και με τους υποστηρικτές των Android να υποστηρίζουν ότι είναι το ίδιο ποιοτικά (αν όχι και περισσότερο) και σε πιο προσιτή τιμή. Στην ανάλυση που θα κάνουμε θα προσπαθήσουμε να δείξουμε σε πιο εύρος τιμών το hardware(processor, RAM) ισοδυναμεί με την τιμή. Αν δηλαδή αυτά που πληρώνεις είναι η «μάρκα» ή όντως η αξία του κινητού σε σύγκριση με τις δυνατότητες του. Και γενικά, ποιο εύρος τιμών είναι το καλύτερο για μια αξιοπρεπής αγορά ενός κινητού. Συνεπώς, ποιο είναι το πιο best value smartphone στην αγορά σήμερα.

# 3. Περιγραφή του συνόλου δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήσαμε περιέχει 11 στήλες και 1021 γραμμές που αφορούν σε γενικές πληροφορίες σχετικά με κινητά τηλέφωνα. Έχει γενικά πληροφορίες για κινητά από πολύ φθηνά έως πολύ ακριβά και όλων των λειτουργικών. Από τις 11 στήλες του dataset εμείς χρησιμοποιήσαμε την στήλη Price (την τιμή σε ινδικά ρούπια την οποία μετατρέψαμε σε ευρώ), την Processor (που έχει το όνομα του επεξεργαστή , τους πυρήνες του και στην συχνότητα του) και την RAM (που έχει την μνήμη RAM και τον αποθηκευτικό builitin χώρο). Οι υπόλοιπες στήλες που υπάρχουν στο αρχείο αλλά δεν χρησιμοποιήσαμε είναι γενικά χαρακτηριστικά που αφορούν στο όνομα του κινητού, το Rating του, την Sim που δέχεται (δηλαδή τον αριθμό των καρτών), το μέγεθος της μπαταρίας του σε mAh, την ανάλυση της οθόνης, την κάμερα που έχει, αν δέχεται έξτρα κάρτα αποθηκευτικού χώρου και το λειτουργικό σύστημα του κάθε κινητού τηλεφώνου.

# 4. Περιγραφή των μεθόδων ανάλυσης δεδομένων

Για την υλοποίηση της ανάλυσης χρησιμοποιήσαμε τρείς τεχνικές. Αρχικά, κάναμε την τεχνική One Hot Encoding στις κατηγορικές τιμές του Processor (και αντίστοιχα της RΑΜ) στην συνέχεια κάναμε Clustering (K-Means) για το grouping των κινητών και τέλος κάναμε την τεχνική του Association Rule Mining (Apriori function) για την εξαγωγή των συσχετίσεων. Ο λόγος που χρησιμοποιήσαμε αυτές τις τεχνικές θα τις παρουσιάσουμε παρακάτω εξηγώντας εν συντομία τι κάνει γενικά η κάθε τεχνική και γιατί την χρησιμοποιούμε στην ανάλυση μας.

Η τεχνική One Hot Encoding χρησιμοποιείται γενικά στη μηχανική μάθηση για τη μετατροπή κατηγορικών δεδομένων σε αριθμητικά δεδομένα. Τα κατηγορικά δεδομένα δηλαδή μετατρέπονται σε στήλες (ανάλογα πόσα είναι) και για κάθε στοιχείο η τιμή της στήλης γίνεται “Hot” (1) ή “Cold” (0) αν η συγκεκριμένη γραμμή έχει ή όχι το στοιχείο αυτό. Αυτή είναι η μορφή που θέλουμε n-dimensions με κάθε κατηγορία να χωρίζεται σε στήλες. Υπάρχει και άλλη μορφή που γίνεται η κωδικοποίηση σε μια στήλη (1-dimension) αναθέτοντας ένα unique value (int) ανάλογα με όλες τις κατηγορίες που έχει μέσα και αν το συναντήσει ξανά θέτει τον ίδιο αριθμό. Αυτήν την μορφή χρησιμοποιήσαμε για να κάνουμε το scatter plot μεταξύ του column Price και του column Processor/RAM οι οποίες στήλες έχουν κατηγορικά δεδομένα. Με αυτόν τον τρόπο, στο σχήμα για τον ίδιο τύπο Processor/RAM (με συγκεκριμένη τιμή int) αντιστοιχίζονται διαφορετικά Prices και βγαίνει σωστό το γράφημα μεταξύ τιμής και Processor/RAM.

