ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ

Ανάλυση Δεδομένων

-ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΗΛΩΝ-

(ΕΡΓΑΣΙΑ 10)

Διδάσκοντες: Ι. Ντζούφρας - Κ Πατέρας

Φοιτητής:

Ονοματεπώνυμο: Παπακίτσος Ανδρέας

Αριθμός Μητρώου: 3200140

Έτος Σπουδών: 4ο

Περιεχόμενα

1 Εισαγωγή – περιγραφή μελέτης και προβλήματος

Η χρήση της ανάλυσης δεδομένων σε οποιονδήποτε επιστημονικό και μη κλάδο διευρύνεται όλο και περισσότερο, με σκοπό την βελτιστοποίηση του αποτελέσματος. Είτε επιδιώκοντας τον χαμηλότερο δυνατό χρόνο παραγωγής, είτε την μεγαλύτερη δυνατή ποιότητα, οι παραγωγοί προιόντων αναζητούν συχνά την λύση στα δεδομένα και τις απαντήσεις που μπορεί να δώσει η διερέυνησή τους. Στην αγορά των μήλων, κύριος σκοπός είναι η καλλιέργεια προιόντων με κατάλληλη γεύση ώστε να είναι προτιμότερα από τους καταναλωτές.

Στόχος της μελέτης αυτής είναι η αξιολόγηση της σχέσης των χαρακτηριστικών ενός μήλου, όπως το βάρος, ο βαθμός ωρίμανσης και το σχήμα, με τις ιδιότητες της γεύσης του. Σε περίπτωση που βρεθεί μια τέτοια συσχέτιση, η επιλογή των καταλληλότερων προιόντων για προώθηση θα γίνεται πολυ γρηγορότερα, κρίνοντας μόνο από τις διαστάσεις και το χρώμα. Το σετ δεδομένων που θα χρησιμοποιήσουμε αποτελείται από 4000 παρατηρήσεις και 9 μεταβλητές που φάινονται παρακάτω.

Πίνακας 1: Πίνακας Δεδομένων

Αριθμός Μεταβλητής	Όνομα Μεταβλητής	Τύπος Μεταβλητής	Σημασία	Τιμές
1	A_id	αριθμητική	Μοναδικός αναγνωριστικός για κάθε μήλο	ακέραιοι
2	Size	αριθμητική	Μέγεθος του μήλου	πραγματικές
3	Weight	αριθμητική	Βάρος του μήλου	πραγματικές
4	Sweetness	αριθμητική	Βαθμός γλυκύτητας του μήλου	πραγματικές
5	Crunchiness	αριθμητική	Υφή που υποδηλώνει την τραγανότητα του μήλου	πραγματικές
6	Juiciness	αριθμητική	Επίπεδο ζουμερότητας μήλου	πραγματικές
7	Ripeness	αριθμητική	Στάδιο ωρίμανσης του μήλου	πραγματικές
8	Acidity	κατηγορική	Επίπεδο οξύτητας του μήλου	πραγματικές
9	Quality	κατηγορική	Συνολική ποιότητα του μήλου	bad, good

2 Περιγραφική Ανάλυση

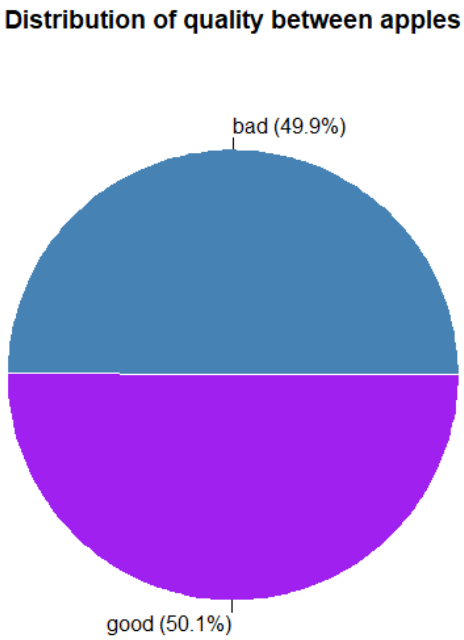
Για την υλοποίηση της Ανάλυσης θα χρησιμοποιήσουμε την Προγραμματιστική Γλώσσα R, σχεδιασμένη ειδικά για στατιστική ανάλυση. Με την χρήση της θα πραγματοποίησουμε τους απαραίτητους ελέγχους και μετασχησματισμούς δεδομένων, εξάγοντας τελικά και τα διαγράμματα. Αρχικά πρέπει να επεξεργαστούμε το σετ δεδομένων για να είναι όσο το δυνατόν ακριβέστερο το αποτέλεσμα της μελέτης. Βλέπουμε ότι από τις 4000 παρατηρήσεις που είχε αρχικά, η μία είναι πλήρως κενή (NA) οπότε την αφαιρούμε. Έπειτα παρατηρούμε ότι ο τύπος της μεταβλητής Acidity είναι «χαρακτήρας» ενώ περιέχει αριθμητικά δεδομένα. Επομένως τελικά έχουμε 1 κατηγορική μεταβλητή (Quality) και 8 αριθμητικές (A_id, Size, Sweetness, Acidity, Juiciness, Ripeness, Weight, Crunchiness). Η Quality δέχεται μόνο 2 τιμές, bad και good, ενώ χρειάστηκε να την μετατρέψουμε σε κατηγορική από χαρακτήρα.

