ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

-Ανάλυση απόδοσης παικτών guard του NBA-

(Εργασία 9)

Διδάσκοντες: Ι.Ντζούφρας – Ξ.Πεντελή

Φοιτητής

Ονοματεπώνυμο: Προμπονάς Αντώνιος

Τμήμα: Πληροφορική

Αριθμός Μητρώου: 3190179

Έτος σπουδών: 4ο

Ημερομηνία: Ιούνιος , 2023

Περιεχόμενα

1. ΕΙΣΑΓΩΓΉ ΠΕΡΙΓΡΑΦΉ ΜΕΛΈΤΗ ΠΡΟΒΛΉΜΑΤΟΣ

Μία από τους σπουδαιότερες και πιο χρήσιμες τεχνικές που χρησιμοποιούνται στην επιστήμη του sports analytics είναι η ανάλυση δεδομένων. Η ανάλυση δεδομένων είναι ένα πολύτιμο εργαλείο τόσο στην εξόρυξη πληροφορίας όσο και στην ερμηνεία και κατανόηση των δεδομένων που συλλέγουμε σε μελέτες και πειράματα. Στην πραγματικότητα, η ανάλυση δεδομένων αποτελεί ένα σύνολο μεθόδων και τεχνικών που χρησιμοποιούνται για την ανακάλυψη προτύπων, τη διαπίστωση συσχετίσεων, την πρόβλεψη μελλοντικών γεγονότων και την εξαγωγή συμπερασμάτων από τα δεδομένα. Ο στόχος της είναι να αναδειχθούν μοτίβα και δομές στα δεδομένα που μπορούν να παράσχουν πληροφορίες για το πρόβλημα που μελετάμε.

Στην προκειμένη περίπτωση θέλουμε να χρησιμοποιήσουμε τις τεχνικές αυτές για να κάνουμε ανάλυση στα δεδομένα 105 καλαθοσφαιριστών που αγωνιζόντουσαν στο πρωτάθλημα του NBA, τη περίοδο 1992-1993. Το σετ δεδομένων που πρέπει να επεξεργαστούμε, περιέχει για κάθε παίκτη κάποιες συγκεκριμένες πληροφορίες. Οι πληροφορίες αυτές έχουν να κάνουν σχετικά με το ύψος του, τον αριθμό των αγώνων και τον συνολικό χρόνο σε λεπτά που αγωνίστηκε, την ηλικία του , τις assists, τα rebounds και τα ποσοστά των καλαθιών που πέτυχε τόσο εντός παιδιάς (2 πόντοι) όσο και από το σημείο των ελευθέρων βολών.

Ο σκοπός μας είναι να βρούμε μέσα από αυτά τα δεδομένα αυτά, ποιος ή ποιοι παίκτες έχουν τις καλύτερες προοπτικές για χρόνο συμμετοχής στους αγώνες, πάνω από 20 λεπτά. Στη προσπάθεια μας αυτή, θα μελετήσουμε τη σχέση που υπάρχει μεταξύ του χρόνου συμμετοχής, ηλικίας και επίδοσης του παίκτη σε πόντους. Με αυτό το τρόπο, είμαστε βέβαιοι ότι θα καταφέρουμε να εξάγουμε ένα όσο το δυνατό πιο ακριβή συμπέρασμα σχετικά με τον ποιο παίκτη αξίζει να προσπαθήσουμε να κάνουμε μεταγραφή.

Πίνακας 1: Πίνακας Δεδομένων

Αριθμός Μεταβλητής	Όνομα Μεταβλητής	Τύπος Μεταβλητής	Σημασία
1.	Player	Κατηγορική	Πως λέγεται ο παίκτης
2.	Height	Ποσοτική	Το ύψος του παίκτη σε cm
3.	Games	Ποσοτική	Το πλήθος των αγώνων που έχει αγωνιστεί ο παίκτης

4.	Minutes	Ποσοτική	Το άθροισμα των λεπτών συμμετοχής
5.	Age	Ποσοτική	Πόσο χρονών είναι ο παίκτης
6.	Points	Ποσοτική	Μέσος όρος πόντων ανά αγώνα
7.	Assists	Ποσοτική	Μέσος όρος τελικών πασών
8.	Rebounds	Ποσοτική	Μέσος όρος rebounds
9.	TPP	Ποσοτική	Ποσοστό καλαθιών που επιτεύχθηκαν εντός παιδιάς
10.	FT	Ποσοτική	Ποσοστό ευστοχίας καλαθιών από τις βολές
11.	Over20	Κατηγορική	Αν αγωνίστηκε ο παίκτης πάνω από 20 λεπτά
12.	PPM	Ποσοτική	Μέσος όρος πόντων ανά λεπτό
13.	Age4	Κατηγορική	Η ηλικία των αθλητών χωρισμένη σε 4 κατηγορίες

2. ΠΕΡΙΓΡΑΦΙΚΉ ΑΝΆΛΥΣΗ

Για να υλοποιήσουμε την ανάλυση μας χρησιμοποιούμε πακέτα και συναρτήσεις που μας προσφέρει το προγραμματιστικό περιβάλλον της γλώσσας R. Αφού εισάγουμε τα δεδομένα μας και ρίξουμε μια γρήγορη ματιά στις μεταβλητές μας , θα παρατηρήσουμε ότι δεν υπάρχει κάποια ξεκάθαρη μεταβλητή που να προσδιορίζει αν ο παίκτης αγωνίζεται πάνω από 20 λεπτά ή όχι. Για το σκοπό αυτό , δημιουργούμε τη μεταβλητή “over20” ,η οποία είναι μία δίτιμη μεταβλητή, καθώς οι μοναδικές τιμές που παίρνει είναι το 0 και το 1. Συγκεκριμένα, όταν παίκτης έχει τιμή μεταβλητής “over20” 0, αυτό σημαίνει ότι αγωνίζεται μέσο όρο κάτω από 20 λεπτά, ενώ αν έχει τιμή 1, αυτό σημαίνει ότι αγωνίζεται πάνω από 20 λεπτά. Για να ενισχύσουμε περισσότερο την ανάλυση μας , δημιουργούμε ακόμα 3 μεταβλητές: I)Την “avg_game_time” που περιέχει τον μέσο χρόνο συμμετοχής των αθλητών στους αγώνες , II) την ”PPM” η οποία περιέχει τον μέσο όρο πόντων των αθλητών ανά λεπτό και III) την “Age4”, η οποία είναι μία κατηγορική μεταβλητή που προήλθε από την ποσοτική μεταβλητή “Age” και περιέχει την ηλικία των αθλητών χωρισμένη σε 4 κατηγορίες. Τα δεδομένα διαμορφώνοντα όπως τα βλέπουμε στον πίνακα 1.

Συνεχίζοντας θέλουμε να εξετάσουμε κάθε μεταβλητή ξεχωριστά και να δούμε τι τιμές περιέχει. Για τις κατηγορικές μεταβλητές μπορούμε να δούμε τις συχνότητες με τις οποίες έχει εμφανιστεί η κατηγορία της κάθε μεταβλητής. Ενώ για τις μεταβλητές που είναι ποιοτικές μπορούμε να δούμε την μέση τιμή την τυπική απόκλιση ,τη διάμεσο την ασυμμετρία αλλά και την κύρτωση. Μέσα από τον πίνακα 2 , βλέπουμε ότι οι μεταβλητές "Minutes" και "TPP" είναι εκείνες που πλησιάζουν όσο καμία άλλη μεταβλητή την κανονική κατανομή, αφού έχουν ασυμμετρία -0.05 και -0.04 αντίστοιχα.

Μεταβλητές	Ελλιπές τιμές	Μέσο	Τυπική απόκλιση	Διάμεσος	Μικρότερη τιμή	Μεγαλύτερη τιμή	ασσυμετρία	κύρτωση
Points	0	10.7	6.058386	9.3	1.7	32.6	0.6346605	3.163343
Minutes	0	1656.486	885.6119	1622	55	3117	-0.0585977	1.887444
Assists	0	3.99619	2.490016	3.3	0.3	3.99619	0.8833885	3.198423
Rebounds	0	2.395238	1.287468	2.2	0.3	6.7	0.7311241	3.501571
Height	0	190.1714	7.074408	191	160	210	-0.9100422	5.907576
Games	0	64.15238	19.05901	73	8	82	-1.167172	3.543449
TPP	0	45.13238	4.39218	45.3	34.8	59.5	-0.0432467	3.537603
FT	0	77.85524	9.176443	79.3	37.5	94.8	-1.575137	6.83902
Age	0	27.53333	3.368596	27	22	37	0.4776691	2.454753
avg_game_time	0	24.30325	9.857443	24.2766	4.35	40.71053	-0.2103572	1.927996
PPM	0	0.2675	0.1514596	0.2325	0.0425	0.815	0.6346605	3.163343

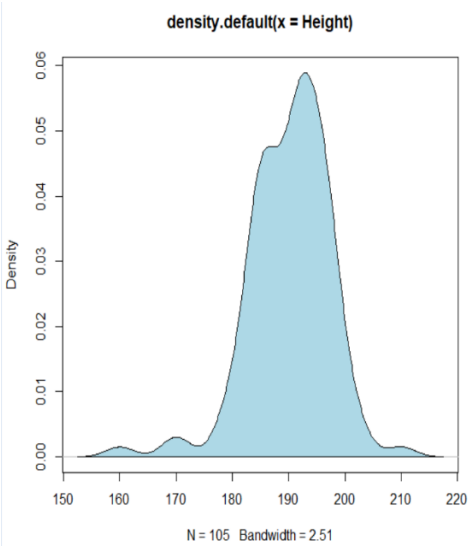
Πίνακας 2: Πίνακας περιγραφικών μέτρων ποιοτικών μεταβλητών

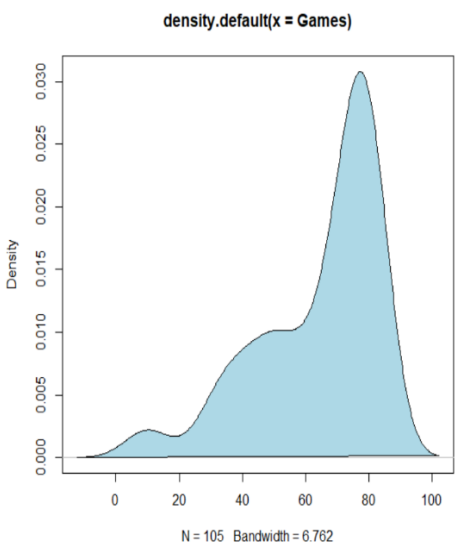
Για να μπορέσουμε να ελέγξουμε ποιες από τις παραπάνω μεταβλητές ακολουθούν κανονική κατανομή , θα εφαρμόσω τους ελέγχους Shapiro-Wilk και Kolmogorov–Smirnov. Μέσω των κατάλληλων συναρτήσεων της R βρίσκουμε ότι σχεδόν όλες οι ποσοτικές μεταβλητές δεν προέρχονται από κανονική κατανομή. Μοναδικές εξαιρέσεις αποτελούν οι μεταβλητές Minutes και TPP , για τις οποίες βρήκαμε ότι :

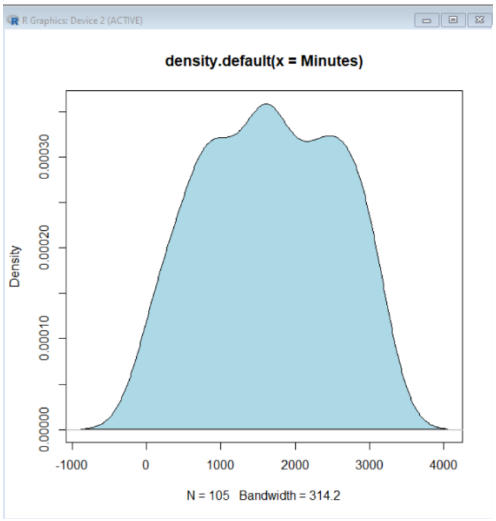
TPP: S-W p value=0.2019 και K-S p value=0.4103

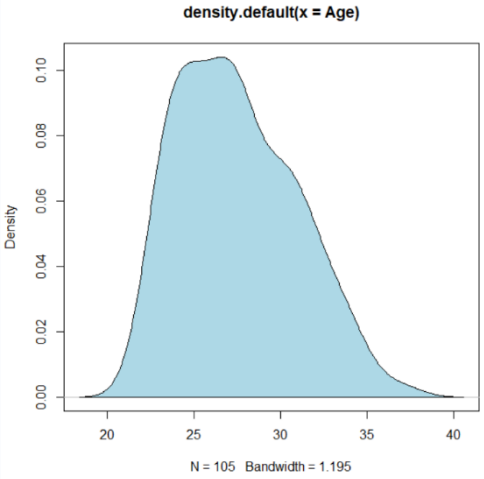
Minutes: S-W p value=0.001958 και K-S p value=0.113

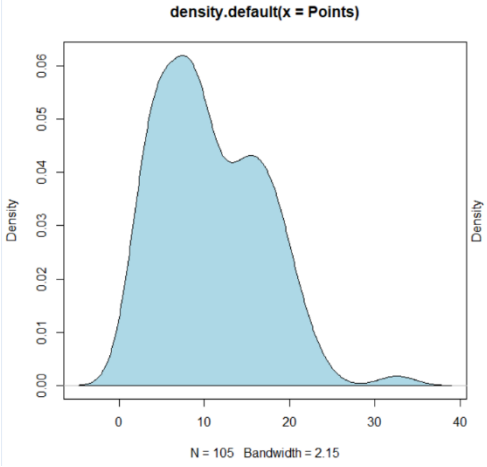
Όλες οι άλλες μεταβλητές είχαν και για τους δύο ελέγχουςp value< 1%.