Η τεχνική Clustering χρησιμοποιείται γενικά για την ανάλυση δεδομένων με σκοπό την ομαδοποίηση ενός συνόλου αντικειμένων σε υποσύνολα (clusters) με βάση την ομοιότητά τους. Κάθε αντικείμενο δηλαδή, σε ένα cluster είναι παρόμοιο με τα αντικείμενα του ίδιου cluster παρά με τα αντικείμενα άλλων clusters. Στην προκειμένη περίπτωση, την χρησιμοποιούμε για να μπορέσουμε να ομαδοποιήσουμε τα κινητά τηλέφωνα με βάση την ομοιότητα στην τιμή τους. Δηλαδή, έχοντας το γράφημα με την τιμή και τους επεξεργαστές ομαδοποιούμε σε ίδια clusters τα κινητά που είναι κοντά μεταξύ τους που από την μορφή του διαγράμματος είναι τα κινητά που είναι κοντά σε τιμές. Επομένως, καταφέραμε να διαχωρίσουμε τα κινητά που είναι παρόμοια ανάλογα με την τιμή τους.

Η τεχνική Association Rule Mining χρησιμοποιείται γενικά για την ανάλυση δεδομένων με σκοπό την ανίχνευση ενδιαφέρουσων συσχετίσεων μεταξύ αντικειμένων από ένα dataset. Στην προκειμένη περίπτωση την χρησιμοποιούμε για να εξάγουμε κοινά χαρακτηριστικά μεταξύ κινητών που βρίσκονται στο ίδιο cluster. Δηλαδή έχοντας χωρίσει κάθε κινητό με βάση την τιμή εξάγουμε αποτελέσματα ανάλογα με τα πιο συχνά χαρακτηριστικά που αφορούν την Processor και την RAM. Συνεπώς, βρίσκουμε τις πιο συχνές και ενδιαφέρουσες συσχετίσεις των κινητών που βρίσκονται στο ίδιο εύρος τιμής. Είμαστε σε θέση δηλαδή, να βγάλουμε ένα γενικό συμπέρασμα για τα χαρακτηριστικά των κινητών και να απαντήσουμε αν είναι τελικά καλά για την τιμή τους ή όχι.

# 5. Πειραματικά Αποτελέσματα

Γενικά, πραγματοποιήθηκαν πολλές πειραματικές εκτελέσεις μέχρι να φτάσουμε στις τελικές αυτές τιμές των παραμέτρων για τον αριθμό των clusters στον K-Means και των minimum support και threshold για το Association Rule Mining. Στην αρχή για τον K-Means, χρησιμοποιήσαμε k = 3 διαχωρίζοντας δηλαδή το εύρος τιμών σε “low-end”, “middle-range” και “high-end” phones όπως διαχωρίζονται με αυτήν την ονομασία στην αγορά γενικότερα. Αλλά μετά παρατηρήσαμε ότι υπάρχουν κάποια outliers (που έχουν δηλαδή τιμή ανάμεσα στα 5000€ με 7000€) και κάναμε τελικώς clustering για k = 4 θεωρώντας μια εξτρά κατηγορία κινητών “luxury” phones που έχουν δηλαδή εξωφρενικά ακραίες τιμές σε σύγκριση με τα υπόλοιπα κινητά.

