Ανάλυση Δεδομένων

Aleksandra Petukhova

3210229

Εισαγωγή – Περιγραφή Μελέτης και Προβλήματος

Σκοπός της μελέτης είναι η διερεύνηση των χαρακτηριστικών ενός ακινήτου που επηρεάζουν την τιμή πώλησής του και η πρόβλεψη της αναμενόμενης αξίας μεταπώλησης ενός σπιτιού με βάση τα συγκεκριμένα χαρακτηριστικά.

Για την εξαγωγή της μελέτης επιλέχθηκε σετ δεδομένων που αποτελείται από τυχαίο δείγμα 117 σπιτιών στην πόλη του Albuquerque, στο Νέο Μεξικό των ΗΠΑ. Τα στοιχεία καταγράφηκαν στο διάστημα μεταξύ 15 Φεβρουαρίου και 30 Απριλίου του 1993. Το σετ περιλαμβάνει πληροφορίες για την τιμή πώλησης του ακινήτου, την ετήσια φορολόγηση του, καθώς και διάφορα άλλα χαρακτηριστικά σπιτιών που καταγράφηκαν.

Πίνακας 1: Πίνακας Δεδομένων

Όνομα	**Τύπος Μεταβλητής**	**Σημασία**	**Τιμές**
price	αριθμητική	Αξία πώλησης (χιλιάδες δολάρια)	
sqft	αριθμητική	Εσωτερικό μέγεθος σπιτιού (τετραγωνικά πόδια)	
age	αριθμητική	Ηλικία σπιτιού (έτη)	
feats	αριθμητική	Αριθμός χαρακτηριστικών (π.χ. πλυντήριο, ψυγείο κλπ.)	0 – 11
ne	κατηγορική	Το σπίτι βρίσκεται στην ΒΑ πλευρά της πόλης	0: όχι, 1: ναι
cor	κατηγορική	Το σπίτι είναι γωνιακό	0: όχι, 1: ναι
resale	κατηγορική	Έχει πουληθεί ξανά στο παρελθόν	0: όχι, 1: ναι
tax	αριθμητική	Ετήσιος φόρος (δολάρια)	

Περιγραφική Ανάλυση

Για να εξετάσουμε τα βασικά χαρακτηριστικά των ποσοτικών μεταβλητών που περιλαμβάνονται στο σύνολο δεδομένων μας, ξεκινάμε με μία περιγραφική ανάλυση. Χρησιμοποιούμε στατιστικά μέτρα όπως η μέση τιμή, η διάμεσος, η τυπική απόκλιση και η ασυμμετρία για να αποκτήσουμε μία πρώτη εικόνα για την κατανομή των τιμών. Παρακάτω παρουσιάζονται τέσσερις μεταβλητές: PRICE, SQFT, AGE και TAX, όπως φαίνονται και στον Πίνακα 2.

Ξεκινώντας με τη μεταβλητή PRICE, η οποία αντιπροσωπεύει την τιμή των ακινήτων, βλέπουμε ότι υπάρχουν συνολικά 117 παρατηρήσεις και δεν υπάρχουν καθόλου απούσες τιμές. Η μέση τιμή είναι 1062.74, ενώ η διάμεσος είναι μικρότερη (960), κάτι που μας δείχνει ότι υπάρχουν μερικές υψηλότερες τιμές που τραβούν τη μέση τιμή προς τα πάνω. Αυτό επιβεβαιώνεται και από τη θετική τιμή της ασυμμετρίας (1.38), η οποία υποδηλώνει δεξιά κατανομή. Το εύρος τιμών είναι μεγάλο (540 έως 2150), ενώ και το ενδοτεταρτημοριακό εύρος είναι 420, δείχνοντας σημαντική διασπορά στις τιμές.

Η μεταβλητή SQFT, που αναφέρεται στα τετραγωνικά μέτρα των ακινήτων, επίσης περιλαμβάνει 117 παρατηρήσεις χωρίς απουσίες. Η μέση τιμή είναι 1653.85 και η διάμεσος είναι 1549, επομένως και εδώ υπάρχει μια ελαφρά δεξιά

ασυμμετρία (skew = 1.19). Οι τιμές κυμαίνονται από 837 έως 3750, με IQR ίσο με 614, δείχνοντας και πάλι ότι υπάρχει σημαντική μεταβλητότητα στο μέγεθος των ακινήτων.