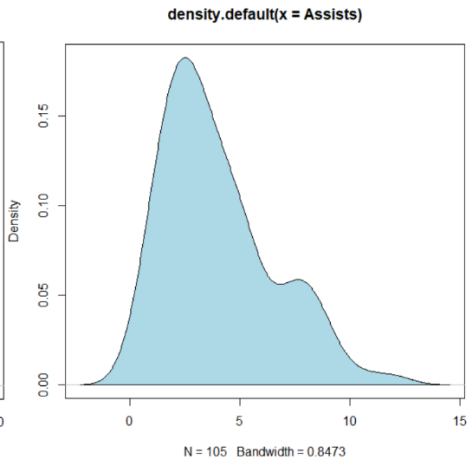

density.default(x = Height)
Density
N = 105 Bandwidth = 2.51

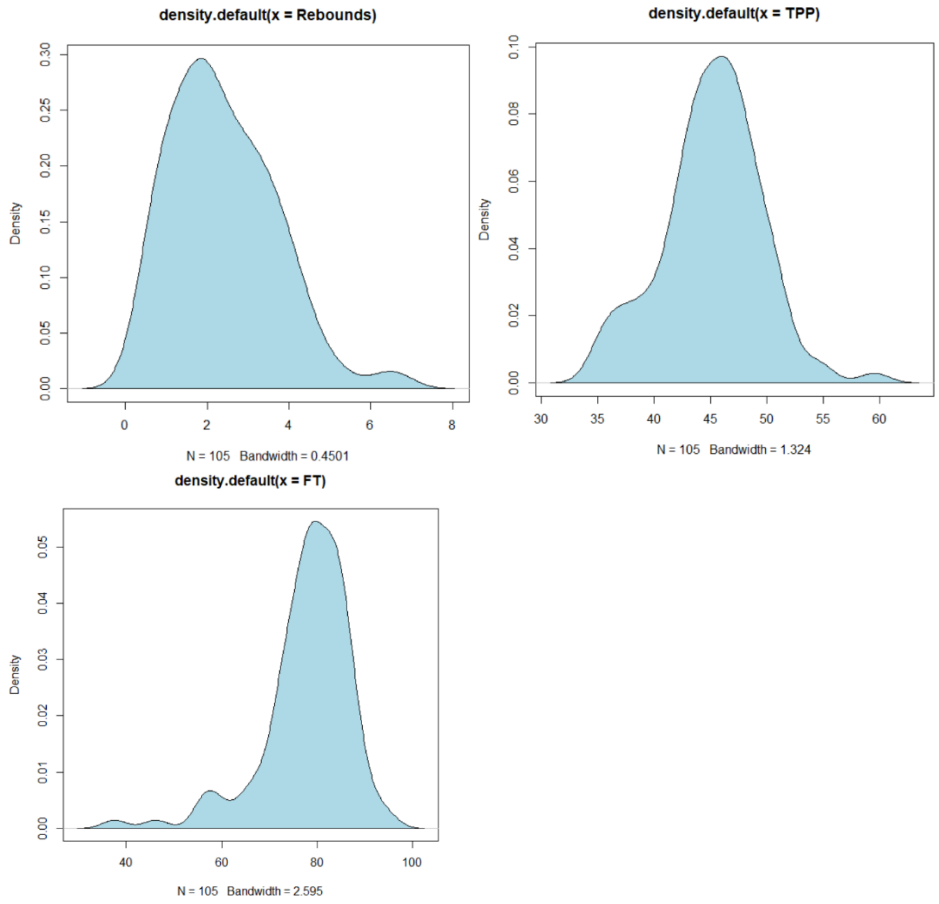

density.default(x = Games)
Density
N = 105 Bandwidth = 6.762


R Graphics: Device 2 (ACTIVE)
density.default(x = Minutes)
Density
N = 105 Bandwidth = 314.2


density.default(x = Age)
Density
N = 105 Bandwidth = 1.195


density.default(x = Points)
Density
N = 105 Bandwidth = 2.15


density.default(x = Assists)
Density
N = 105 Bandwidth = 0.8473



Σχήμα1: Διαγράμματα πυκνότητας πιθανότητας

Μέσα από τα παραπάνω διαγράμματα πυκνότητας επιβεβαιώνεται ότι μας έδωσαν και οι έλεγχοι σημαντικότητας. Βλέπουμε ότι οι μοναδικές καμπύλες που είναι συμμετρικές και αποτελούνται από δύο ομοιόμορφές ουρές που εκτείνονται προς τα αριστερά και τα δεξιά. , δηλαδή έχουν τη μορφή κανονικής κατανομής, είναι εκείνες των μεταβλητών Minutes και TPP. Όλες οι άλλες καμπύλες δείχνουν καθαρά ότι τα δεδομένα των μεταβλητών τους, δεν προήλθαν από κανονική κατανομή.

3. Σχέσεις ανά δύο

Σε αυτή τη φάση της ανάλυσης διερευνούμε τη σχέση που υπάρχει μεταξύ των ποσοτικών μεταβλητών που επεξεργαζόμαστε και θεωρούμε ότι είναι τα σημεία κλειδιά προκειμένου να βγάλουμε ένα τελικό συμπέρασμα σχετικά με το ποιο παίκτη πρέπει να προσπαθήσουμε να κάνουμε μεταγραφή. Αρχικά εφαρμόζουμε τον πίνακα συσχετίσεων του Pearson, μέσω του οποίου διαπιστώνουμε ότι υπάρχει ισχυρή συσχέτιση μεταξύ των μεταβλητών: Minutes και Points(0.7886281), over20 και Points(0.692046) και avg_game_time και Points(0.9010207) . Στη συνέχεια, φτιάχνουμε boxplots ανά κατηγορία για όλες τις κατηγορικές μεταβλητές(βλ.

παράρτημα, σελ 16). Οι σχέσεις των κατηγορικών μεταβλητών που θα είχε ενδιαφέρον να ερευνήσουμε είναι οι εξής:

1. Ηλικία ~ Πόντοι(Age ~ Points)
2. Ύψος ~ Πόντοι (Height ~ Points)
3. Λεπτά ~ Πόντοι (Minutes ~ Points)
4. Παιχνίδια ~ Πόντοι (Games ~ Points)
5. Πάνω από 20 Λεπτά Συμμετοχής ~ Πόντοι (over20~ Points)
6. Μέσος χρόνος Συμμετοχής ~ Πόντοι (avg_game_time ~ Points)
7. Λεπτά Συμμετοχής ~ Πόντοι (Minutes ~ Points)
8. Παιχνίδια ~ Πόντοι (Games ~ Points)
9. Ηλικία ~ Ποσοστό ευστοχίας στα δίποντα (Age ~ TPP, Age4 ~ TPP)
10. Ηλικία ~ Ποσοστό ευστοχίας στις βολές (Age ~ FT , Age4 ~ FT)
11. Πάνω από 20 λεπτά συμμετοχής ~ Πόντοι ανά λεπτό (over20 ~ PPM)
12. Πάνω από 20 λεπτά συμμετοχής ~ Ηλικία (over20 ~ Age, over20~ Age4)
13. Πάνω από 20 λεπτά συμμετοχής ~ Μέσος χρόνος συμμετοχής (over20 ~ avg_game_time)

Οι σχέσεις μεταξύ των παραπάνω μεταβλητών ερμηνεύονται μέσα από τα ραβδογράμματα, που έχουν παρατεθεί στο παράρτημα(σελ 20-23) ,τους ελέγχους t test προκειμένου να ελέγξουμε τη σχέση μεταξύ των ποσοτικών μεταβλητών και τους ελέγχους X square προκειμένου να ελέγξουμε τη σχέση μεταξύ των κατηγορικών μεταβλητών καθώς και τη σχέση μεταξύ κατηγορικών και ποσοτικών μεταβλητών .

Μέσα, λοιπόν, από τους συγκεκριμένους ελέγχους βρίσκουμε ότι όλες οι παραπάνω σχέσεις έχουν ισχυρή συσχέτιση μεταξύ τους, αφού σε όλες απορρίπτεται η μηδενική υπόθεση(p value<5%). Συνεπώς, συνειδητοποιούμε το ότι αν ο παίκτης έχει προοπτικές να αγωνίζεται πάνω από 20 λεπτά δεν εξαρτάται από ένα ή δύο μεταβλητές, αλλά από όλες τις μεταβλητές που αφορούν τα χαρακτηριστικά ή την επίδοση του.

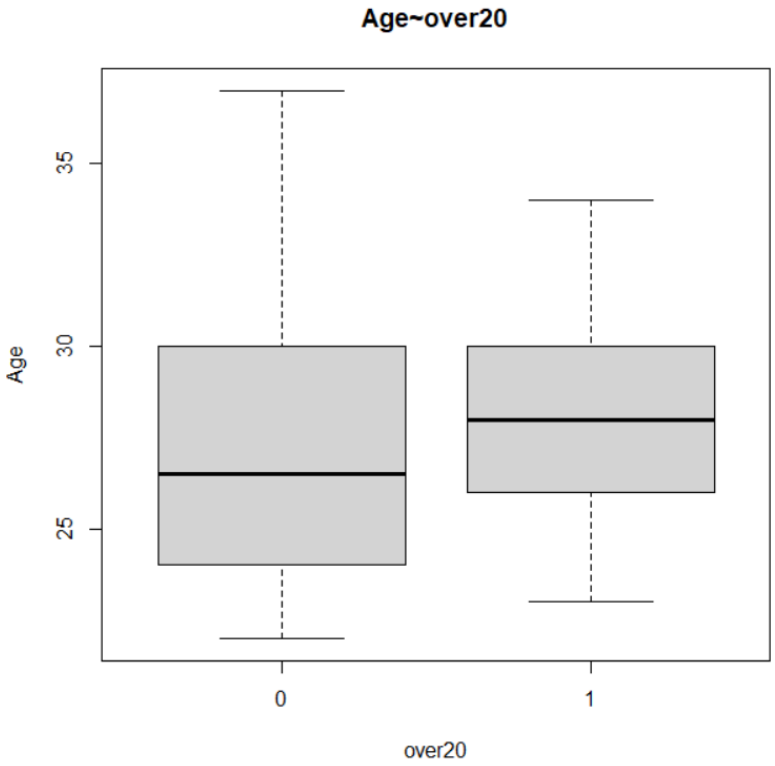
4. Προβλεπτικά ή Ερμηνευτικά μοντέλα

Αφού μελετήσαμε τις σχέσεις ανά δύο μεταξύ των μεταβλητών, ήρθε η ώρα να χρησιμοποιήσουμε κάποια συγκεκριμένα ερμηνευτικά μοντέλα προκειμένου να διεξάγουμε ένα συμπέρασμα σχετικά με τον ποιο παίκτη πρέπει να προσπαθήσουμε να κάνουμε μεταγραφή.

Ωστόσο, πριν ξεκινήσουμε τη διαδικασία αυτή, θα πραγματοποιήσουμε μερικούς ελέγχους για να διερευνήσουμε τις σχέσεις της ηλικίας με το αν παίκτης έχει χρόνο συμμετοχής μικρότερο ή μεγαλύτερο από 20 λεπτά καθώς και αν οι πόντοι του παίκτη επηρεάζουν το χρόνο συμμετοχής στις επιμέρους κατηγορίες.

Για να δούμε ποια είναι η σχέση μεταξύ της ηλικίας και χρόνου συμμετοχής

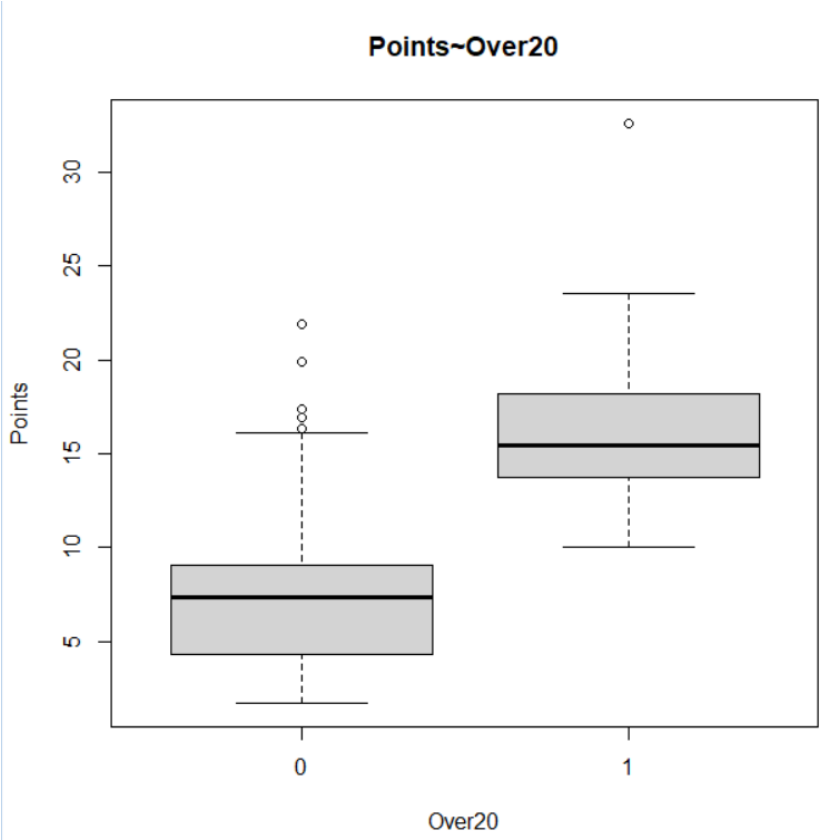
για λιγότερο ή περισσότερο από 20 λεπτά, θέλουμε πρώτα να ελέγξουμε την κανονικότητα της ηλικίας για κάθε μία από τις δύο περιπτώσεις. Πραγματοποιώντας τους ελέγχους κανονικότητας που εφαρμόσαμε στα προηγούμενα ερωτήματα βλέπουμε ότι η μηδενική υπόθεση απορρίπτεται σίγουρα για χρόνο συμμετοχής κάτω από 20 λεπτά , ενώ το αποτέλεσμα είναι διφορούμενο για χρόνο συμμετοχής πάνω από 20 λεπτά. Συγκεκριμένα , για χρόνο συμμετοχής πάνω από 20 λεπτά έχουμε ότι: S-W = 0.04702(<5%) και K-S= 0.109(>5%), και για χρόνο συμμετοχής κάτω από 20 λεπτά έχουμε ότι: S-W = 0.0008959 και K-S= 0.002799 . Συνεπώς, δεδομένου ότι δεν απορρίφθηκε και με τους 2 ελέγχους η κανονικότητα για τη πρώτη περίπτωση αλλά απορρίφθηκε και με τους 2 ελέγχους για τη δεύτερη περίπτωση και συνυπολογίζοντας ότι τα δείγματα δεν ακολουθούν κανονική κατανομή, συμπεραίνουμε ότι ο μέσος δεν είναι κατάλληλο μέτρο περιγραφής της κεντρικής θέσης για τις δύο ομάδες. .Για αυτό το λόγο πραγματοποιούμε τους μη παραμετρικούς ελέγχους Wilcoxon–test και Kruskal-Wallis για να ελέγξουμε την ισότητα των διαμέσων από όπου και συμπεραίνουμε ότι δεν είναι στατιστικά σημαντική η διαφορά μεταξύ των διαμέσων(W:pvalue = 0.07283(>5%) K-W:pvalue= 0.07229 (>5%)),δηλαδή η ηλικία του παίκτη δεν επηρεάζει στατιστικά σημαντικά το αν θα αγωνιστεί περισσότερα ή λιγότερα από 20 λεπτά. Παρακάτω, έχουμε παραθέσει ένα πολύ ενδιαφέρον boxplot.