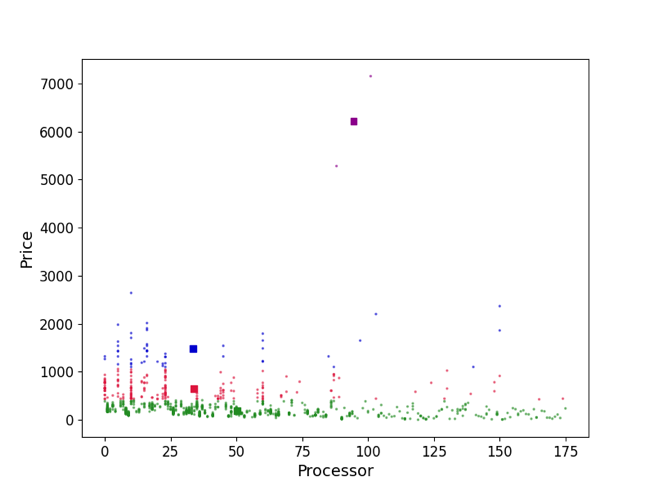
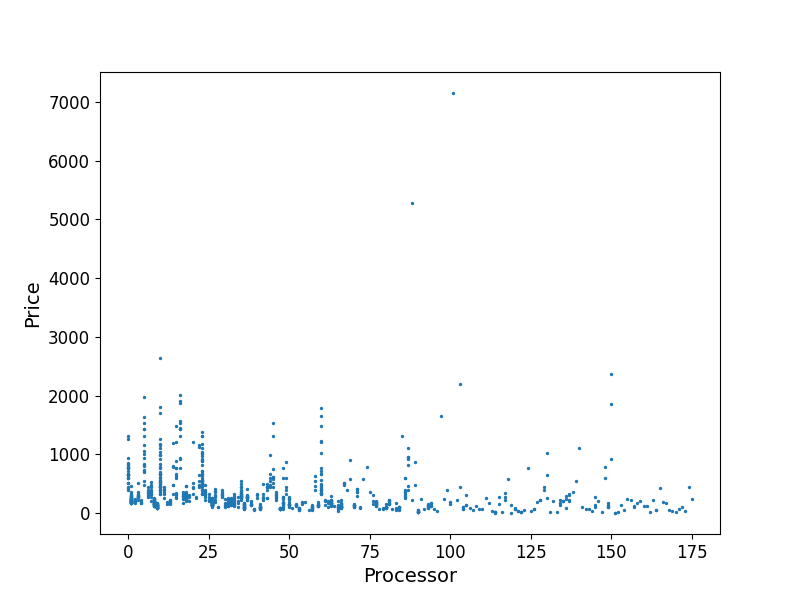
Για τις παραμέτρους min support και threshold έχουμε θέσει σαν default τις τιμές 0.2 και 0.6 αντίστοιχα οι οποίες είναι σχετικά χαμηλές για την εξαγωγή περισσότερων αποτελεσμάτων/συσχετίσεων. Γενικά, όσο χαμηλώνουμε την παράμετρο για το minimum support τόσο αυξάνεται ο αριθμός των frequent itemsets και συνεπώς ο αριθμός των συσχετίσεων. Αντίστοιχα, αν αυξήσουμε το minimum threshold τότε λαμβάνουμε πιο ισχυρές και ενδιαφέρουσες συσχετίσεις καθώς «φιλτράρονται» οι αδύναμες συσχετίσεις και συνεπώς έχουμε μικρότερο αριθμό συσχετίσεων αλλά πιο ουσιαστικό. Συνεπώς, αλλάζοντας αυτές τις τιμές λαμβάνουμε λιγότερα ή περισσότερα αποτελέσματα.