Στην περίπτωση της AGE, που αφορά την ηλικία των ακινήτων, παρατηρούμε ότι υπάρχουν αρκετές απούσες τιμές (περίπου 42% των παρατηρήσεων λείπουν). Από τις τιμές που διαθέτουμε, η μέση ηλικία είναι 14.97 έτη και η διάμεσος είναι ελαφρώς μικρότερη (13 έτη), ενώ και εδώ εντοπίζεται δεξιά ασυμμετρία (1.27). Το εύρος ηλικιών είναι από 1 έως 53 έτη, με IQR ίσο με 13.5.

Τέλος, η μεταβλητή TAX, η οποία αναφέρεται στον φόρο ακινήτου, έχει 107 παρατηρήσεις, με ένα ποσοστό απουσιών περίπου 8.5%. Η μέση τιμή είναι 793.49, ενώ η διάμεσος είναι λίγο μικρότερη (731), δείχνοντας και εδώ ήπια δεξιά ασυμμετρία (skew = 1.06). Το εύρος τιμών είναι από 223 έως 1765 και το IQR είναι 319, που δείχνει επίσης αρκετή διακύμανση.

Παρατηρούμε λοιπόν ότι και οι τέσσερις ποσοτικές μεταβλητές παρουσιάζουν δεξιά ασυμμετρία,αυτό τονίζεται και διαγραμματικά βλ. σχήμα 1, κάτι που σημαίνει ότι υπάρχουν υψηλές τιμές που επηρεάζουν τη μορφή της κατανομής. Αν και δεν μπορούμε να πούμε με βεβαιότητα ότι οι κατανομές τους αποκλίνουν έντονα από την κανονική, η παρουσία ασυμμετρίας αποτελεί ένδειξη ότι ίσως χρειάζεται μετασχηματισμός των δεδομένων σε επόμενα στάδια της ανάλυσης.

Η περιγραφική ανάλυση των κατηγορικών μεταβλητών δείχνει πως οι μεταβλητές AREA, COR και RESALE δεν έχουν ομοιόμορφη κατανομη. Η μεταβλητή **FEATS** (αριθμός παροχών ή χαρακτηριστικών) εμφανίζει κατανομή τύπου καμπάνας, με τη συχνότερη τιμή να είναι τα **4 χαρακτηριστικά**, ενώ οι περισσότερες κατοικίες διαθέτουν από 3 έως 5. Τιμές μικρότερες του 2 και μεγαλύτερες του 6 εμφανίζονται σπάνια, κάτι που υποδεικνύει συμμετρική κατανομή γύρω από τη μέση τιμή. Η μεταβλητή **AREA**, η οποία πιθανώς υποδεικνύει κατηγορία ζώνης, παρουσιάζει **ασύμμετρη κατανομή**, καθώς η πλειοψηφία των παρατηρήσεων ανήκει στην κατηγορία 1. Η μεταβλητή **COR**, που πιθανότατα σχετίζεται με γωνιακότητα οικοπέδου, δείχνει ότι περίπου τα δύο τρίτα των κατοικιών **δεν είναι γωνιακά** (COR = 0), ενώ το υπόλοιπο ένα τρίτο αφορά γωνιακά ακίνητα. Παρομοίως, η μεταβλητή **RESALE**, που δηλώνει αν ένα ακίνητο αποτελεί μεταπώληση ή όχι, καταγράφει σαφή υπεροχή της τιμής 0, δηλαδή **πρωτογενείς πωλήσεις**, έναντι των μεταπωλήσεων. Συνολικά, οι μεταβλητές **AREA**, **COR** και **RESALE** είναι ανισοκατανεμημένες, γεγονός που θα πρέπει να ληφθεί υπόψη στην επόμενη φάση του υποδείγματος, καθώς η μειωμένη ποικιλία κατηγοριών ενδέχεται να επηρεάσει τη στατιστική ισχύ. Από την άλλη, η μεταβλητή **FEATS** εμφανίζει ισορροπημένη κατανομή και πιθανόν να συμβάλει σημαντικά στην ερμηνεία της τιμής του ακινήτου (`PRICE`).

Η περιγραφική ανάλυση των κατηγορικών μεταβλητών δείχνει πως οι μεταβλητές **AREA**, **COR** και RESALE δεν έχουν ομοιόμορφη κατανομη με κάποιες κατηγορίες να κυριαρχούν. Αυτό μπορεί να έχει επίπτωση στη στατιστική ισχύ κατά την ενταξή τους σε μοντέλο πρόβλεψης. Αντίθετα, η μεταβλητή **FEATS** παρουσιάζει αρκετά κανονική κατανομή και θα μπορούσε να έχει σημαντική ερμηνευτική αξία, εφόσον εξεταστεί στατιστικά με την τιμή,για λεπτομέρειες βλ. σχήμα 2 στο Παράρτημα.