Κοιτάζοντας κάθε μεταβλητή ξεχωριστά θα μπορέσουμε να δούμε τις ιδιότητες των τιμών τους και να αντλήσουμε περισσότερες πληροφορίες. Για την κατηγορική μεταβλητή, μπορούμε να δούμε την συχνότητα εμφάνισης της κάθε κατηγορίας που είναι μοιρασμένη σχεδόν ίσα, όπως φαίνεται στο Σχήμα 1. Η A_id αποτελεί μοναδικό αναγνωριστικό των μήλων, οπότε δεν προσφέρει κάποια πληροφορία και θα την αγνοήσουμε κατά την μελέτη. Οι υπόλοιπες ποσοτικές μεταβλητές, παρατηρούμε πως έχουν κάποιες ιδιαιτερότητες. Αρχικά, φαίνεται ότι λαμβάνουν αρνητικές τιμές, γεγονός φαινομενικά ανησυχητικό για μεταβλητές όπβς το βάρος και το μέγεθος. Ωστόσο αυτό δεν προέρχεται από λανθασμένες τιμές στα δεδομένα, αλλα από κανονικοποίηση που έχουν δεχθεί νωρίτερα. ¨Ετσι, είναι όλες στην ίδια κλίμακα και μάλιστα πιθανώς ακολουθούν κανονική κατανομή, όπως φαίνεται από τις τιμές της κύρτωσης ή της ασυμμετρίας στον Πίνακα 2 και τα QQ-plots. Πραγματοποιώντας ελεγχους Shapiro-Wilks και Kolmogorov-Smirnov βλέπουμε ότι κάποιες υποθέσεις απορρίπτονται, πιθανότατα λόγω του μεγάλου δείγματος (n=4000), μιας συνθήκης στην οποία είναι ευαίσθητα αυτά τα tests. Για την συνέχεια της μελέτης θα υποθέσουμε κανονικότητα όλων των μεταβλητών.

Πίνακας 2: Περιγραφικά Μέτρα Μεταβλητών

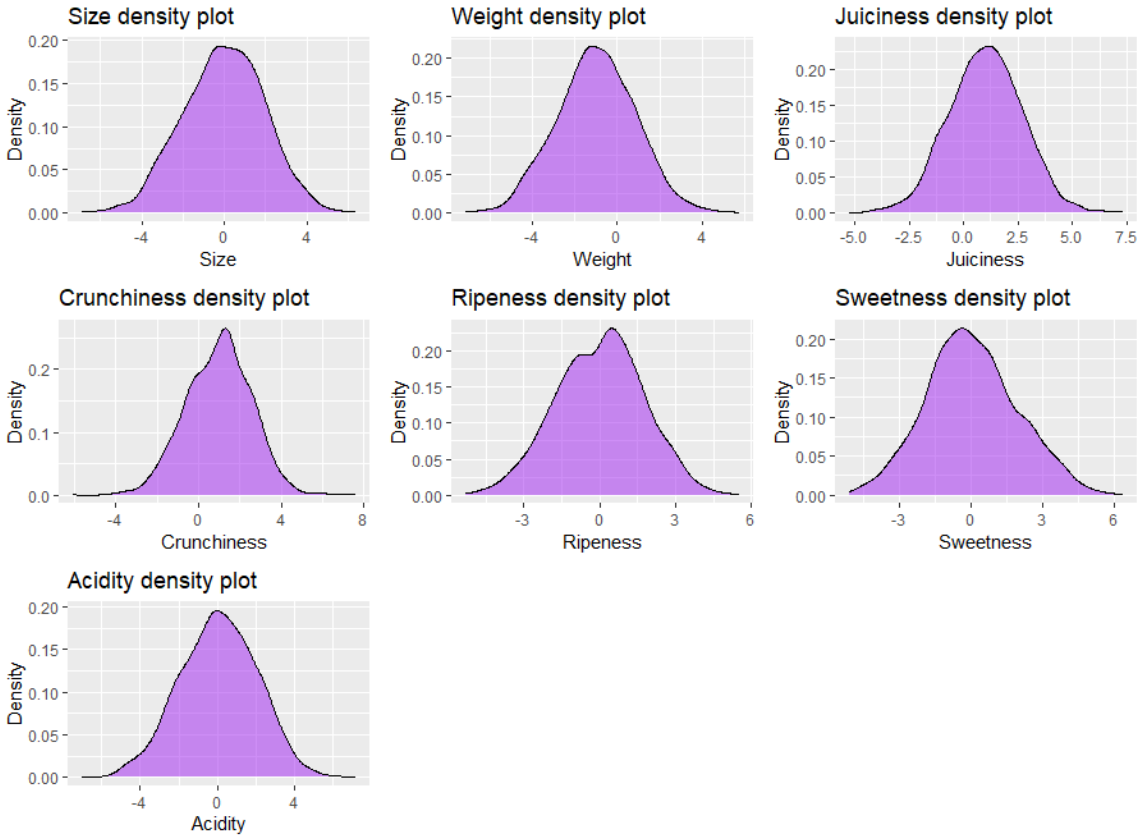
Όνομα Μεταβλητής	Μέσο	Τυπική Απόκλιση	Διάμεσος	Μικρότερη Τιμή	Μεγαλύτερη Τιμή	Εύρος	Ασυμμετρία	Κύρτωση
size	-0.50	1.93	-0.51	-7.15	6.41	13.56	0	-0.09
weight	-0.99	1.60	-0.98	-7.15	5.79	12.94	0	0.36
sweetness	-0.47	1.94	-0.5	-6.89	6.37	13.27	0.08	0.01
crunchiness	0.99	1.40	1	-6.06	7.62	13.67	0	0.72
juiciness	0.51	1.93	0.53	-5.96	7.36	13.33	-0.11	0.03
ripeness	0.50	1.87	0.5	-5.86	7.24	13.10	-0.01	-0.07
acidity	0.08	2.11	0.02	-7.01	7.4	14.42	0.06	-0.1

Διάγραμμα 1: Συχνότητες εμφάνισης κακών/καλών μήλων



Αν περιοριστούμε μόνο στα μήλα κατανεμημένα στην κατηγορία “good” της Quality, η οποία και είναι αυτή που μας ενδιαφέρει περισσότερο, παρατηρούμε πως οι υπόλοιπες μεταβλητές εξακολοθούν να είναι κατανονικά κατανεμημένες γενικά (Διάγραμμα 2) . Το πλήθος των καλών μήλων είναι πάλι αρκετά μεγάλο (2004) ώστε να μην μας εμποδίσει κάποια απώλεια στην κανονικότητα.