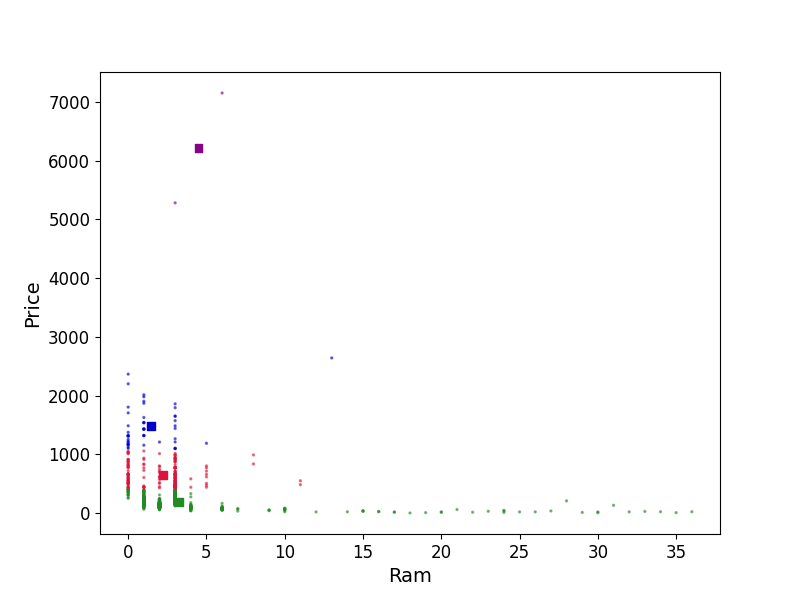
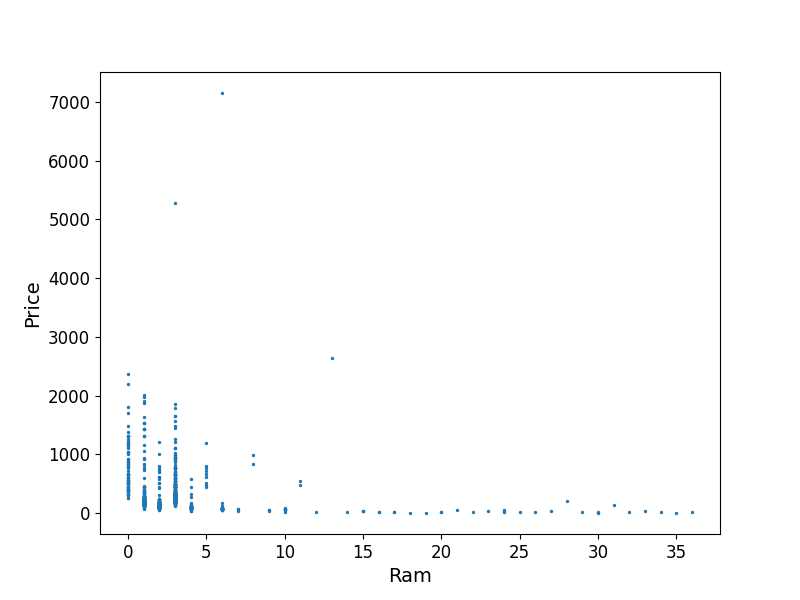
Σε γενικές γραμμές ανάλογα σε ποιο cluster κινητών βρισκόμασταν, προσπαθήσαμε να αλλάξουμε αυτές τις δύο παραμέτρους για να βγαίνουν και αρκετά αποτελέσματα αλλά και αρκετές σχέσεις. Για παράδειγμα, στο cluster για τα «luxury phones» (όπως φαίνεται και στο Σχήμα 2, 4) έχει μικρό αριθμό εγγραφών είναι προτιμητέο να έχεις τις default τιμές γιατί γενικά οι συσχετίσεις που θα βγουν θα είναι συχνές (0.5 και άνω) και ισχυρές λόγω του χαμηλού πλήθους των στοιχείων. Αντίθετα, στο cluster για τα low-end phones επειδή έχει τις περισσότερες εγγραφές (όπως φαίνεται πάλι από το Σχήμα 2, 4) αν έχουμε τις default τιμές δεν θα παρθούν πολλά αποτελέσματα γιατί θα είναι πιο μικρό το support για τα frequent itemsets καθώς και το confidence των συσχετίσεων.