Οι πληροφορίες αυτές είναι χρήσιμες προκειμένου να αποφασίσουμε ποιοι στατιστικοι ελεγχοι είναι κατάλληλοι για τις συγκεκριμένες μεταβλητές. Στην μεταβλητή AGE συγκεντρώσαμε έναν μεγάλο αριθμό άγνωστων/ελλιπών τιμών

Πίνακας 2: Πίνακας περιγραφικών μέτρων ποσοτικών μεταβλητών

Μεταβλητή	label	n	NA.prc	mean	sd	se	md	trimmed	range	iqr	skew
2	PRICE	117	0.00	1062.74	380.44	35.17	960	1003.94	1610 (540-2150)	420.0	1.38
3	SQFT	117	0.00	1653.85	523.72	48.42	1549	1594.33	2913 (837-3750)	614.0	1.19
1	AGE	68	41.88	14.97	12.67	1.54	13	13.14	52 (1-53)	13.5	1.27
4	TAX	107	8.55	793.49	308.18	29.79	731	764.48	1542 (223-1765)	319.0	1.06

βλ. πίνακα 5. Θα εξετάσουμε όλο το σετ δεδομένων για παρουσία τέτοιων τιμών.

Στον παρακάτω πίνακα (βλ. πίνακας 3) παρατηρούμε ότι υπάρχουν ελλιπείς τιμές σε δύο μεταβλητές: στην AGE και στην TAX.Το ποσοστό αυτών αντιστοιχεί στο 41.9% των παρατηρήσεων και 8.55% αντίστοιχα.

Η ύπαρξη τόσο μεγάλου ποσοστού ελλιπών τιμών στη μεταβλητή AGE αποτελεί πρόβλημα για την ανάλυση. Αν διατηρήσουμε αυτή τη μεταβλητή και επιλέξουμε να αναλύσουμε μόνο τις παρατηρήσεις όπου υπάρχουν διαθέσιμες τιμές για όλες τις μεταβλητές (μέθοδος listwise deletion), τότε αυτομάτως χάνουμε σχεδόν τις μισές παρατηρήσεις. Αυτό μπορεί να μειώσει τη στατιστική ισχύ της ανάλυσής μας, καθώς και να δημιουργήσει σφάλμα στα αποτελέσματα, εάν οι τιμές που λείπουν δεν είναι τυχαίες (δηλαδή αν υπάρχει κάποιο μοτίβο).

Από την άλλη πλευρά, αν επιλέξουμε να εξαιρέσουμε εντελώς τη μεταβλητή AGE από το μοντέλο μας, τότε μπορούμε να διατηρήσουμε περισσότερες παρατηρήσεις. Αυτό είναι μπορεί να είναι θετικό διότι ενδεχομένως να οδηγήσει σε στατιστικά πιο σταθερά αποτελέσματα, ειδικά αν η μεταβλητή AGE δεν έχει ισχυρή συσχέτιση με τη μεταβλητή εξόδου του μοντέλου. Ωστόσο, υπάρχει και το ενδεχόμενο να χάσουμε πολύτιμη πληροφορία, αν η ηλικία του ακινήτου είναι ένας σημαντικός προβλεπτικός παράγοντας.

Για να αποφασίσουμε αν η εξαίρεση της μεταβλητής AGE αλλάζει τα αποτελέσματα, μπορούμε να εκτελέσουμε την ανάλυσή μας δύο φορές: μία φορά με όλες τις μεταβλητές (και άρα με λιγότερες παρατηρήσεις) και μία χωρίς την μεταβλητή AGE (αλλά με περισσότερες παρατηρήσεις). Αν διαπιστώσουμε ότι τα αποτελέσματα δεν αλλάζουν σημαντικά, τότε ίσως να είναι προτιμότερο να την αφαιρέσουμε. Αν όμως υπάρχουν ουσιαστικές διαφοροποιήσεις, θα πρέπει να εξετάσουμε μεθόδους αντιμετώπισης των ελλιπών τιμών, όπως η πολλαπλή συμπλήρωση (multiple imputation).

Θα προβούμε στην μελέτη των ακραίων τιμών, διαγραμματικά και για καλύτερη κατανόηση οι τιμές αναδυκνύονται συνοπτικά στον παραπάνω πίνακα (βλ. πίνακα 3).