Διάγραμμα 2 : Πυκνότητα μεταβλητών καλών μήλων



3 Σχέσεις Μεταβλητών ανά δύο

Η πληροφορία που προσπαθούμε να αντλήσουμε βρίσκεται στην σχέση μεταξύ των μεταβλητών. Αρχικά, πρέπει να ελέγξουμε αν υπάρχει κάποια συσχέτιση σε οποιοδήποτε ζευγάρι ποσοτικών μεταβλητών. Σύμφωνα με την μέθοδο συσχέτισης του Pearson δεν φαίνεται να υπάρχει κάποια δυνατή συσχέτιση σε οποιοδήποτε συνδυασμό. Στην συνέχεια πρέπει να κοιτάξουμε τις πιθανές σχέσεις μεταξύ των ποσοτικών μεταβλητών και της κατηγορικής Quality. Η μεταβλητή Sweetness εμφανίζει την ισχυρότερη σχέση του dataset με την μεταβλητή Size, χωρίς ωστόσο να είναι πολύ δυνατή, ενώ λιγότερο ισχυρές σχέσεις έχει με το βάρος και τον βαθμό ωρίμανσης (βλέπε Πίνακα 3). Ενδιαφέρον έχει η εξέταση της σχέσης Sweetness~Quality, μιας και ο t.test έλεγχος έδειξε πως υπάρχει σημαντική διαφορά στην μέση τιμή γλυκήτητας ανάμεσα στα καλά και στα κακά μήλα (p-value < 2.2e-16) .

Πίνακας 3: Συσχέτιση Pearson Ποσοτικών μεταβλητών

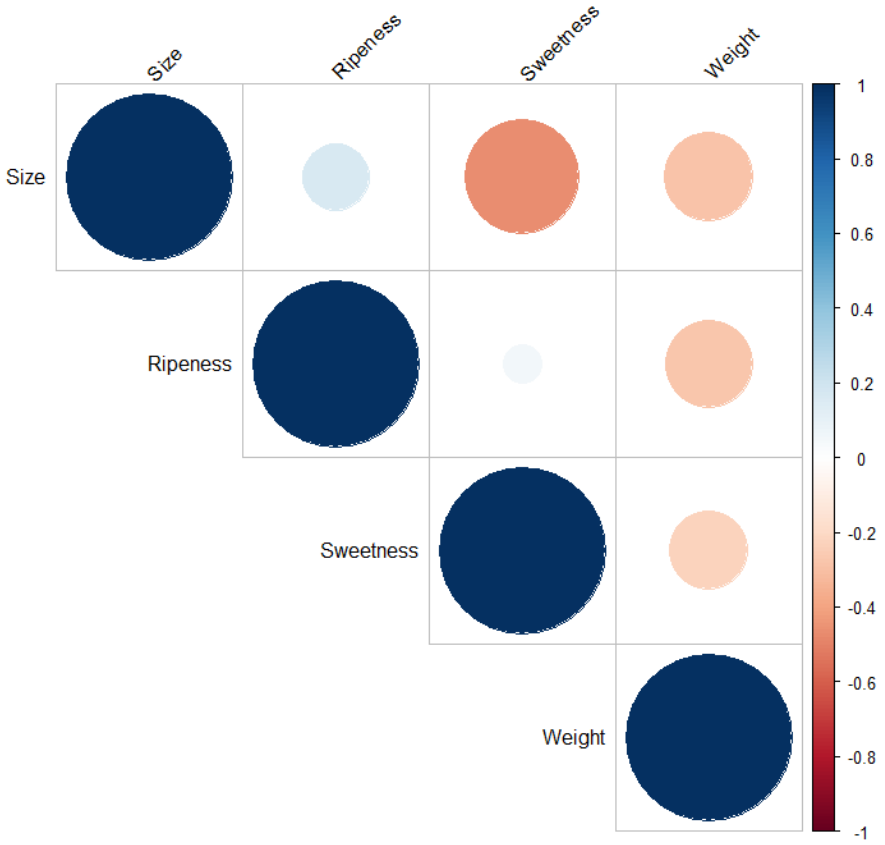
Συσχέτιση μεταβλητών Ανά 2

Συνδυασμός	*Συντελεστής Συσχέτισης*
Size~Weight	-0.17
Size~Sweetness	-0.32
Size~Crunchiness	0.17
Size~Juiciness	-0.02
Size~Ripeness	-0.13
Size~Acidity	0.20
Weight~Sweetness	-0.15
Weight~Crunchiness	-0.10
Weight~Juiciness	-0.09
Weight~Ripeness	-0.24
Weight~Acidity	0.02

Συνδυασμός	*Συντελεστής Συσχέτισης*
Sweetness~Crunchiness	-0.04
Sweetness~Juiciness	0.10
Sweetness~Ripeness	-0.27
Sweetness~Acidity	0.09
Crunchiness~Juiciness	-0.26
Crunchiness~Ripeness	-0.20
Crunchiness~Acidity	0.07
Juiciness~Ripeness	-0.10
Juiciness~Acidity	0.25
Ripeness~Acidity	-0.20

Λαμβάνοντας μόνο τα καλά μήλα υπόψιν, η σχέση Sweetness~Size φαίνεται ισχυρότερη, ενώ η εξάρτηση της Γλυκύτητας από το στάδιο ωρίμανσης φαίνεται να εξαφανίζεται τελείως. Η σχέση Sweetness~Weight είναι επίσης ισχυρότερη από ότι πριν. Συνεπώς η πιο σημαντική σχέση για τα μήλα γενικότερα είναι η Size~Sweetness (Pearson's correlation = -0.32), που για τα καλά μήλα είναι ακόμα ισχυρότερη (Pearson's correlation = -0.47).