Αυτό συμβαίνει και λόγω του πλήθους όπως είπαμε, αλλά επίσης η στήλη Processor π.χ. για την στήλη τύπου processor και clock speed έχει πολύ συγκεκριμένες τιμές (συγκεκριμένα μοντέλα επεξεργαστών και συγκεκριμένες τιμές clock speed) και στην RAM (αν και γενικά οι τιμές τις RAM σε κινητά αν και είναι αριθμοί είναι συγκεκριμένες 4GB, 6GB, 8GB όπως και ο χώρος αποθήκευσης κτλ.). Συνεπώς, προσπαθήσαμε να προσαρμόσουμε κάθε φορά τις τιμές για να έχουμε μια γκάμα τιμών και συσχετίσεων για τα χαρακτηριστικά κάθε κινητού και ας μην είναι τόσο ισχυρές οι συσχετίσεις κάθε φορά.

Οι χρόνοι εκτελέσεις γενικά λόγω ότι δεν είναι εξωφρενικά μεγάλο του dataset ήταν αρκετά μικροί. Δηλαδή συγκριτικά μεγάλοι ήταν οι χρόνοι εκτέλεσης για το Clustering της Price με τον Processor και της RAM και με τις δύο να είναι περίπου στα 30 δευτερόλεπτα η καθεμιά. Αν ήταν αρκετά μεγαλύτερο το dataset οι χρόνοι εκτέλεσης θα ήταν πιθανόν πολύ μεγαλύτεροι λόγω της μεγαλύτερης πολυπλοκότητας του clustering.

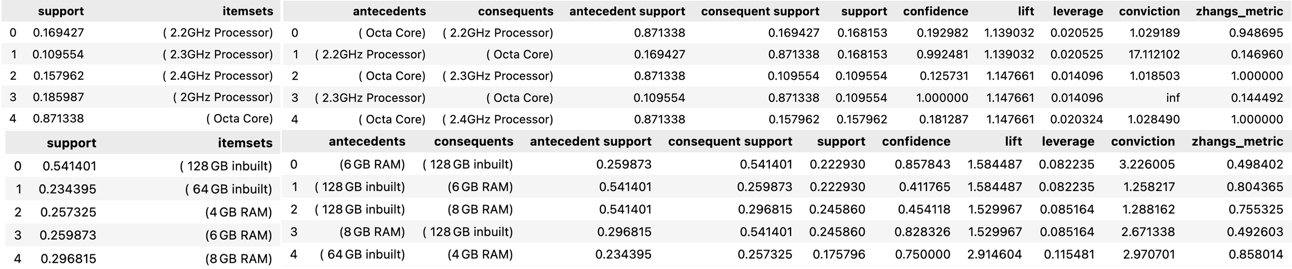
Στις παρακάτω εικόνες, βλέπουμε τα δύο γραφήματα για το Price με την Processor και την RAM αντίστοιχα και το αποτέλεσμα των Clustering και των δύο (k = 4). Επίσης, παραθέτονται και ενδεικτικά οι πρώτες 5 εγγραφές των πινάκων για τα frequent itemsets και τα Association Rules όσον αφορά τα τέσσερα cluster για την ανάλυση Processor και για την ανάλυση RAM αντίστοιχα.

**

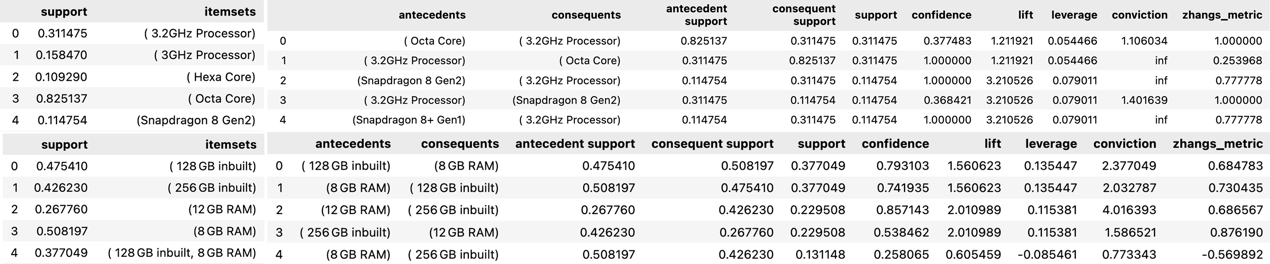
**

*Εικόνα 1: Απλό scatter γράφημα (Price – Processor) Εικόνα 2: Γράφημα Cluster (Price – Processor)*