Οι ακραίες τιμές μπορούν να είναι λάθη καταγραφής και σε αυτή την περίπτωση τα διορθώνουμε ή τα αφαιρούμε. Αν είναι σπάνιες αλλά πραγματικές τιμές, κρίνουμε ανάλογα με τον σκοπό της ανάλυσης. Αν επηρεάζουν σημαντικά τα αποτελέσματα ή «τραβούν» το μοντέλο προς μη ρεαλιστικές προβλέψεις, μπορεί να είναι απαραίτητο να τις εξαιρέσουμε.

Εντοπίζοντας τις μεταβλητές που περιέχουν ακραίες τιμές και εξετάζοντάς τες σε σχέση με το συνολικό εύρος των τιμών, διαπιστώνουμε ότι όλες εμπίπτουν σε ρεαλιστικά πλαίσια. Για παράδειγμα, στη μεταβλητή PRICE, τιμές που προσεγγίζουν τα 2 εκατομμύρια θεωρούνται εύλογες στην πραγματική αγορά. Αντίστοιχα, και οι υπόλοιπες ποσοτικές

Πίνακας 3: Πίνακας ακραίων τιμών

PRICE	SQFT	AGE	TAX
2050	2921	53	1639
2080	2931	41	1635
2150	2848	46	1732
2150	3750	43	1534
1999		40	1765
1900		40	1487
2150		45	
2100			
1844			

μεταβλητές παρουσιάζουν τιμές που κρίνονται αποδεκτές. Συνεπώς, δεν κρίνεται απαραίτητο να εξαιρέσουμε τις ακραίες τιμές.

Σχέσεις ανά δύο

Κυριότερος στόχος είναι να διαπισώσουμε ποιοι παράγοντες επηρεάζουν την τελική τιμή πώλησης ακινήτων στην αγορά, προκειμένου να μπορέσουμε να προβλέψουμε τυχόν τιμές γνωρίζοντας κοποια χαρακτηριστικά σπιτιού. Για να επιτευχθεί ο στόχος αυτός θέτουμε ερωτήματα όπως; Η περιοχή προβλέπει την τιμή; Συσχετίζεται ο φόρος με την τιμή;Τα παλαιότερα σπίτια κοστίζουν λιγότερο; Από τα ερωτήματα που αφορούν την τιμή προκύπτουν οι παρακάτω σχέσεις, που πρέπει να μελετηθουν, δηλαδή πρέπει να εκτιμηθεί αν υπάρχει σχέση μεταξύ της ποσοτικής μεταβλητής-τιμής και κάθε άλλης τιμής παρούσας στο σετ δεδομένων.Παράλληλα όμως, θα εξεταστούν και τυχόν σχέσεις μεταξύ ποιοτικών μεταβλητών.Επιπλέον με βάση των , θα αναλυθεί πώς επηρεάζεται ο φόρος από το μέγεθος ακινήτου.

Σχέσεις προς μελέτη

- Τιμή – Τοποθεσία
- Τιμή – Εσωτερικό μέγεθος σπιτιού
- Τιμή – Χαρακτηριστικά
- Τιμή – Φόρος
- Τιμή – Ηλικία ακινήτου
- Τιμή – Πώληση στο παρελθόν
- Φόρος – Εσωτερικό μέγεθος σπιτιού

Σε κάθε ζεύγος μεταβλητών, εφαρμόστηκαν κατάλληλα tests συνοδευόμενα από διαγράμματα για οπτικοποίηση σχέσεων. Αναλυτικότερα, για να διαπιστώσουμε καλύτερα τις σχέσεις ποσοτικών μεταβλητών, χρησιμοποιήθηκε pearson correlation test. Το αποτέλεσμα έδειξε ότι η τιμή **PRICE** συσχετίζεται θετικά με το εσωτερικό μέγεθος σπιτιού **SQFT** και επίσης έχει ισχυρή θετική συσχετίση με τον ετήσιο φόρο **TAX**, για λεπτομέρειες συχέτισης βλ. σχήμα 5 στο Παράρτημα.

Για την σχέση τιμής και χαρακτηριστικών χρησιμοποιήθηκε Kruskal Wallis test, διότι η μεταβλητή χαρακτηριστικά είναι πολυ επίπεδη, και ό έλεγχος συνοδεύεται από ραβδογράμματα. για λεπτομέρειες βλ. σχήμα 6 στο Παράρτημα.Παρατηρούμε ότι απορρίπτεται η μηδενική υπόθεση(p p-value = 0.001757< 0.05) αυτό υποδηλώνει ότι τα χαρακτηριστικά ακινήτων σχετίζονται με διαφορές στην τιμή.