Διάγραμμα 3:Συσχέτιση Pearson για καλά μήλα



4 Προβλεπτικά ή ερμηνευτικά μοντέλα

Όπως είδαμε προηγουμένως, η υψηλότερη γλυκύτητα πιθανώς να οδηγεί σε μεγαλύτερη ποιότητα ή το αντρίστροφο. Στόχος μας πλέον είναι η προσπάθεια εκτίμησης της γλυκητήτας μέσω τιμών για κάποιες από τις άλλες μεταβλητές. Πρώτα θα δοκιμάσουμε τα απλά γραμμικά μοντέλα με κάθε μια από τις αριθμητικές μεταβλητές.

Πίνακας 4 : Περιγραφή γραμμικών μοντέλων μίας μεταβλητής για την εκτίμηση της Sweetness

	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.64	-0.69 – -0.58	**<0.001**
Size	-0.33	-0.36 – -0.30	**<0.001**
Observations	4000		
R^2 / R^2 adjusted	0.105 / 0.105		

	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.48	-0.54 – -0.42	**<0.001**
Acidity	0.08	0.05 – 0.11	**<0.001**
Observations	4000		
R^2 / R^2 adjusted	0.007 / 0.007		

	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.66	-0.73 – -0.59	**<0.001**
Weight	-0.19	-0.22 – -0.15	**<0.001**
Observations	4000		
R^2 / R^2 adjusted	0.024 / 0.024		

	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.42	-0.49 – -0.35	**<0.001**
Crunchiness	-0.05	-0.09 – -0.01	**0.018**
Observations	4000		
R^2 / R^2 adjusted	0.001 / 0.001		

	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.33	-0.39 – -0.27	**<0.001**
Ripeness	-0.28	-0.31 – -0.25	**<0.001**
Observations	4000		
R^2 / R^2 adjusted	0.075 / 0.075		

	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.52	-0.58 – -0.46	**<0.001**
Juiciness	0.10	0.07 – 0.13	**<0.001**
Observations	4000		
R^2 / R^2 adjusted	0.009 / 0.009		

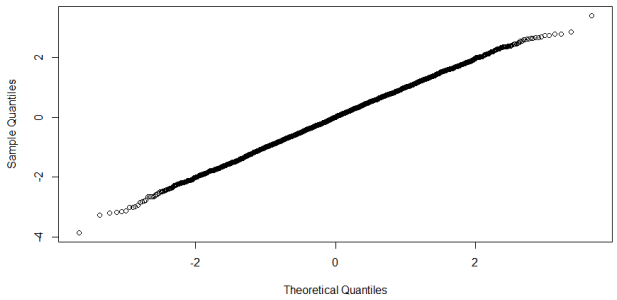
Βλέπουμε στον Πίνακα 4 ότι το πρώτο μοντέλο Sweetness = -0.64 -0.33* Size έχει την καλύτερη προσαρμοή, χωρίς ωστόσο να είναι επαρκής για την επίλυση του προβλήματος. Καθώς όλες οι μεταβλητές φαίνονται στατιστικά σημαντικές για την πρόβλεψη της γλυκύτητας, (με p-value<0.001) θα συνεχίσουμε με το full μοντέλο.

Το μοντέλο είναι Sweetness = β_0 + β_1*Size + β_2*Weight + β_3*Acidity + β_4*Ripeness + β_5 * Juiciness + β_6* Crunchiness + β_7*(Quality=good). Εφαρμόζοντας stepwise επιλογή με BIC το καλύτερο ήταν το αρχικό (BIC = 3291.34) και στον Πίνακα 5 φαίνονται οι παράγοντες του. Η ερμηνεία του μοντέλου σύμφωνα με την πραγματική έννοια των μεταβλητών είναι παραπλανητική, καθώς τα δεδομένα λαμβάνουν εξαρχής αρνητικές τιμές, όντας επεξεργασμένα. Η σταθερά β_0 είναι -1.44 και δείχνει την τιμή του μήλου που έχει 0 σε όλες τις μεταβλητές. Ο συντελεστής β_1 δείχνει την αύξηση που θα δεχθεί η γλυκύτητα αν αυξηθεί κατά μία μονάδα το μέγεθος. Ο συντελεστής β_2 δείχνει την αύξηση που θα δεχθεί αν αυξηθεί κατά μία μονάδα το βάρος. Αντίστοιχα οι συντελεστές β_3,β_4,β_5 και β_6 δείχνουν την αύξηση της γλυκήτητας αν αυξηθεί κατα μία μονάδα μία απο τις Acidity,Ripeness,Juiciness, Crunchiness. Σύμφωνα με τον β_7 τα καλά μήλα έχουν 1.26 παραπάνω μονάδες στην Sweetness. Τα κατάλοιπα είναι κανονικά κατανεμημένα (βλέπετε Διάγραμμα 4) και το R^2 = 0.407 οπότε η προσαρμογή στα δεδομένα φαίνεται να είναι καλή.

Πίνακας 5: Τελικό μοντέλο εκτίμησης της Sweetness

	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-1.44	-1.53 – -1.35	**<0.001**
Size	-0.53	-0.56 – -0.51	**<0.001**
Weight	-0.43	-0.46 – -0.40	**<0.001**
Acidity	0.15	0.13 – 0.18	**<0.001**
Ripeness	-0.36	-0.38 – -0.33	**<0.001**
Juiciness	-0.13	-0.16 – -0.10	**<0.001**
Crunchiness	-0.13	-0.16 – -0.09	**<0.001**
Quality [good]	1.26	1.16 – 1.36	**<0.001**
Observations	4000		
R^2 / R^2 adjusted	0.407 / 0.406		

Διάγραμμα 4: QQ-plot καταλοίπων του μοντέλου



Αν επαναλάβουμε την ίδια διαδικασία για τα μήλα κατηγοριοποιημένα ως καλά, το μοντέλο που προκύπτει ως καλύτερο είναι το Sweetness = β_0 + β_1*Size + β_2*Weight + β_3*Acidity + β_4*Ripeness + β_5* Crunchiness .. Το R^2 αυξήθηκε και τα κατάλοιπα παραμένουν κανονικά κατανεμημένα, ενώ η μεταβλητή Juiciness αποκλείστηκε.