Σχήμα 2: Boxplot που δείχνει τη σχέση μεταξύ της ηλικίας και με το αν ο παίκτης αγωνίζεται πάνω από 20 λεπτά

Στη συνέχεια θέλουμε να διερευνήσουμε τη σχέση μεταξύ των πόντων που σκοράρουν οι αθλητές με το αν αγωνίζονται πάνω από 20 λεπτά. Για το σκοπό αυτό, θα ελέγξουμε την

κανονικότητα τους. Πραγματοποιώντας τους κατάλληλους ελέγχους βλέπουμε ότι η μηδενική υπόθεση απορρίπτεται σίγουρα για χρόνο συμμετοχής κάτω από 20 λεπτά , ενώ το αποτέλεσμα παρόλο που είναι λίγο διφορούμενο για χρόνο συμμετοχής πάνω από 20 λεπτά, επειδή η απόκλιση είναι πολύ κοντά στο 5%, δεχόμαστε ότι απορρίπτεται . Συγκεκριμένα , για χρόνο συμμετοχής πάνω από 20 λεπτά έχουμε ότι: S-W = 0.001738 (<5%) και K-S= 0.07593 (>5%), και για χρόνο συμμετοχής κάτω από 20 λεπτά έχουμε ότι: S-W = 0.01596 και K-S= 0.02807. Συνεπώς, αφού όπως και πριν , δεν απορρίφθηκε και με τους 2 ελέγχους η κανονικότητα για τη πρώτη περίπτωση αλλά απορρίφθηκε και με τους 2 ελέγχους για τη δεύτερη περίπτωση και συνυπολογίζοντας ότι τα δείγματα δεν ακολουθούν κανονική κατανομή, συμπεραίνουμε ότι ο μέσος δεν είναι κατάλληλο μέτρο περιγραφής της κεντρικής θέσης για τις δύο ομάδες. Για αυτό το λόγο πραγματοποιούμε τους μη παραμετρικούς ελέγχους Wilcoxon – test και Kruskal-Wallis για να ελέγξουμε την ισότητα των διαμέσων από όπου και συμπεραίνουμε ότι είναι στατιστικά σημαντική η διαφορά μεταξύ των διαμέσων(W:pvalue = 1.969e-15(<5%) και K-W:pvalue= 1.916e-15 (<5%)),δηλαδή η επίδοση του παίκτη σε πόντους επηρεάζει στατιστικά σημαντικά το αν θα αγωνιστεί περισσότερα ή λιγότερα από 20 λεπτά. Παρακάτω, έχουμε παραθέσει ένα επίσης πολύ ενδιαφέρον boxplot.



Σχήμα 3: Boxplot που δείχνει τη σχέση μεταξύ των πόντων και με το αν ο παίκτης αγωνίζεται πάνω από 20 λεπτά

Αφού έχουμε ελέγξει τα δεδομένα μας εκτενώς και έχουμε κατανοήσεις τις σχέσεις μεταξύ των μεταβλητών μπορούμε να προχωρήσουμε στη προσαρμογή ενός μοντέλου για να ερμηνεύσουμε την επιρροή των μεταβλητών στις προοπτικές που έχει ο παίκτης για αύξηση του χρόνου συμμετοχής του πάνω από 20 λεπτά. Αρχικά παίρνουμε το πλήρες μοντέλο για το αν ο παίκτης αγωνίζεται περισσότερο ή λιγότερο από 20 λεπτά, όπου θέλουμε να είναι της μορφής:

Over20(χρόνος συμμετοχής πάνω από 20λεπτά)= βο +β1*Points(πόντοι) + β2*Assists(τελικές πάσες) + β3*Rebounds(ριμπάουντ) + β4*TTP(ποσοστό ευστοχίας στα 2ποντα) + β5*FT(ποσοστό ευστοχίας στις ελεύθερες βολές) + β6*Games(παιχνίδια) + β7*Height(ύψος) + β8*Minutes(Λεπτά) + β9*avg_game_time(μέσος χρόνος συμμετοχής) + β10*Age(Ηλικία~ Ποσοτική) + β11*Age4(Ηλικία~ Κατηγορική)

Το μοντέλο που θα εφαρμόσουμε είναι το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης. Η εξαρτημένη μεταβλητή μας, δηλαδή η μεταβλητή που θέλουμε να μελετήσουμε και να προβλέψουμε είναι η over20. Οι υπόλοιπες μεταβλητές αποτελούν ανεξάρτητες μεταβλητές και τώρα μέσω των αποτελεσμάτων της παλινδρόμησής θα εξηγήσουμε ποια και πόσο σημαντική είναι η επιρροή τους στην εξαρτημένη μεταβλητή.
Κάνουμε χρήση και της απλής αλλά και της γενικευμένης γραμμικής παλινδρόμησής.
Οι παράγοντες που θα εξετάσουμε είναι το pvalue,το Std. Error,που είναι το εκτιμώμενο σφάλμα της πρόβλεψης για τον μέσο όρο της εξαρτημένης μεταβλητής, το Estimate, που είναι η εκτιμώμενη τιμή του συντελεστή παλινδρόμησης για κάθε ανεξάρτητη μεταβλητή στο μοντέλο και το T value,η οποία αναφέρεται στην τιμή του στατιστικού τεστ t για τον κάθε συντελεστή παλινδρόμησης. Εκτελούμε, λοιπόν τη παλινδρόμηση και τα αποτελέσματα που παίρνουμε είναι τα εξής:

Μεταβλητές	Estimate	Std. Error	T value	Pr(>|t|)
Σταθερά	-2.99344	1.1655	-2.568	0.01185<5%
Points	-0.01863	0.0116	-1.577	0.11837
Assists	0.02012	0.0198	1.015	0.31277
Rebounds	0.012457	0.0387	0.322	0.74853
TPP	0.01378	0.0067	2.037	0.04454<5%
FT	-0.0030	0.0033	-0.895	0.37314
Games	0.00454	0.0034	1.327	0.18775
Minutes	-0.0001	0.0001	-0.719	0.47389
avg_game_time	0.0484	0.0118	4.105	8.8e-05<1%
Height	0.0142	0.0047	3.036	0.00312<1%
Age	-0.0267	0.0283	-0.946	0.34656
Age4(25.8,29.5]	0.0616	0.1142	0.540	0.59062
Age4(29.5,33.2]	0.1601	0.2140	0.748	0.45641
Age4(33.2,37]	0.2089	0.3231	0.647	0.51940

Residual standard error	0.2619
Multiple R-squared	0.7362
Adjusted R-squared	0.6986
AIC	31.557

Στο παραπάνω αποτέλεσμα βλέπουμε ένα μοντέλο το οποίο έχει μία αρκετά καλή προσαρμογή(R-squared=70%),το οποίο σημαίνει ότι οι ανεξάρτητες μεταβλητές που έχουμε συμπεριλάβει είναι καλές προγνωστικές μεταβλητές και εξηγούν μεγάλο μέρος της διακύμανσης της απόκρισης. Επίσης, παίρνουμε AIC=31.557 και Residual standard error=0.2619, τιμές οι οποίες είναι πάρα πολύ χαμηλές. Ο συνδυασμός των παραπάνω παραγόντων υποδηλώνει ότι το μοντέλο έχει πολύ μικρή απόκλιση από τις πραγματικές τιμές και πολύ καλή προσαρμογή στα δεδομένα. Και από τις 2 εφαρμογές παλινδρόμησής μας γίνεται εύκολα αντιληπτό ότι υπάρχουν 4 μεταβλητές που σε πρώτο χρόνο τουλάχιστον ασκούν υψηλή επιρροή στη διαμόρφωση του μοντέλου. Οι μεταβλητές αυτές είναι η σταθερά (β0,over20) , η TPP(β4) , η avg_game_time(β8) και η Height(β9).

Μέσα από εκτενές model selection βρήκαμε ότι το καλύτερο μοντέλο όσο αφορά την επίδραση των μεταβλητών είναι το παρακάτω:

Μεταβλητές	Estimate	Std.Error	t value	Pr(>|t|)
Σταθερά	-3.063099	0.822371	-3.725	0.000324<1%
Points	-0.024621	0.010301	-2.390	0.018713<5%
TPP	0.010937	0.006302	1.736	0.085723
avg_game_time	0.052053	0.006152	8.462	2.29e-13<1%
Height	0.011700	0.003797	3.082	0.002660<1%

Residual standard error	0.258
Multiple R-squared	0.7187
Adjusted R-squared	0.7075
AIC	20.319

Το συγκεκριμένο μοντέλο προέκυψε μέσα από Model Selection BIC. Συγκεκριμένα, αρχικά εμφανίσαμε ένα μοντέλο το οποίο περιείχε όλες τις μεταβλητές και είχε R-squared=70% και AIC=31.557. Στη συνέχεια αρχίσαμε να αφαιρούμε μία μία κάποιες μεταβλητές και τώρα, το μοντέλο μας έχει R-squared=71% και AIC=20.319. Το συγκεκριμένο AIC είναι το χαμηλότερο που μπορούμε να πετύχουμε και συνεπώς ο αλγόριθμος τώρα σταματάει. Επίσης έχουμε residual standard error ίσο με 0.258, το οποίο δείχνει το πόσο μικρή απόκλιση υπάρχει από τις πραγματικές τιμές και το πόσο καλή προσαρμογή έχει το μοντέλο στα δεδομένα μας. Μπορούμε να ερμηνεύσουμε πολύ πιο εύκολα τις μεταβλητές και την επίδραση τους στην εξαρτημένη μεταβλητή και συνεπώς να εξάγουμε πολύ πιο εύκολα συμπέρασμα. Η

σταθερά(β0), ο μέσος χρόνος συμμετοχής(β3) και το ύψος(β4) δείχνουν να έχουν την ύψιστη επίδραση στην εξαρτημένη μας μεταβλητή (με pvalue<1%) , ενώ και οι πόντο(β1) ι επηρεάζουν σημαντικά(pvalue<5%). Η μεταβλητή TPP(β2) δεν φαίνεται να έχει την ίδια επίδραση, αλλά επειδή είναι πολύ κοντά στο να απορριφθεί ως στατιστικά σημαντική , θα την λάβουμε υπόψιν μας(pvalue=0.0857. 0.0857-0.05=3%). Για να υποστηρίξουμε τα παραπάνω επιχειρήματα μας, πραγματοποιούμε και έναν έλεγχο πολλαπλής συνδεσιμότητας(vif).

Points	TPP	avg_game_time	Height
6.086565	1.197302	5.746828	1.127435

Το παραπάνω αποτέλεσμα αποδεικνύει τη σχέση που έχουν μεταξύ τους οι μεταβλητές. Σε γενικές γραμμές, οι επιθυμητές τιμές για το GVIF είναι κοντά στο 1 ή χαμηλά, γύρω από αυτό. Συνεπώς, το αποτέλεσμα αυτό υποδηλώνει ότι οι μεταβλητές είναι ανεξάρτητες μεταξύ τους.

Τέλος, θα δείξουμε τι συμβαίνει στην ερμηνεία του αρχικού μας μοντέλου αν αφαιρέσουμε τη σταθερά.

Μεταβλητές	Estimate	Std. Error	t value	Pr(>t)
Points	-0.018	0.0116	-1.577	0.11837
Assists	0.020	0.0198	1.015	0.31277
Rebounds	0.012	0.038	0.322	0.74853
TPP	0.013	0.0067	2.037	0.04454<5%
FT	-0.003	0.0033	-0.895	0.37314
Games	0.004	0.0034	1.327	0.18775
Minutes	-0.0001	0.0118	-0.719	0.47389
avg_game_time	0.048	0.0001	4.105	8.8e-05<1%
Height	0.014	0.0047	3.036	0.00312<1%
Age	-0.026	0.028	-0.946	0.34656
Age4(22,25.8)	-2.9934	1.165	-2.568	0.01185<5%
Age4(25.8,29.5)	-2.9317	1.214	-2.414	0.01777<5%
Age4(29.5,33.2)	-2.8333	1.274	-2.223	0.02872<5%
Age4(33.2,37)	-2.7844	1.354	-2.056	0.04261<5%

Residual standard error	0.2619
Multiple R-squared	0.9096
Adjusted R-squared	0.8957
AIC	31.557

Όπως είναι εύκολα αντιληπτό, έχουμε για πρώτη φορά προσαρμογή πολύ κοντά στο 90%. Οι τιμές AIC και residual standard error παραμένουν και πάλι σε πολύ καλά επίπεδα. Η επίδραση των μεταβλητών είναι η ίδια με το αρχικό μοντέλο, αλλά βλέπουμε ότι πρώτη φόρα

, η ηλικία ασκεί σημαντική επιρροή στο μοντέλο(p-value<5%). Στην τελική μας ανάλυση, δείξαμε ότι ο ρόλος της δεν είναι σημαντικός.

Σε αυτό το σημείο θέλουμε να τονίσουμε ότι στο παράρτημα έχουμε παραθέσει τα κατάλληλα διαγράμματα(βλ. Παράρτημα, σελ.27-28-29), τα οποία αποδεικνύουν τη χαμηλή ομοσκεδαστικότητα των μοντέλων και συνεπώς τις μικρές αποκλίσεις που έχουν οι τιμές τους με την πραγματικότητα.