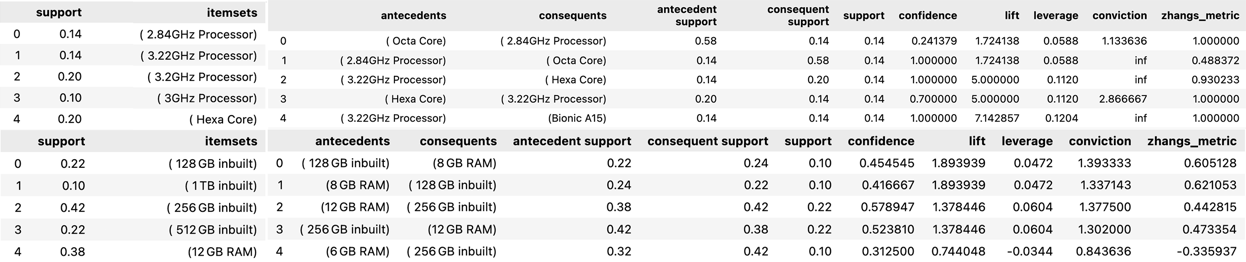
*Εικόνα 3: Απλό scatter γράφημα (Price – RAM) Εικόνα 4: Γράφημα Cluster (Price – RAM)*

**

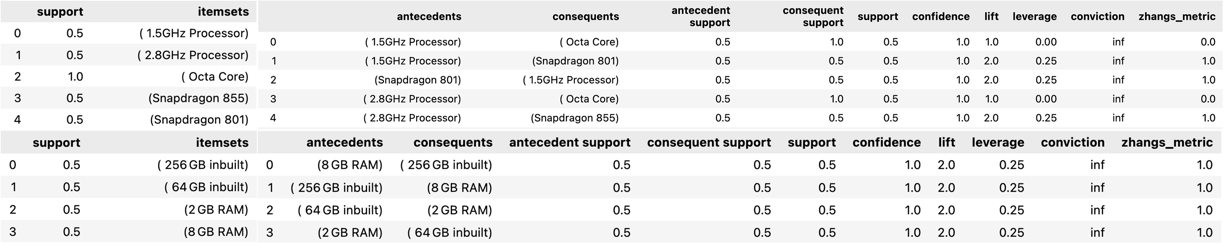
*Εικόνα 5: Association Rule Mining (Processor, RAM) για Low-End smartphones*

**

*Εικόνα 6: Association Rule Mining (Processor, RAM) για Middle Range smartphones*

**

*Εικόνα 7: Association Rule Mining (Processor, RAM) για High-End smartphones*

**

*Εικόνα 8: Association Rule Mining (Processor, RAM) για Luxury smartphones*

# 6. Κριτική αποτίμηση αποτελεσμάτων

ΦΡΟΣΩ για price processor

ΑΓΓΕΛΟΣ: για price ram (συνοπτικά για κάθε cluster, ποια είναι τα frequent items και οι συσχετίσεις του (συνοπτικές αυτές) και αν βγάζει νόημα να έχουν αυτά από προσωπική άποψη)

# 7. Συμπεράσματα

Τι συμπεράσματα βγάζουμε για τα διαφορετικά είδους κινητά (low-end, middle-end, higher-end, luxury). Απο τα αποτελέσματα ποιο είναι το καλύτερο? Σε σύγκριση με την τιμή δηλαδή πάντα. Γενικά αν εμπίπτουν τα αποτελέσματα με την πραγματικότητα… Αρκετά συνοπτικά!