Διάγραμμα 5: Μοντέλο εκτίμησης της Sweetness λαμβάνοντας υπόψιν μόνο τα καλά μήλα

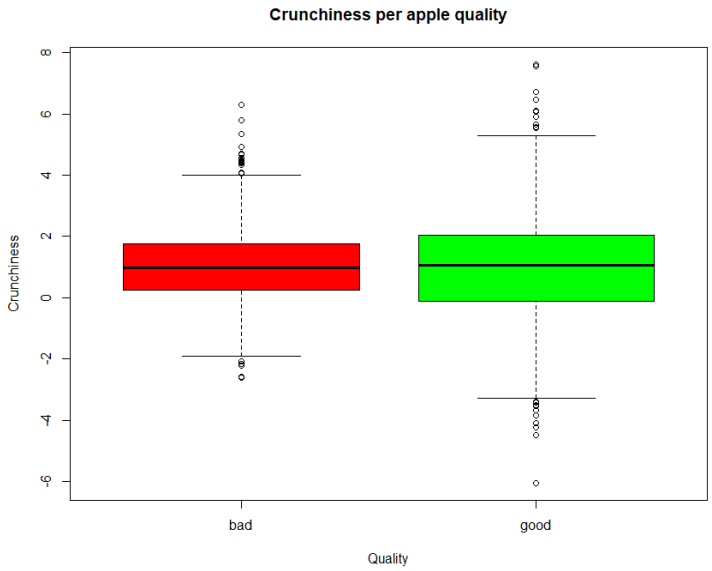
	Sweetness		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.25	-0.33 – -0.17	**<0.001**
Size	-0.70	-0.74 – -0.66	**<0.001**
Weight	-0.39	-0.43 – -0.35	**<0.001**
Acidity	0.27	0.24 – 0.31	**<0.001**
Ripeness	0.11	0.06 – 0.15	**<0.001**
Crunchiness	-0.17	-0.21 – -0.12	**<0.001**
Observations	2004		
R^2 / R^2 adjusted	0.434 / 0.432		

5 Συμπεράσματα και Συζήτηση

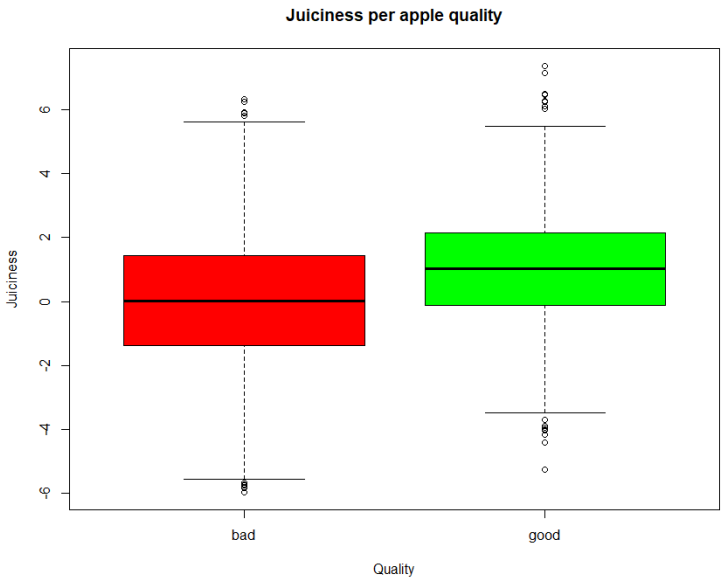
Μεταξύ των χαρακτηριστικών ενός μήλου υπάρχουν σχέσεις αρκετά ισχυρές για να μπορέσουμε να αντλήσουμε κάποιες πληροφορίες. Το μέτρο της γλυκύτητας της γέυσης του μπορεί να εκτιμηθεί με βάση κάποια παραπάνω στοιχεία για αυτό. Το μοντέλο δεν είναι ιδανικό, θα μπορούσε να υπάξει καλύτερη προσαρμογή στα δεδομένα, αλλά κάλυψε τις ανάγκες της έρευνας. Ένα καλό μήλο, είναι αρκετά πιο γλυκο ενώ η οξύτητα συμβάλλει θετικά στην πιο γλυκή γεύση. Το μέγεθος του μήλου είναι σημαντικό, ενώ από ότι φαίνεται πιο μικρά και ελαφριά μήλα είναι πιο γλυκά. Αντίστοιχη έρευνα μπορεί να πραγματοποιηθεί και σε άλλα αγροτικά προϊόντα, ώστε να αποκτηθεί βαθύτερη γνώση της γεύσης τους και της προτίμησης που έχουν οι άνθρωποι σε κάποια από αυτά.