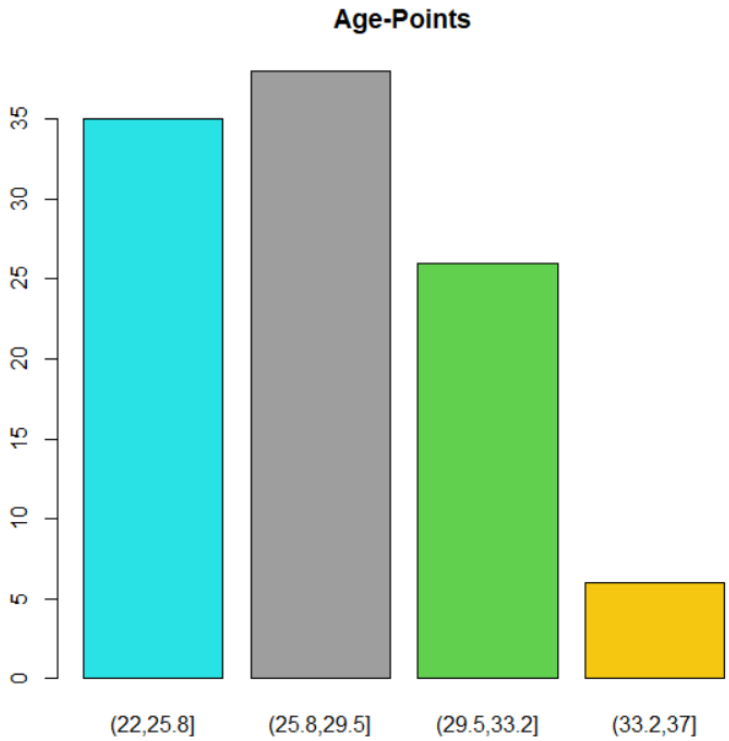
5. Συμπέρασμα και συζήτηση

Η παραπάνω μελέτη είχε ως σκοπό την εκτίμηση ενός υποδείγματος για τους παράγοντες που επηρεάζουν το αν ο αθλητής έχει προοπτικές να αγωνίζεται πάνω από 20 λεπτά αλλά και την σχέση των παραγόντων αυτών μεταξύ τους. Το μοντέλο μας δείχνει να έχει μία αρκετά καλή προσαρμογή (R2adj = 0.72 και φτάνει και σε επίπεδα του 90% αν αφαιρέσουμε τη σταθερά) και ερμηνεύει τις σχέσεις μεταξύ των μεταβλητών αρκετά καλά.

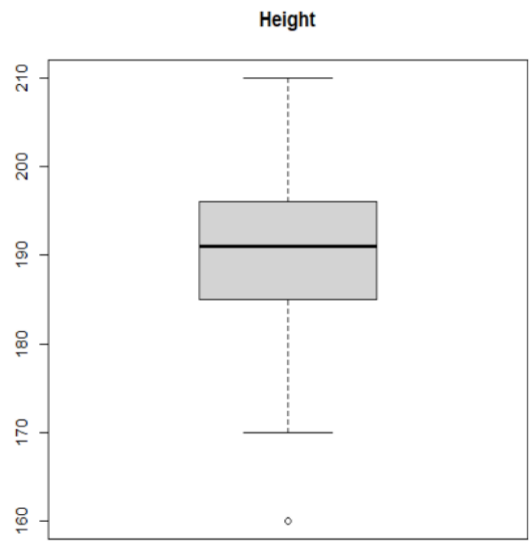
Το συμπέρασμα που βγάζουμε από την παραπάνω ανάλυση είναι ότι η χρησιμοποίηση ή όχι του αθλητή πάνω από 20 λεπτά επηρεάζεται δραματικά από τους πόντους που πετυχαίνει , το μέσο χρόνο συμμετοχής του καθώς και από το ύψος του. Επίσης φαίνεται ότι η ευστοχία του παίκτη στα σουτ εντός πεδιάς (ποσοστό στα δίποντα δηλαδή) επηρεάζει τις προοπτικές αυξημένου χρόνου συμμετοχής. Το ύψος είναι ένα στοιχείο του ανθρώπου που δεν μπορείς να αλλάξεις. Ο μέσος χρόνος συμμετοχής είναι επίσης ένας παράγοντας που δεν μπορεί να αλλάξει από μόνος του και πρέπει να συμβούν ένας συνδυασμός πραγμάτων για να αυξηθεί. Ωστόσο, το σκοράρισμα και η ευστοχία στα δίποντα είναι παράγοντες που δεν μπορούμε να αγνοήσουμε.

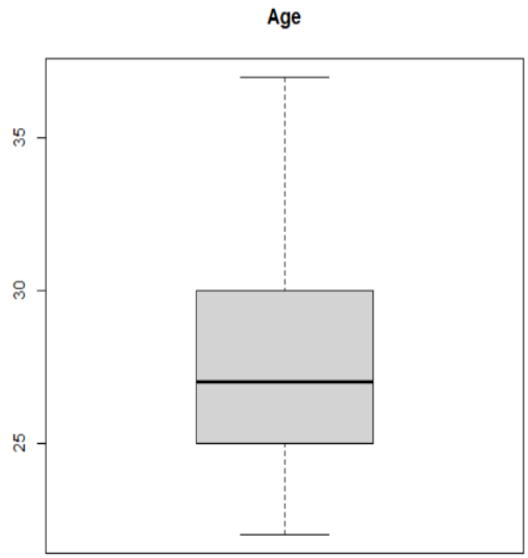
Συνεπώς, προκειμένου να αυξήσει ένας αθλητής τις πιθανότητες για αυξημένο χρόνο συμμετοχής πρέπει να βελτιώσει αισθητά την επίδοση του στο σκοράρισμα και στην ευστοχία του στα δίποντα. Εμείς για να μπορέσουμε να αποφασίσουμε ποιον αθλητή πρέπει να προσπαθήσουμε να υπογράψουμε, δημιουργήσαμε ένα data frame με τους πόντους , την ευστοχία στα δίποντα, το μέσο χρόνο συμμετοχής και το ύψος των αθλητών που έχουν μέσο χρόνο συμμετοχής πάνω από 20 λεπτά και πόντους ανά αγώνα πάνω από τον γενικό μέσο όρο πόντων. Με βάση λοιπόν, τα υπάρχων δεδομένα, εμείς προτείνουμε για μεταγραφή τον καλαθοσφαιριστή Michael Jordan. Ο Jordan με βάση τα στοιχεία μας για τις σχετικές κατηγορίες είναι πρώτος σε πόντους, $3^{ος}$ σε μέσο χρόνο συμμετοχής , $5^{ος}$ σε ύψος και $7^{ος}$ σε ποσοστά ευστοχίας στα δίποντα με ποσοστό 49.5% και απόσταση 5% από τον πρώτο της συγκεκριμένης λίστας. Κανένας άλλος αθλητής δεν έχει τόσο μεγάλη συνέπεια στις 4 αυτές κατηγορίες. Συνεπώς, ο συγκεκριμένος παίκτης μπορεί να κάνει τη διαφορά για οποιαδήποτε ομάδα.

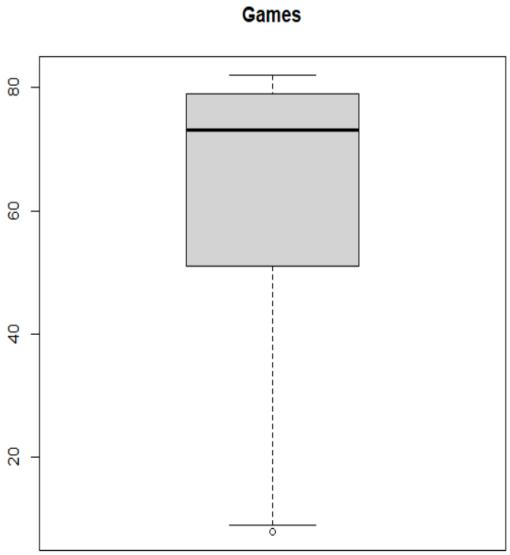
Παράρτημα

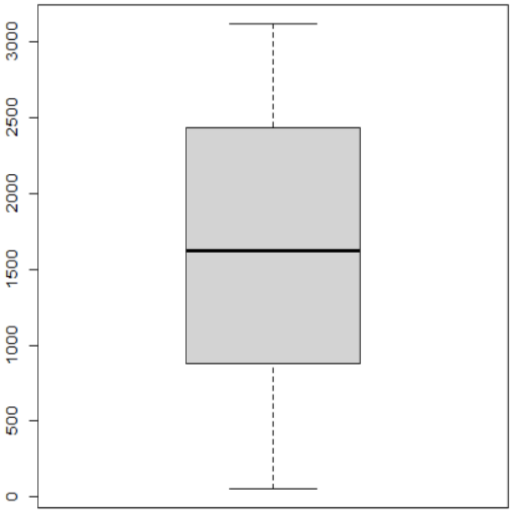


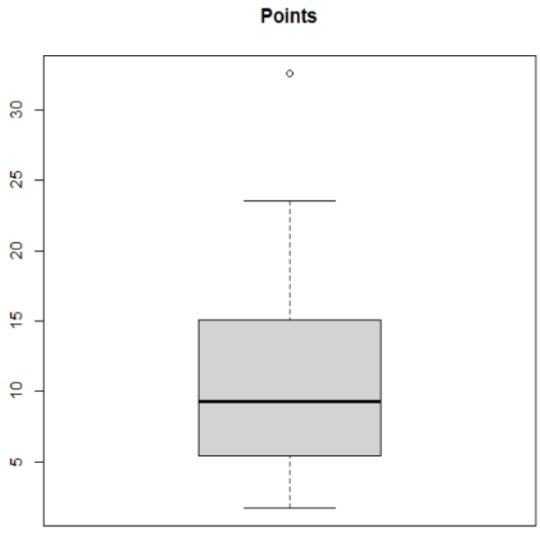
Σχήμα 4: Barplot που δείχνει χαρακτηριστικά πόσο κομβικό ρόλο παίζει η ηλικία στην επίδοση ενός αθλητή , όσο αφορά το τομέα του σκοραρίσματος.