Παράρτημα

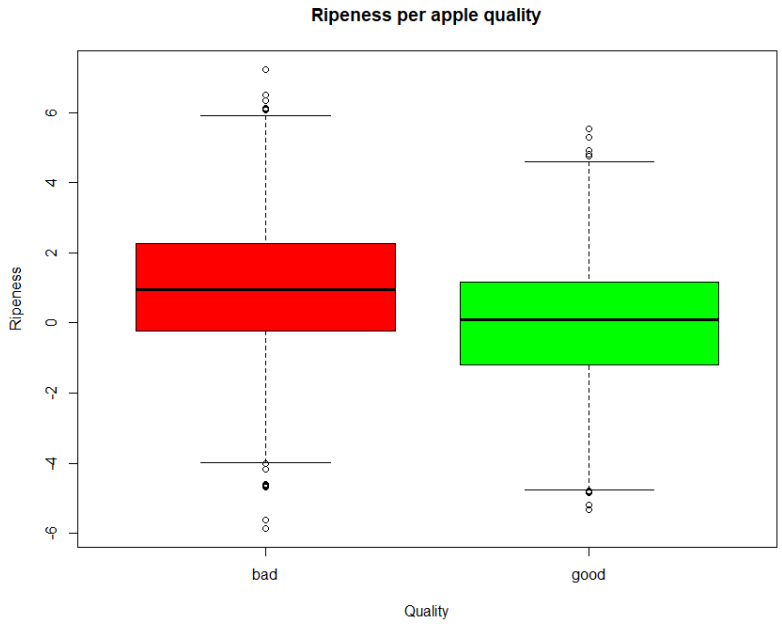
Διάγραμμα 7: Boxplots Crunchiness για κάθε κατηγορία της Quality



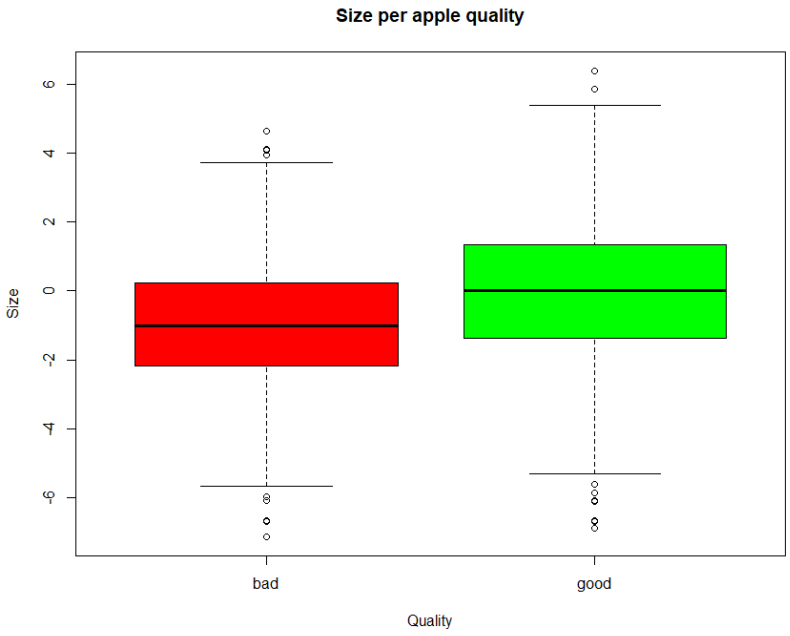
Διάγραμμα 6 : Boxplots Juiciiness για κάθε κατηγορία της Quality



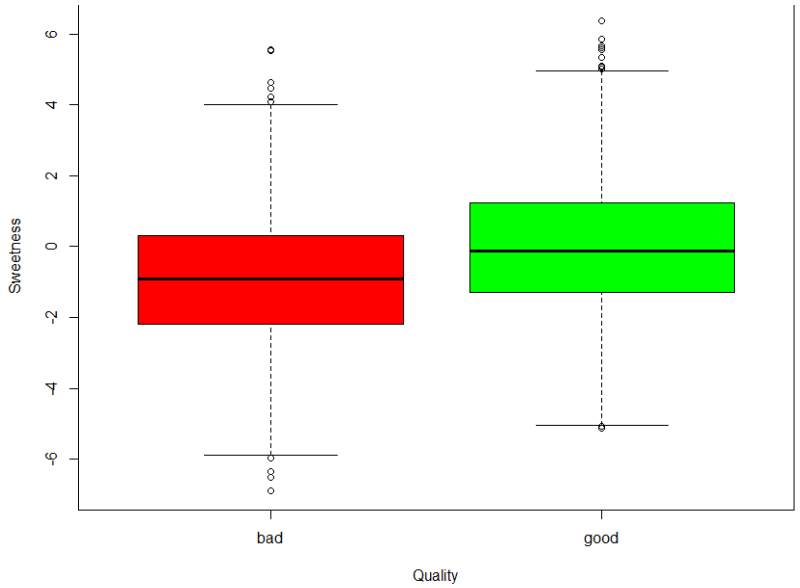
Διάγραμμα 8 : Boxplots Ripeness για κάθε κατηγορία της Quality



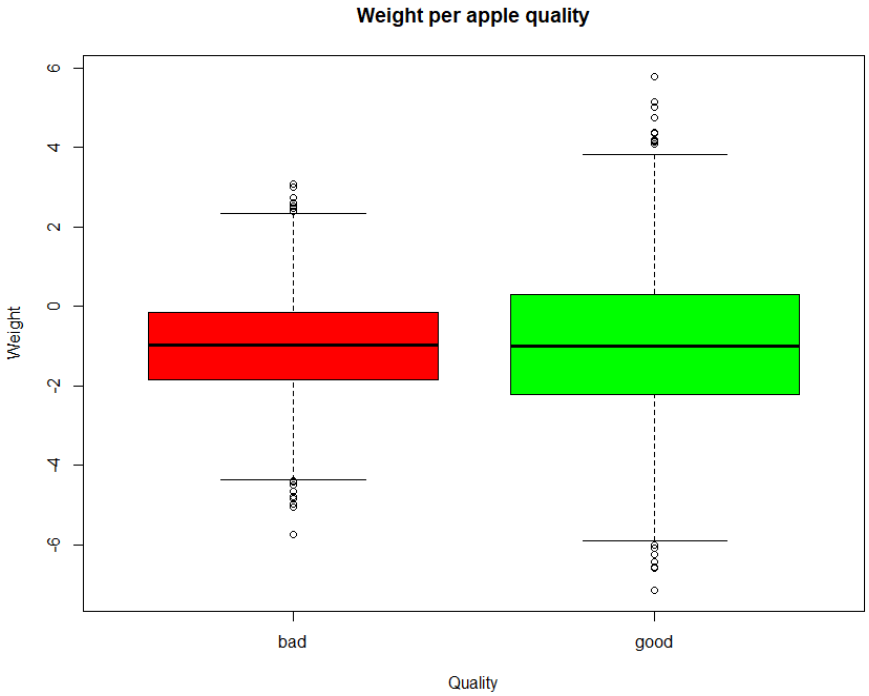
Διάγραμμα 9:Boxplots Size για κάθε κατηγορία της Quality



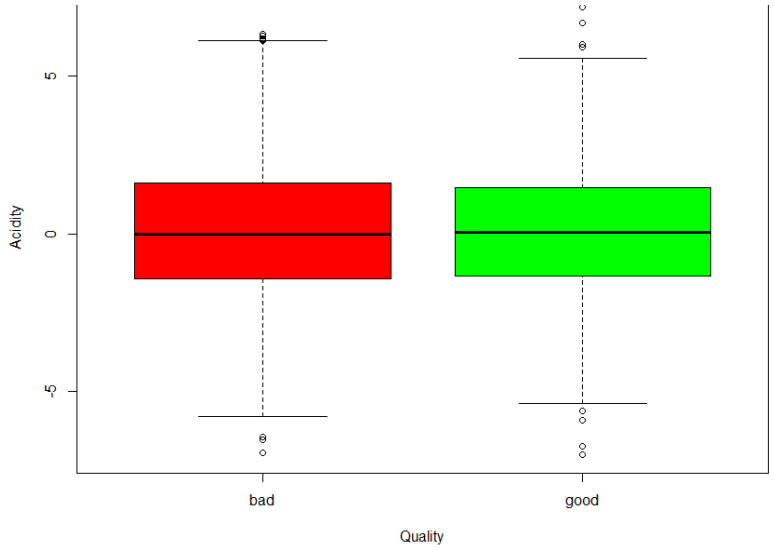
Διάγραμμα 11: Boxplots Sweetness για κάθε κατηγορία της Quality



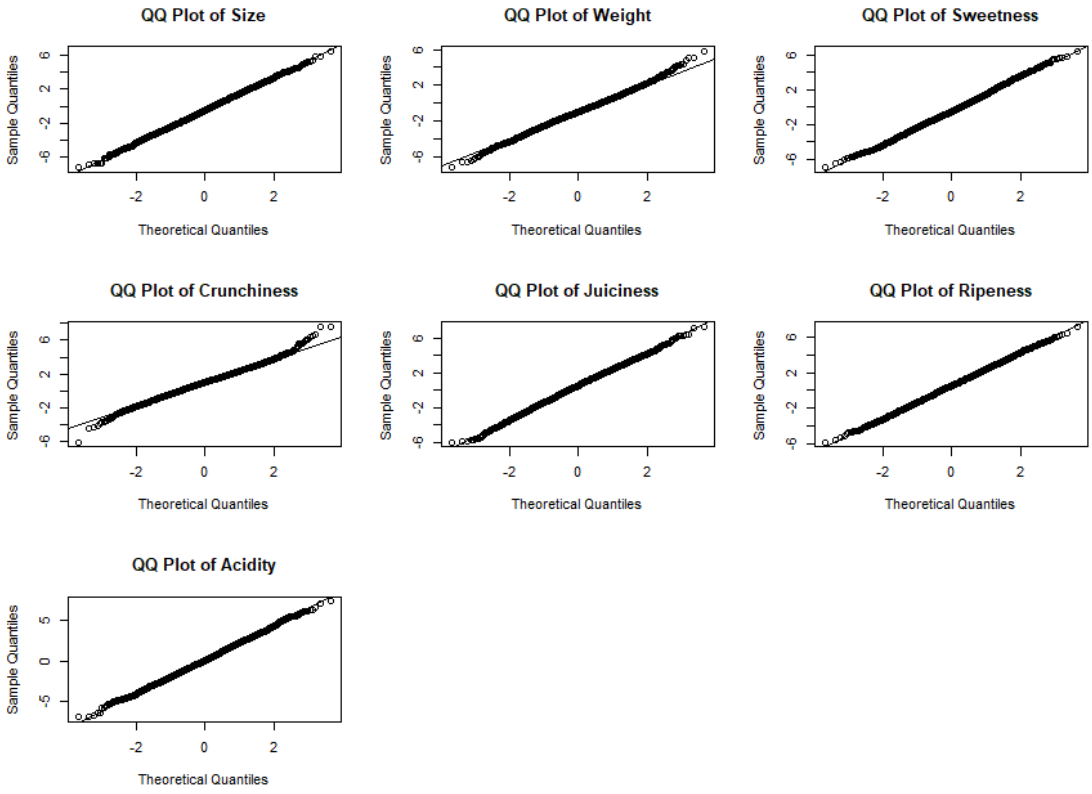
Διάγραμμα 10: Boxplots Weight για κάθε κατηγορία της Quality