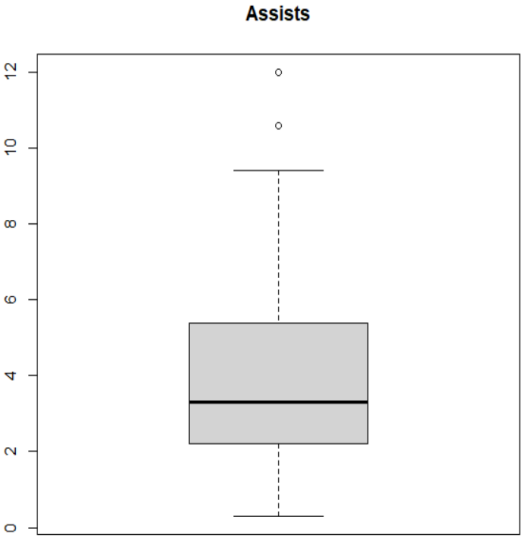

Height
210
200
190
180
170
160

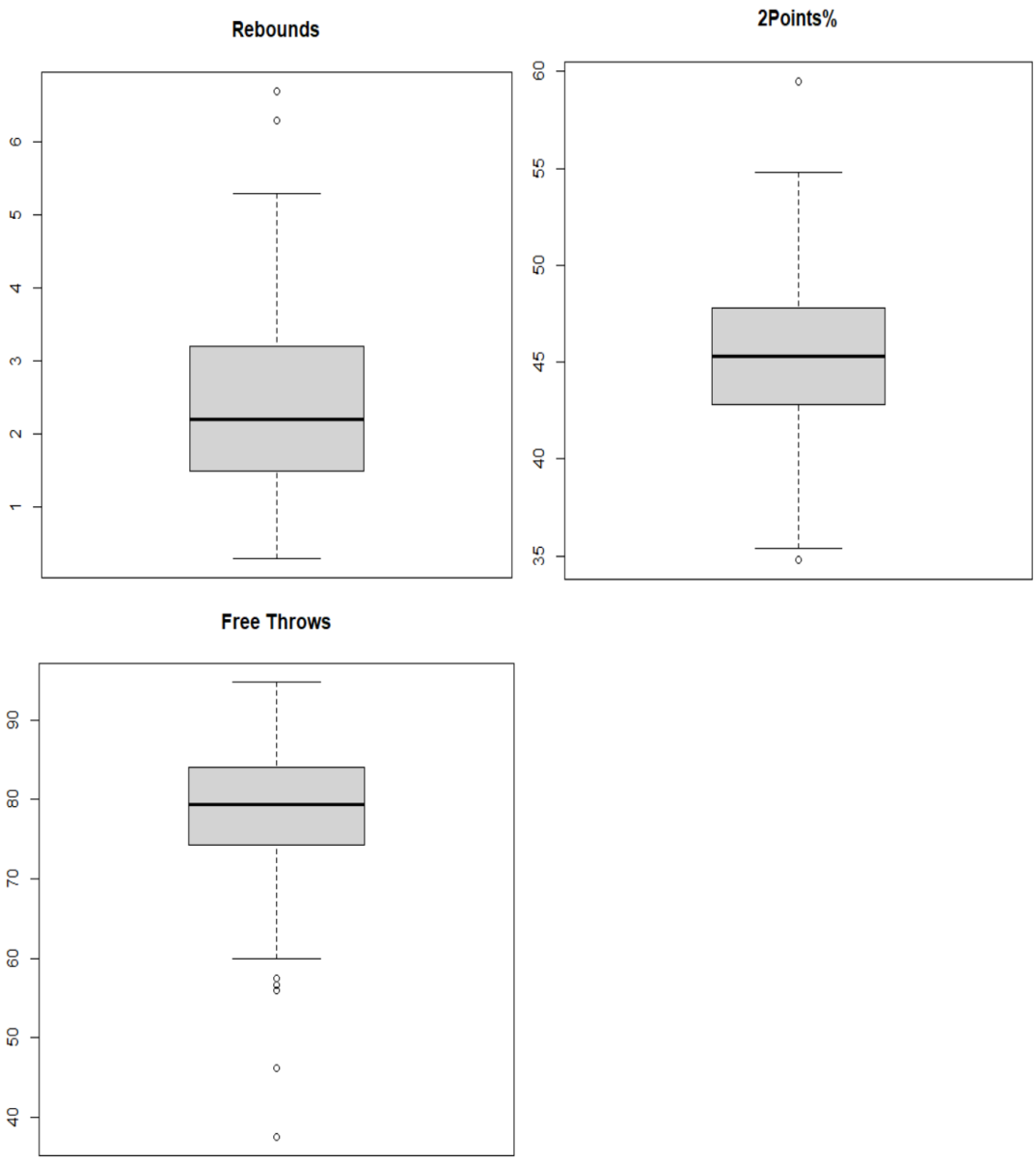

Age
35
30
25


Games
80
60
40
20

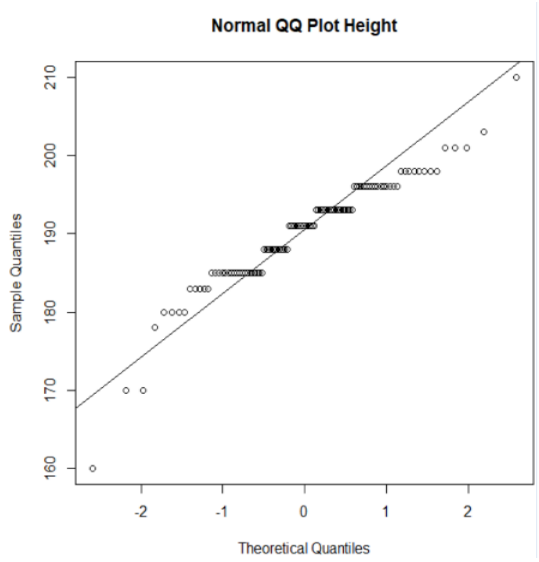

Minutes
3000
2500
2000
1500
1000
500
0

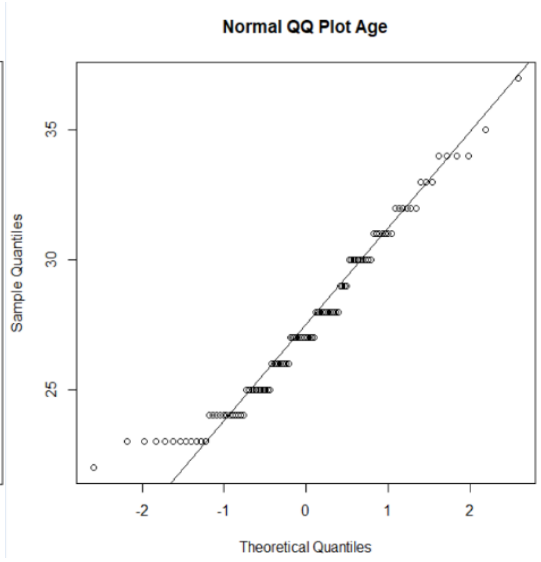

Points
30
25
20
15
10
5

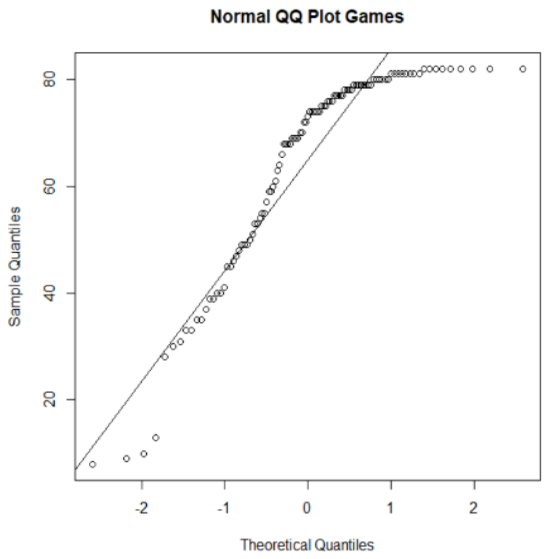

Assists
12
10
8
6
4
2
0

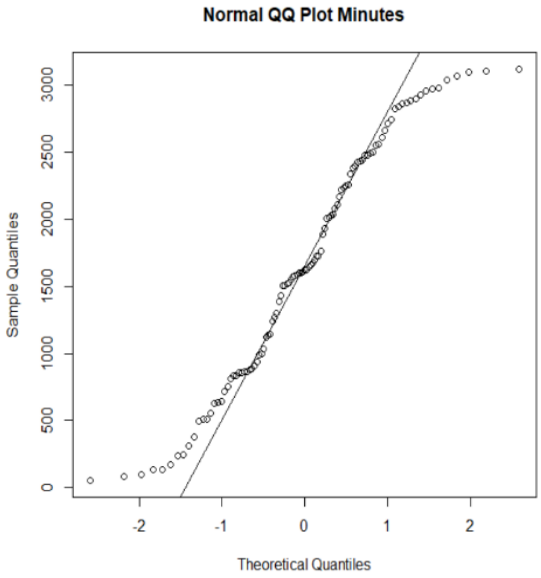


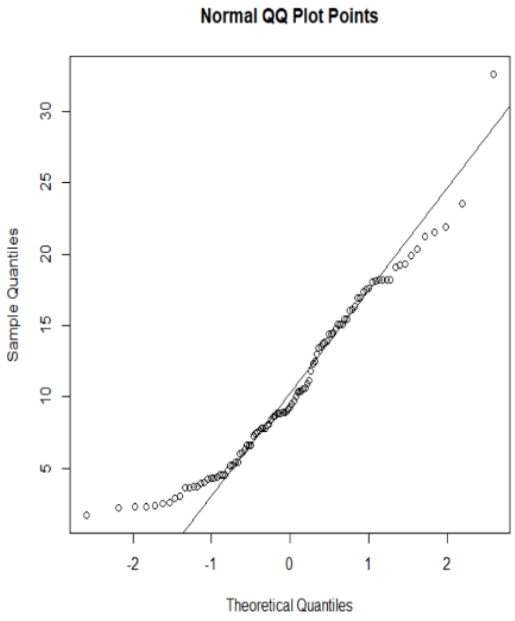
Σχήμα 5: Boxplot της κάθε ποιοτικής μεταβλητής ξεχωριστά

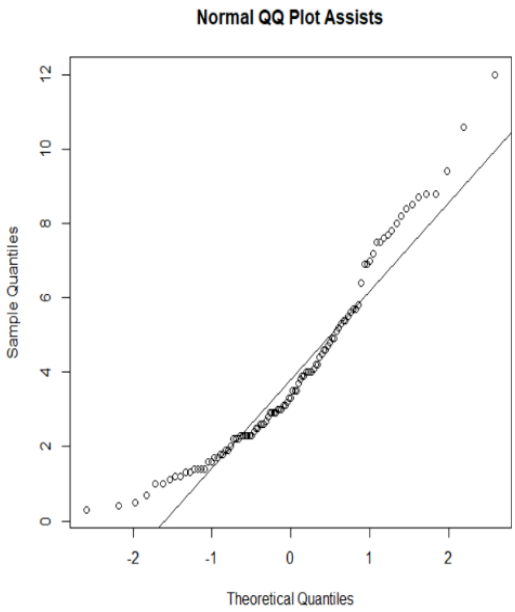

Normal QQ Plot Height
Sample Quantiles
Theoretical Quantiles
160
170
180
190
200
210
-2
-1
0
1
2

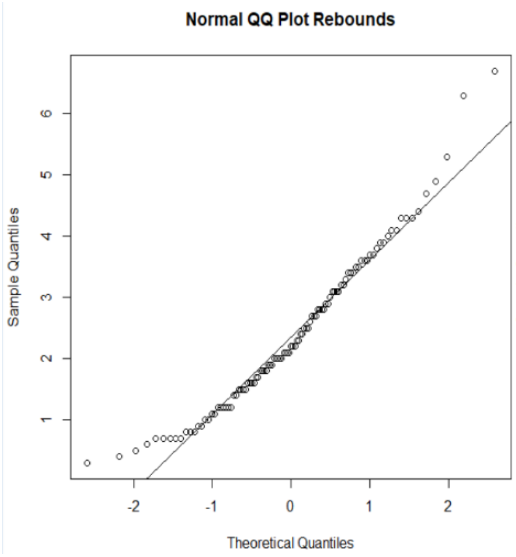

Normal QQ Plot Age
Sample Quantiles
Theoretical Quantiles
25
30
35
-2
-1
0
1
2

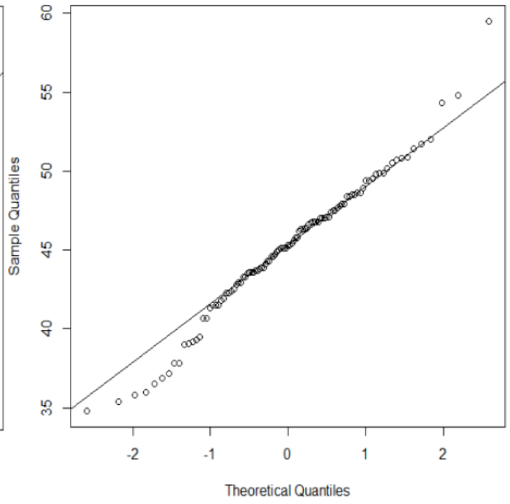

Normal QQ Plot Games
Sample Quantiles
Theoretical Quantiles
20
40
60
80
-2
-1
0
1
2

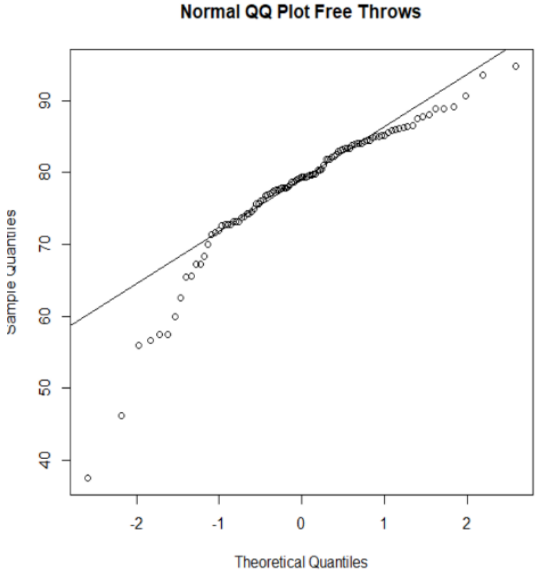

Normal QQ Plot Minutes
Sample Quantiles
Theoretical Quantiles
0
500
1000
1500
2000
2500
3000
-2
-1
0
1
2


Normal QQ Plot Points
Sample Quantiles
Theoretical Quantiles
5
10
15
20
25
30
-2
-1
0
1
2

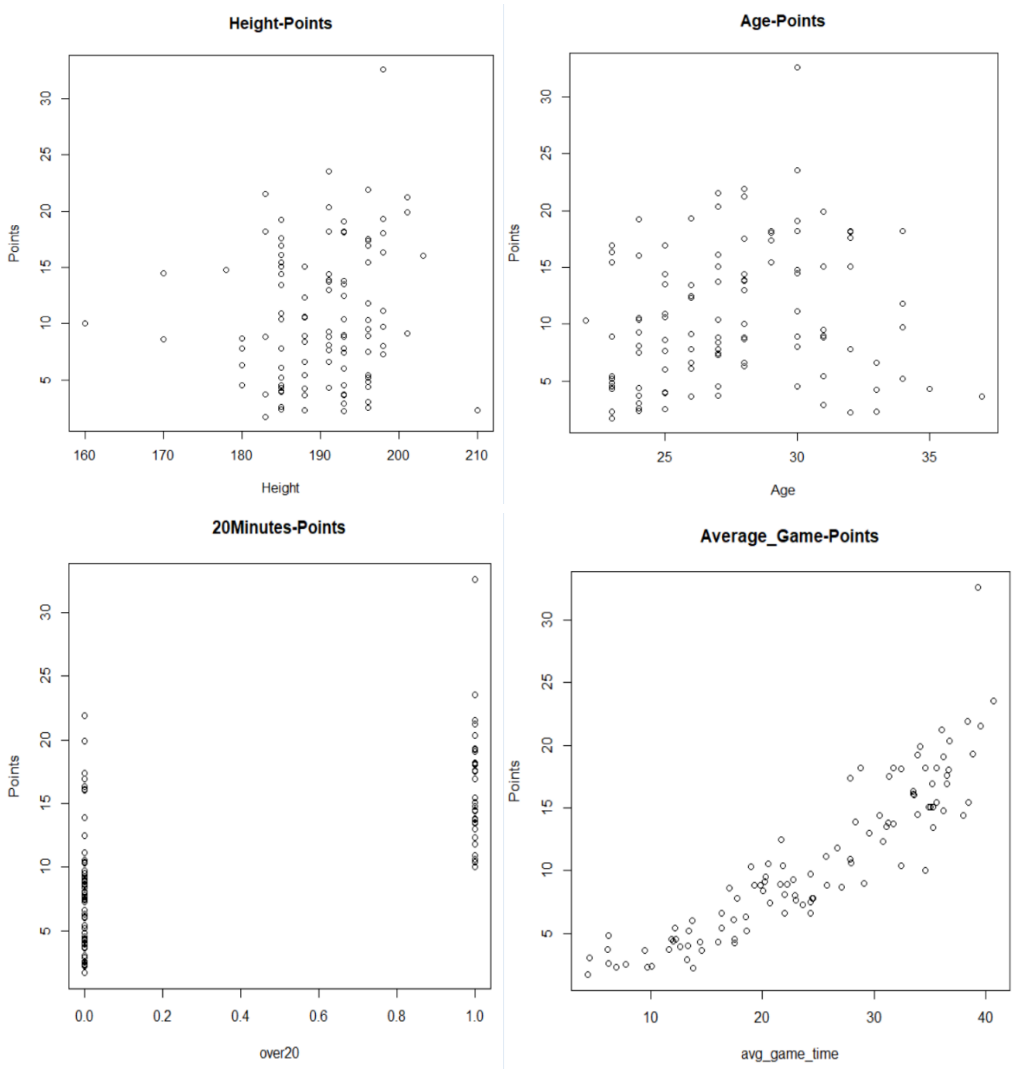

Normal QQ Plot Assists
Sample Quantiles
Theoretical Quantiles
0
2
4
6
8
10
12
-2
-1
0
1
2


Normal QQ Plot Rebounds
Sample Quantiles
Theoretical Quantiles
1
2
3
4
5
6
-2
-1
0
1
2