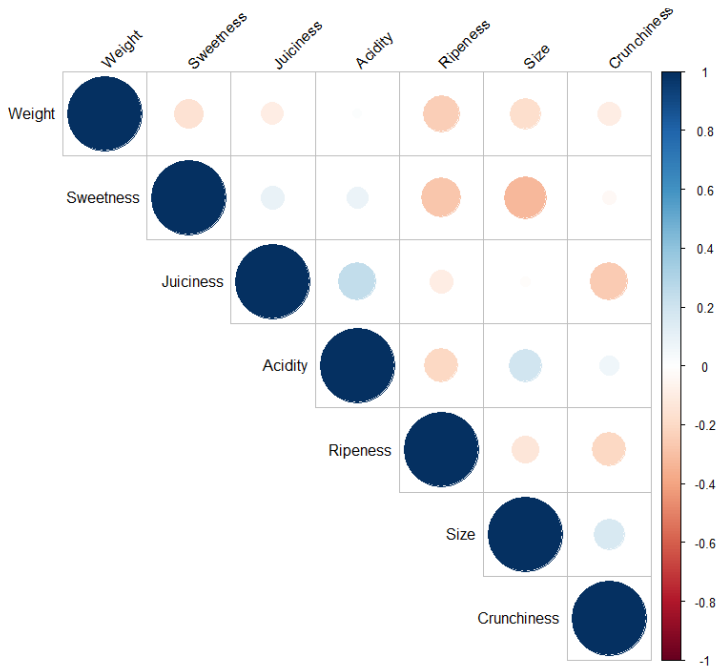
Διάγραμμα 12: Boxplots Acidity για κάθε κατηγορία της Quality



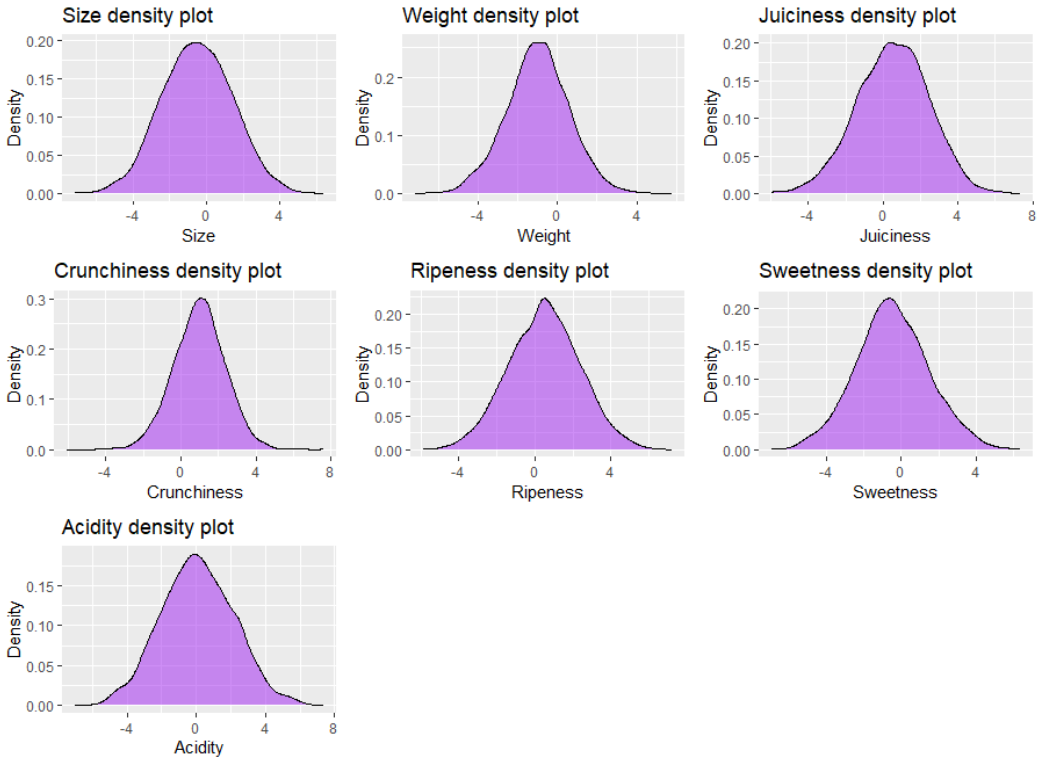
Διάγραμμα 13: QQ-plots κάθε μεταβλητής



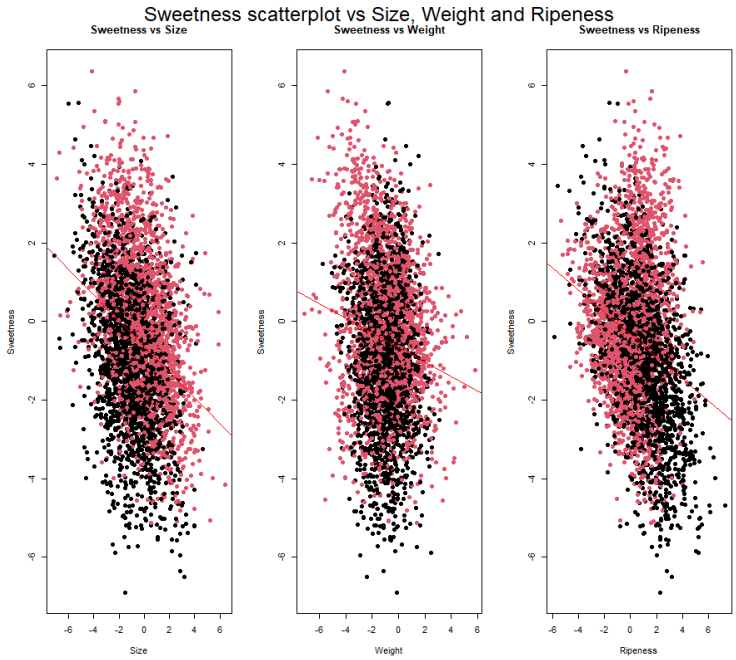
Διάγραμμα 15: Πίνακας Συσχέτισης Pearson όλων των αριθμητικών μεταβλητών



Διάγραμμα 14: Πυκνότητα της κάθε αριθμητικής μεταβλητής για το υποσύνολο των καλών μήλων



Διάγραμμα 16: Scatterplots της Sweetness με Size, Weight, Ripeness. Η γραμμή είναι γραμμική παλινδρόμηση



Διάγραμμα 17: Πυκνότητα καταλοίπων του πρώτου μοντέλου