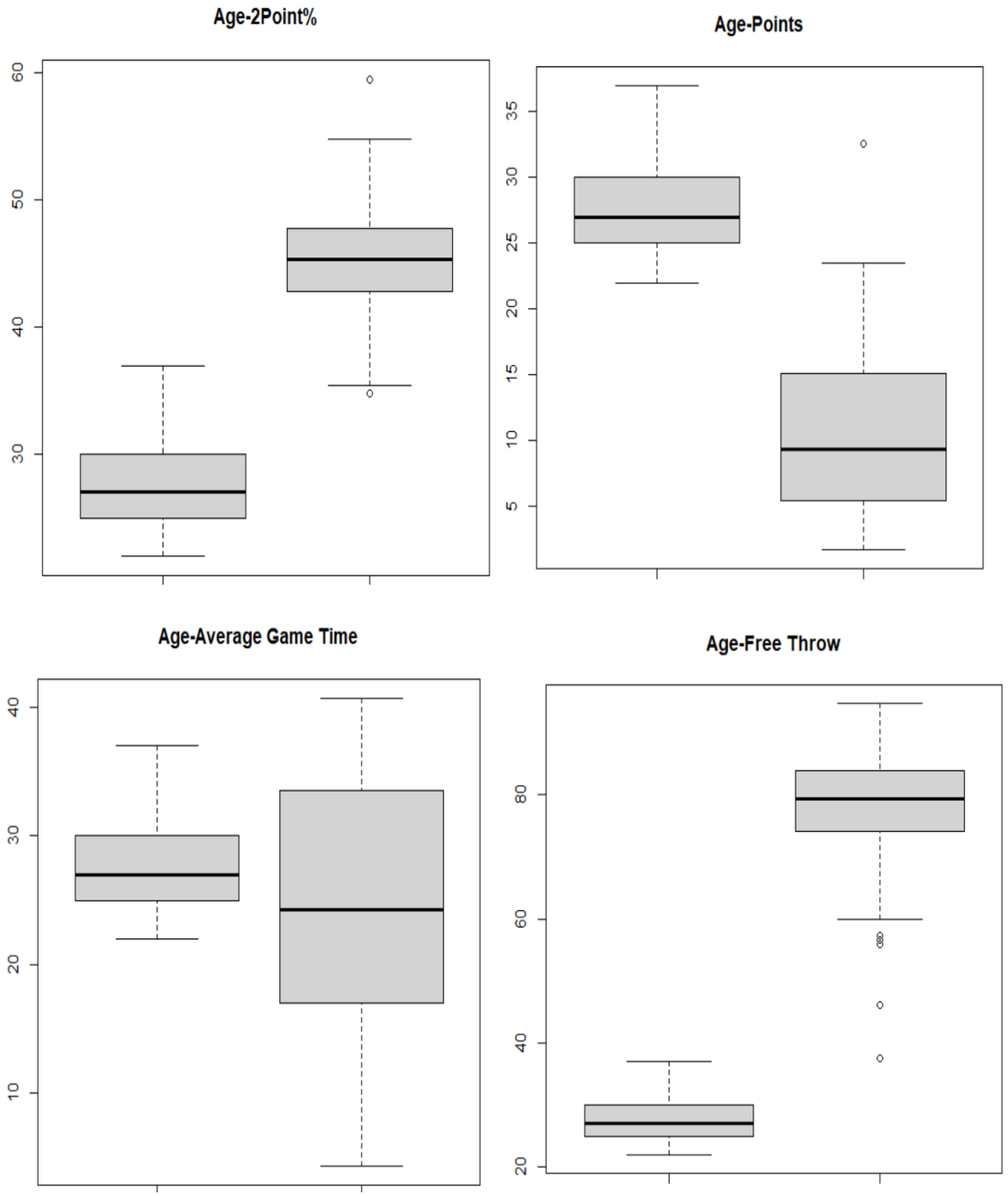

Normal QQ Plot 2Points%
Sample Quantiles
Theoretical Quantiles
35
40
45
50
55
60
-2
-1
0
1
2


Normal QQ Plot Free Throws
Sample Quantiles
Theoretical Quantiles
40
50
60
70
80
90
-2
-1
0
1
2

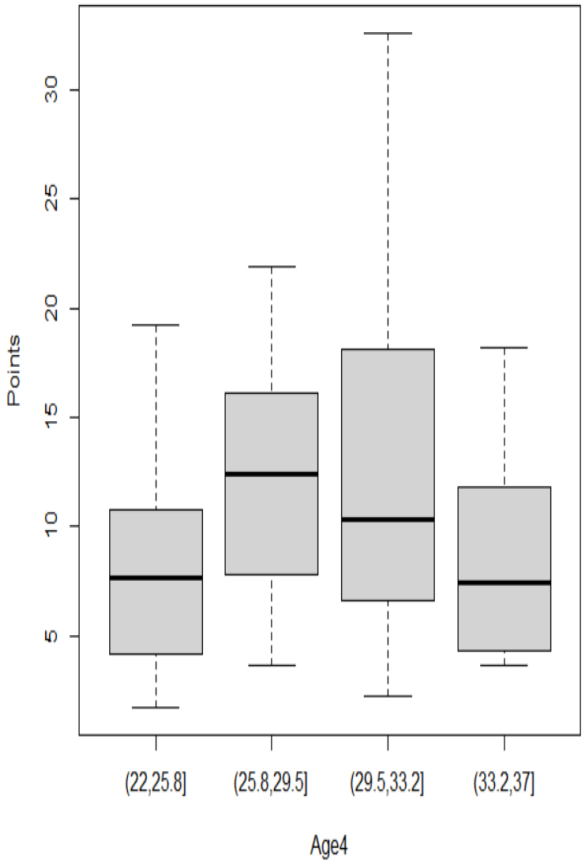
Σχήμα 6: QQplots για κάθε ποιοτική μεταβλητή ξεχωριστά όπου παρατηρείται απόκλιση από την κανονική κατανομή για όλες τις μεταβλητές εκτός από την μεταβλητή TPP(Two Points Percentage) και ίσως και για τη μεταβλητή Minutes.

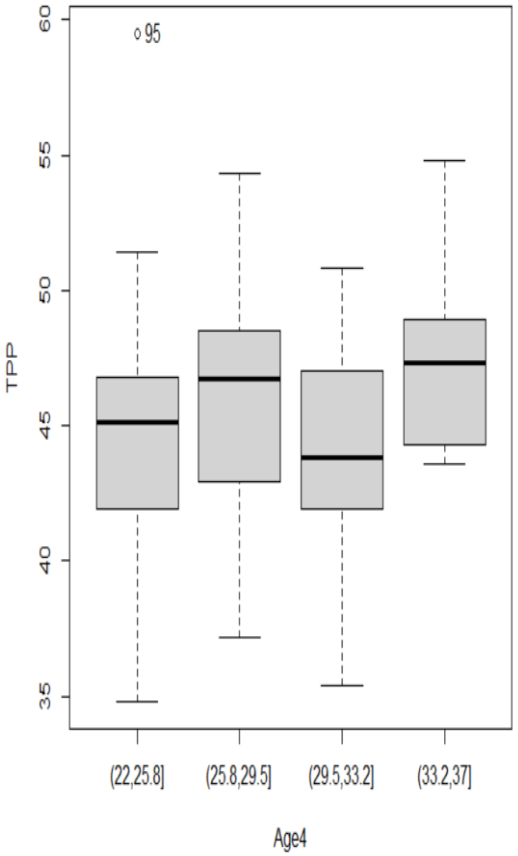


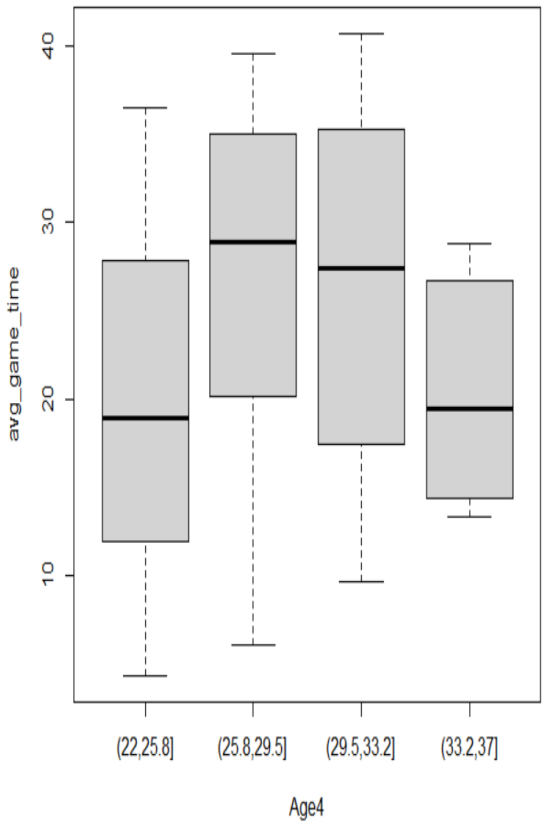
Σχήμα 7: Μέσα από τα παραπάνω scatterplots συγκρίνουμε τις τιμές της μεταβλητής Points με τις τιμές των μεταβλητών που θεωρούμε ότι έχουν τη μεγαλύτερη επίδραση.

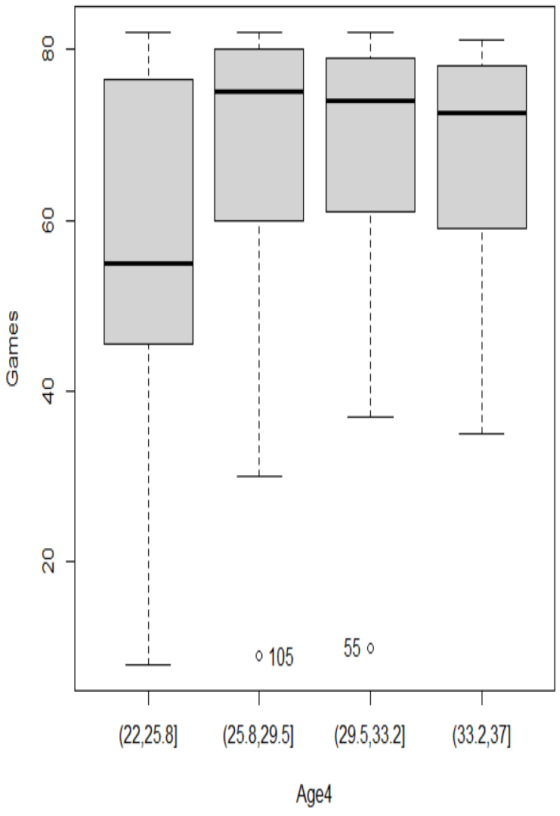


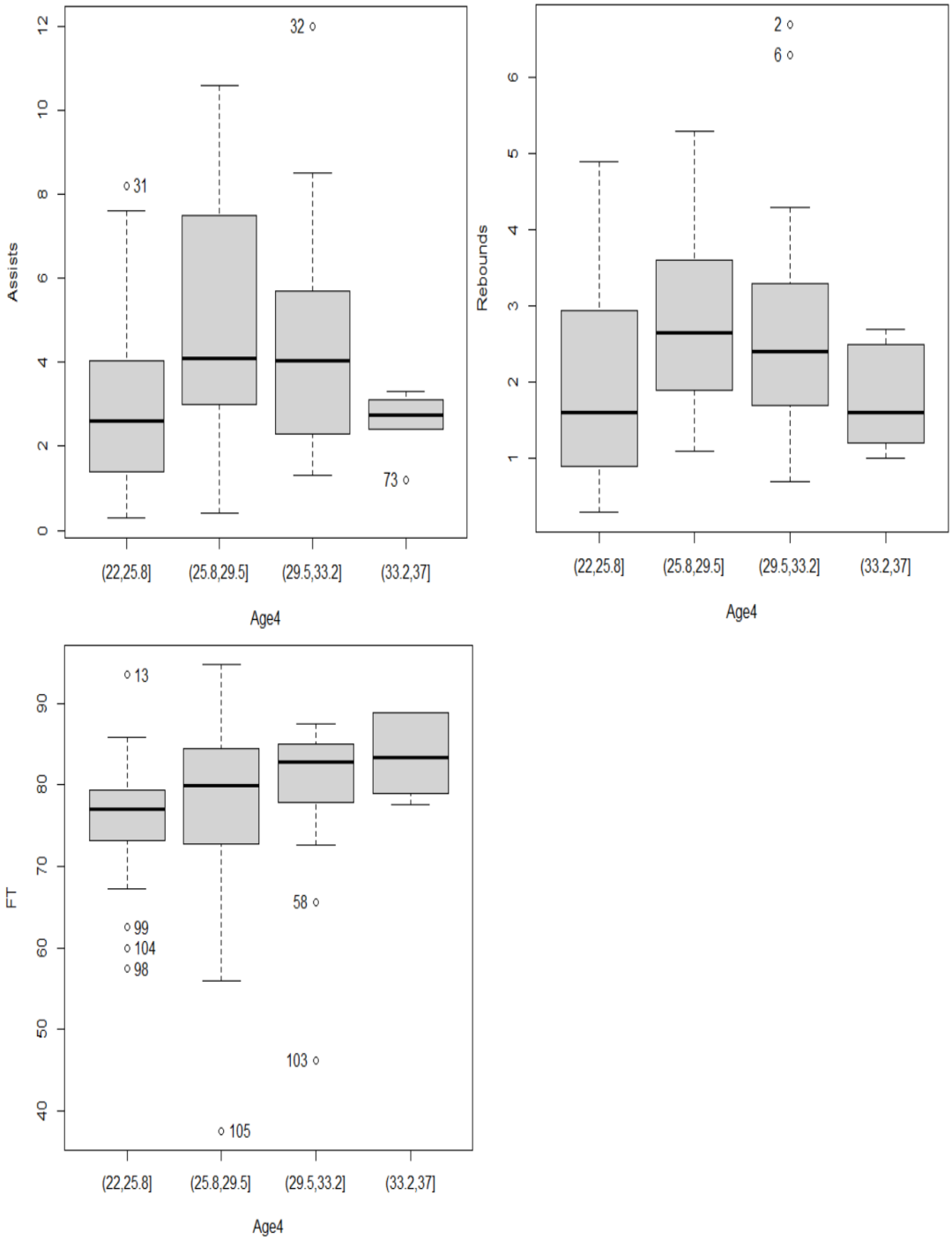
Σχήμα 8: Μέσα από τα παραπάνω boxplots συγκρίνουμε τις τιμές της μεταβλητής Age με τις τιμές των μεταβλητών που θεωρούμε ότι έχουν τη μεγαλύτερη επίδραση.


Points
5
10
15
20
25
30
(22,25.8]
(25.8,29.5]
(29.5,33.2]
(33.2,37]
Age4

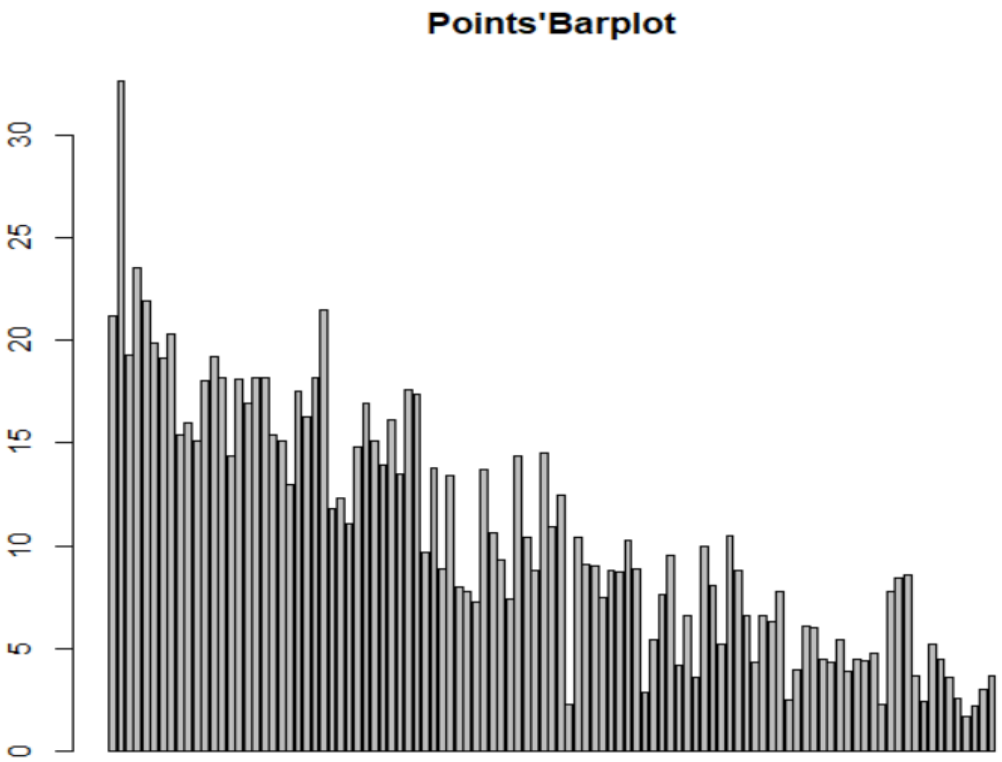

TPP
35
40
45
50
55
60
95
(22,25.8]
(25.8,29.5]
(29.5,33.2]
(33.2,37]
Age4

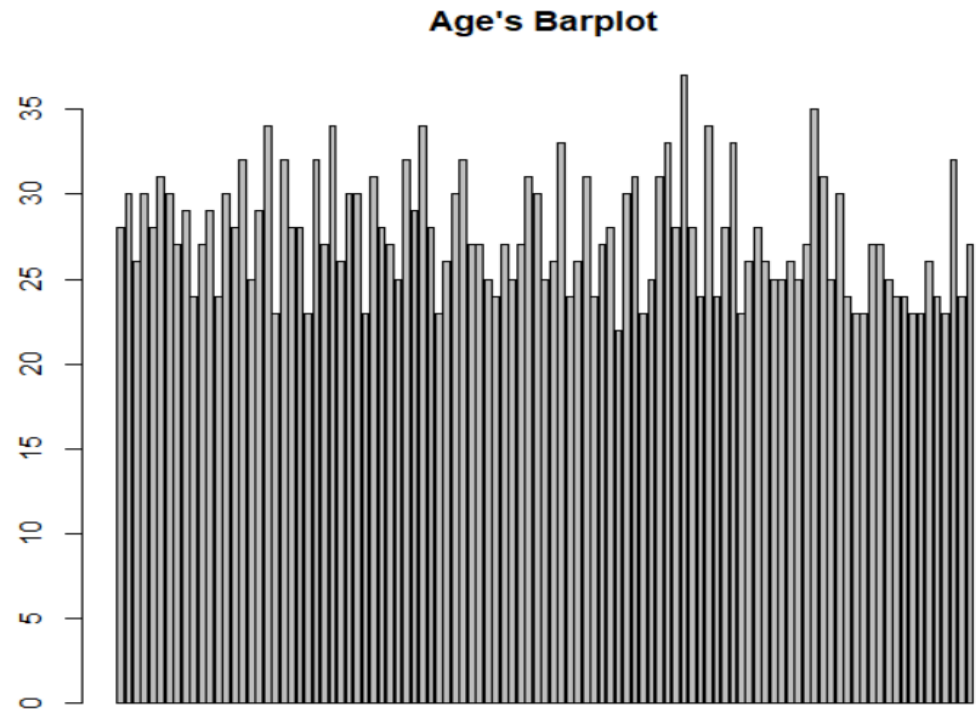

avg_game_time
10
20
30
40
(22,25.8]
(25.8,29.5]
(29.5,33.2]
(33.2,37]
Age4


Games
20
40
60
80
105
55
(22,25.8]
(25.8,29.5]
(29.5,33.2]
(33.2,37]
Age4



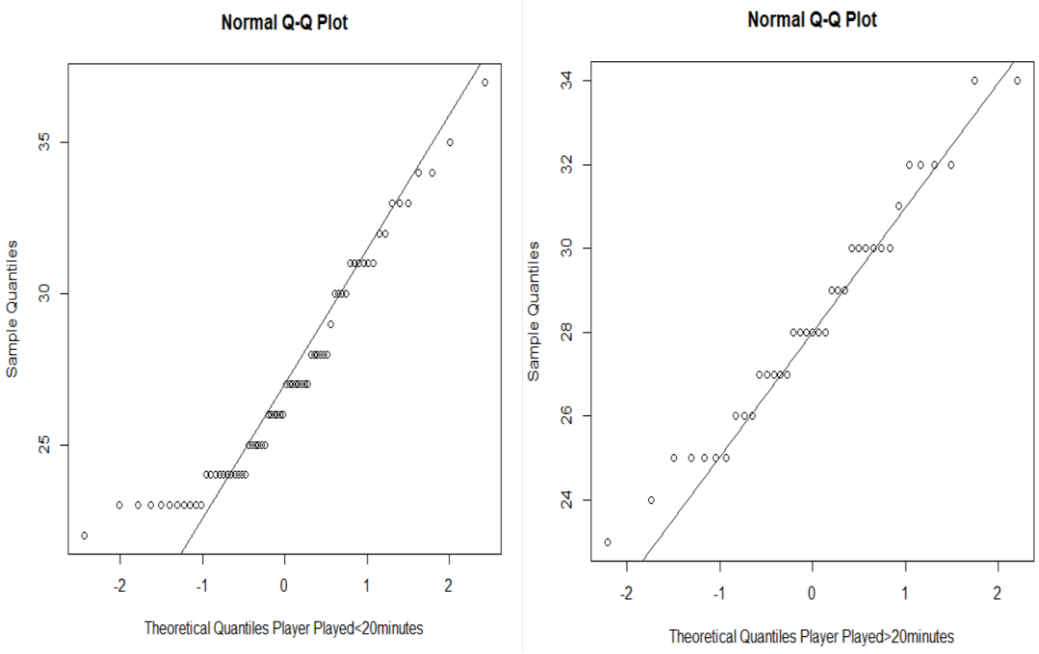
Σχήμα 9: Μέσα από τα παραπάνω scaterplots συγκρίνουμε τα επιμέρους τμήματα των τιμών της κατηγορικής μεταβλητής Age με τις τιμές των υπόλοιπων μεταβλητών.



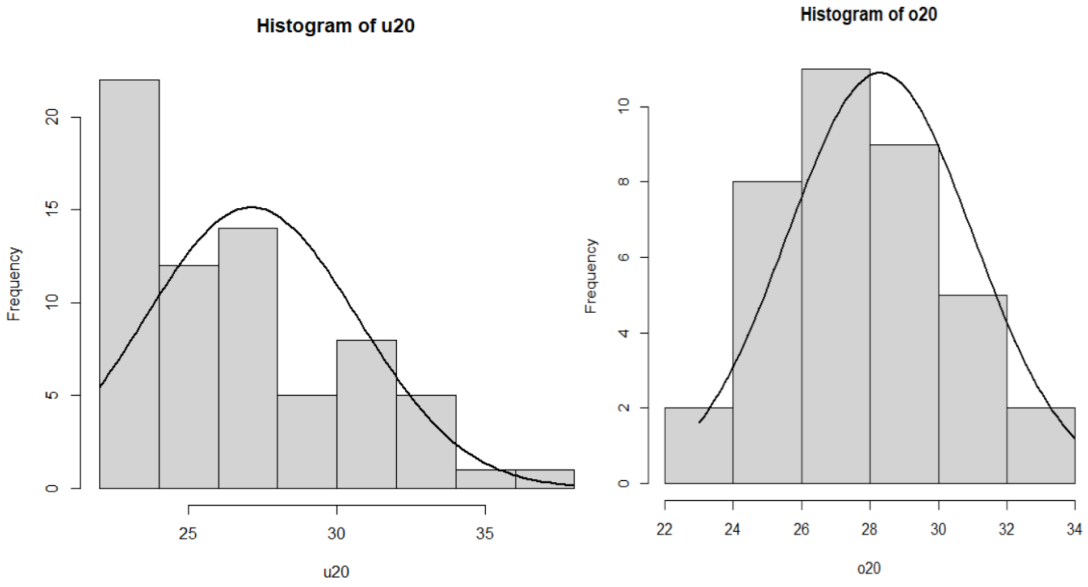


Σχήμα 10: Barplots που δείχνουν τις κατανομές των μεταβλητών Points & Age

Σύγκριση της ηλικίας των παικτών με την κατηγορική μεταβλητή του αν έπαιξε ο αθλητής πάνω από 20 λεπτά

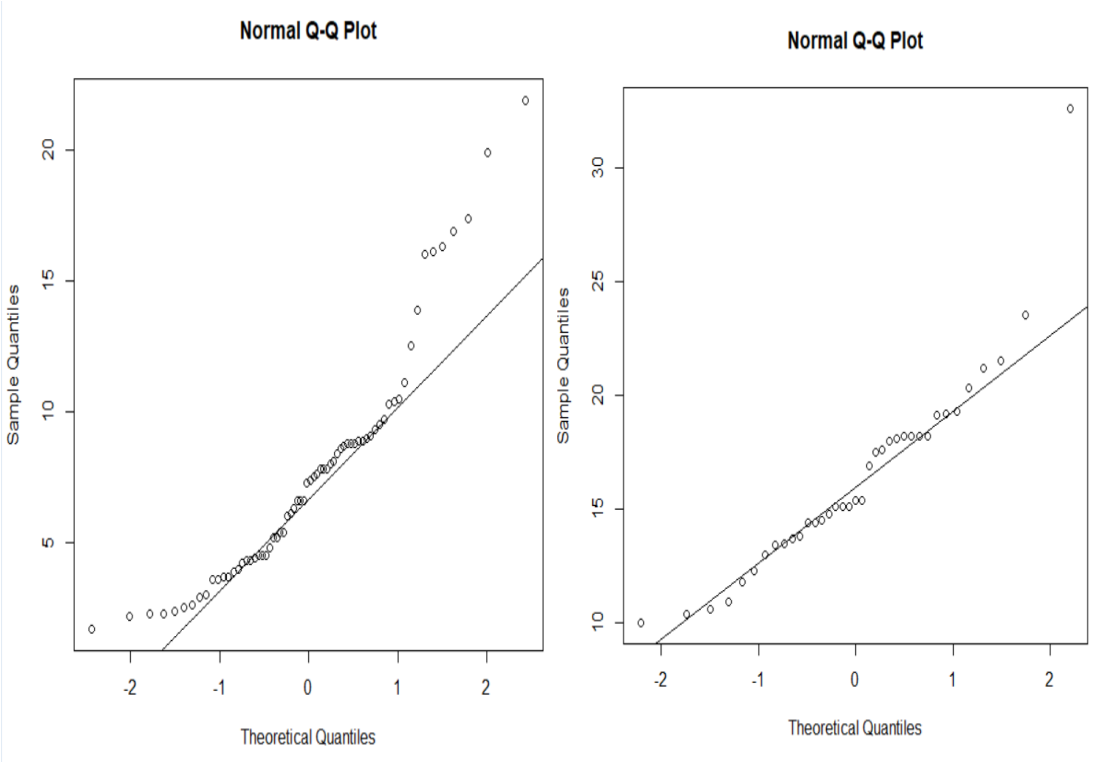


Σχήμα : QQplots για να δούμε την κατανομή της ηλικίας των παικτών για χρόνο συμμετοχής λιγότερο από 20 λεπτά και για χρόνο συμμετοχής μεγαλύτερο από 20 λεπτά.

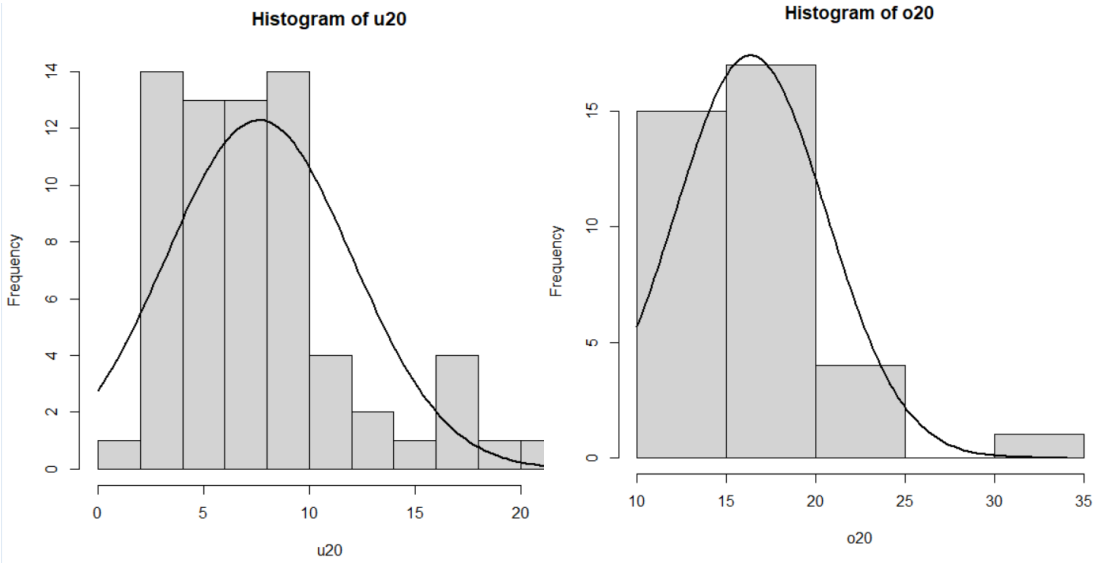


Σχήμα 11: Histograms για να δούμε την κατανομή της ηλικίας των παικτών για χρόνο συμμετοχής λιγότερο από 20 λεπτά και για χρόνο συμμετοχής μεγαλύτερο από 20 λεπτά.

Σύγκριση των πόντων των παικτών με την κατηγορική μεταβλητή του αν έπαιξε ο αθλητής πάνω από 20 λεπτά

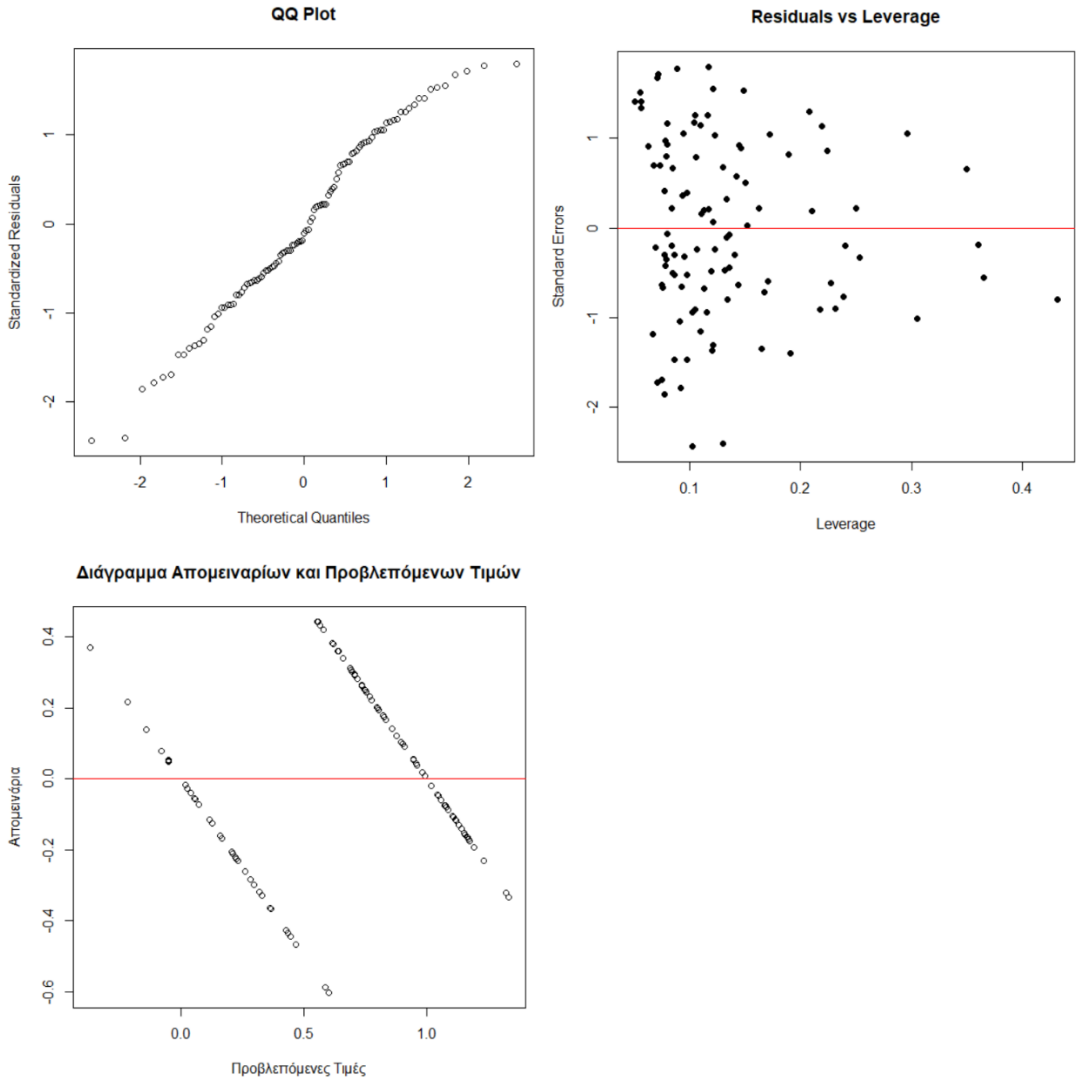


Σχήμα 12: QQplots για να δούμε την κατανομή των πόντων των παικτών για χρόνο συμμετοχής λιγότερο από 20 λεπτά και για χρόνο συμμετοχής μεγαλύτερο από 20 λεπτά.



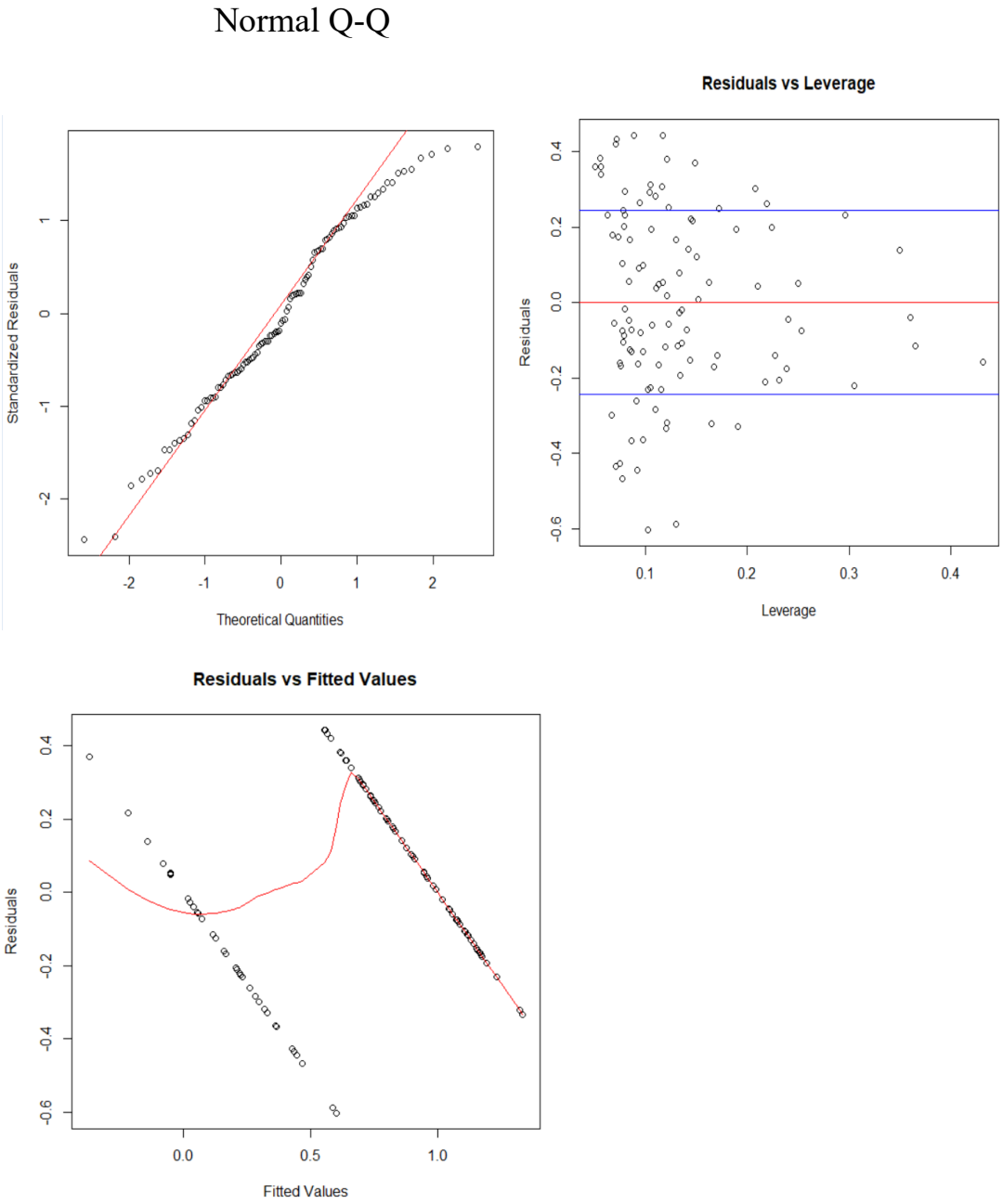
Σχήμα 13: Histograms για να δούμε την κατανομή των πόντων των παικτών για χρόνο συμμετοχής λιγότερο από 20 λεπτά και για χρόνο συμμετοχής μεγαλύτερο από 20 λεπτά.

Μοντέλο Παλινδρόμησης με όλες τις μεταβλητές



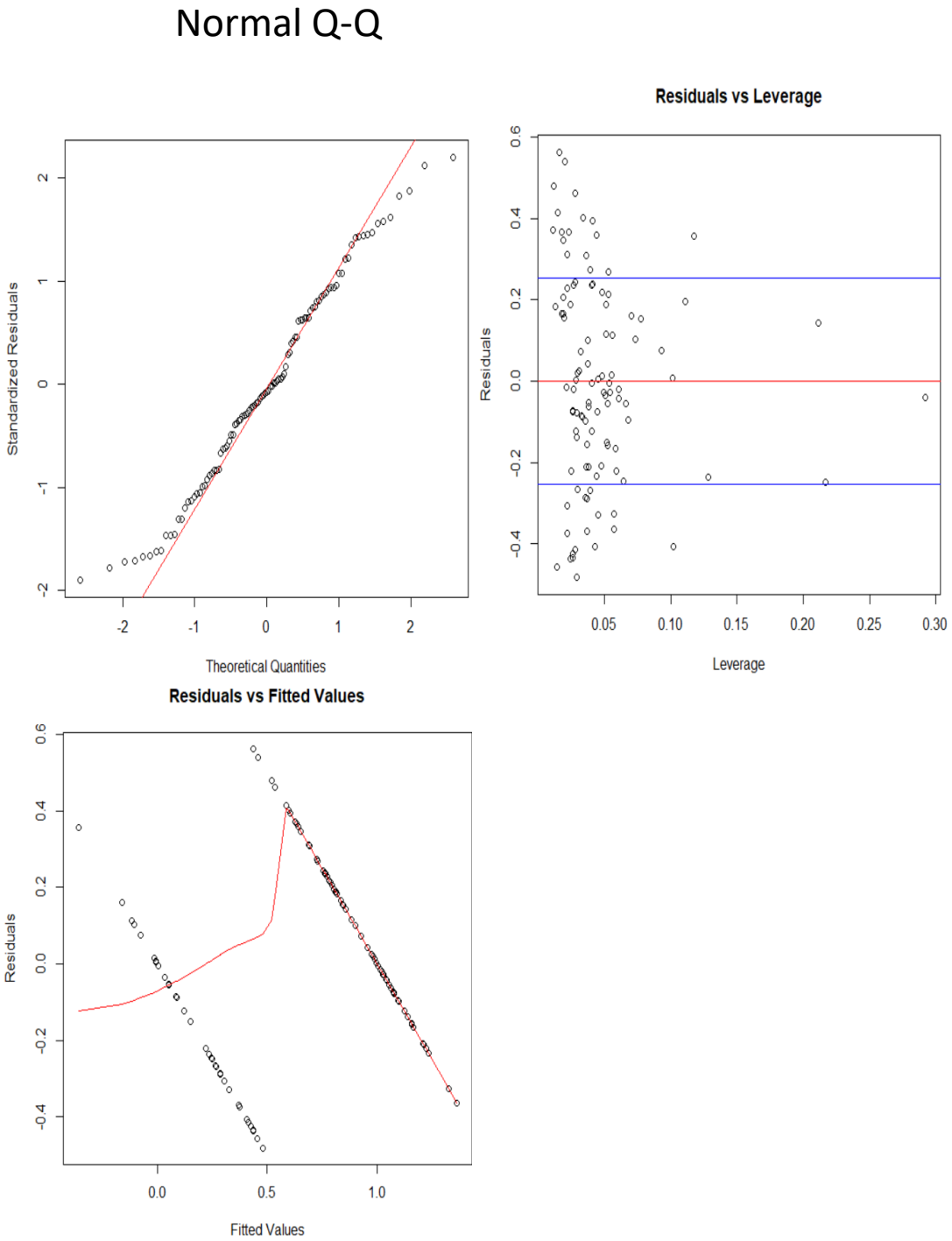
Σχήμα 14: Το plot για τους ελέγχους υποθέσεων του μοντέλου με όλες τις μεταβλητές

Μοντέλο Παλινδρόμησης με όλες τις μεταβλητές πλην της σταθερά



Σχήμα 15: Το plot για τους ελέγχους υποθέσεων του μοντέλου με όλες τις μεταβλητές πλην της σταθερά. Στο διάγραμμα Residuals-Leverage, η οριζόντια γραμμή με κόκκινο χρώμα είναι για τη μέση τιμή των υπολοίπων, ενώ οι οριζόντιες γραμμές με μπλε χρώμα είναι για το standard deviation των υπολοίπων

Το τελικό Μοντέλο Παλινδρόμησης μετά τις μεθόδους επιλογής μεταβλητών



Σχήμα 16: Το plot για τους ελέγχους υποθέσεων του τελικού μοντέλου μετά τις μεθόδους επιλογής μεταβλητών. Στο διάγραμμα Residuals-Leverage, η οριζόντια γραμμή με κόκκινο χρώμα είναι για τη μέση τιμή των υπολοίπων, ενώ οι οριζόντιες γραμμές με μπλε χρώμα είναι για το standard deviation των υπολοίπων