Για την σχέση τιμής και τοποθεσίας, χρησιμοποιήθηκε Wilcoxon εφόσον η μεταβλητή τοποθεσία αποτελεί μια δίτιμη μεταβλητή και η προϋποθεση κανονικότητας απορρίπτεται και για τις 2 τιμές 0 και 1(p-value = 0.0001416 και p-value = 5.789e-07). Το αποτέλεσμα δείχνει την απουσία καποιας σχέσης μεταξύ των συγκεκριμένων μεταβλητών,για λεπτομέρειες συχέτισης βλ. σχήμα 4 στο Παράρτημα.

Για την σχέση τιμής και γεγονότος πώλησης στο παρελθόν, επαναλήφθηκε η ίδια διαδικασία και το αποτέλεσμα του Wilcoxon (p-value = 0.8044) έδειξε ότι δεν υπάρχει κάποια σχέση μεταξύ τιμής και του γεγονότος ότι πουλήθηκε/ δεν πουλήθηκε το ακίνητο στο παρελθόν, για λεπτομέρειες συχέτισης βλ. σχήμα 8 στο Παράρτημα.

΄Εχοντας κάνει τους απαραίτητους ελέγχους εμηνεύσει και τα αντίστοιχα διαγράμματα καταλήγουμε ότι οι σημαντικότερες σχέσεις που ενδέχεται να μοντελοποιούν την τιμή και μπορούν να μελετηθούν περεταίρω είναι:

- Τιμή – Εσωτερικό μέγεθος σπιτιού
- Τιμή – Χαρακτηριστικά
- Τιμή – Φόρος

Πίνακας 4: Pearson correlation matrix with p-values

	PRICE	SQFT	TAX
PRICE			
SQFT	0.845 (< .001)		
TAX	0.876 (< .001)	0.859 (< .001)	

Όσον αφορά την μεταβλητή TAX (ετήσιος φόρος) παρουσιάζει ισχυρή γραμμική συσχέτιση τόσο με την τιμή αγοράς της κατοικίας (PRICE) όσο και με το εμβαδόν του ακινήτου (SQFT). Όπως φάνηκε στα αντίστοιχα διαγράμματα (για λεπτομέρειες συσχέτισης βλ. πίνακες A4 και A5 στο Παράρτημα), η αύξηση της τιμής συνδέεται με αντίστοιχη αύξηση του ετήσιου φόρου, γεγονός που υποδηλώνει ότι ο φόρος αντανακλά τη συνολική αξία της ακίνητης περιουσίας.

Παράλληλα, διαπιστώνεται ότι και το εμβαδόν του ακινήτου επηρεάζει θετικά τον φόρο, γεγονός αναμενόμενο, καθώς μεγαλύτερα ακίνητα τείνουν να έχουν υψηλότερη αγοραία αξία και, συνεπώς, μεγαλύτερη φορολογική επιβάρυνση.Τα αποτελέσματα αυτά καθιστούν τη μεταβλητή TAX ιδιαίτερα χρήσιμη στην πρόβλεψη ή την κατηγοριοποίηση ακινήτων με βάση χαρακτηριστικά που σχετίζονται με την αγορά,

Προβλεπτικά ή Ερμηνευτικά μοντέλα

Οι ανωτέρω έλεγχοι μας καθοδήγησαν στην επιλογή των μεταβλητών που επηρεάζουν καθοριστικά την τιμή ενός ακινήτου. Στη συνέχεια, προχωρούμε στην κατασκευή στατιστικού μοντέλου πρόβλεψης της αξίας μεταπώλησης κατοικιών, το οποίο θα βασιστεί στις πλέον σημαντικές μεταβλητές. Σκοπός είναι η ακριβής αποτύπωση της αξίας μέσω του μοντέλου. Ωστόσο, σε πρώτη φάση, κατασκευάζουμε το μοντέλο έτσι ώστε να περιέχει όλες τις μεταβλητές και βλέπουμε αν τηρούνται οι υποθέσεις για την εκτέλεση πολλαπλής παλινδρόμησης.

Για την προϋποθεση κανονικότητας χρησιμοποιήθηκε ο έλεγχος κανονικότητας Shapiro–Wilk (W = 0.966, p = 0.065) δεν έδειξε στατιστικά σημαντική απόκλιση από την κανονικότητα. Συνεπώς, τα κατάλοιπα του μοντέλου κατανέμονται περίπου κανονικά.

Το διάγραμμα καταλοίπων έναντι των προσαρμοσμένων τιμών δεν παρουσίασε μοτίβο μεταβαλλόμενης διασποράς, κάτι που υποστηρίζει την παραδοχή της ομοσκεδαστικότητας (ίσης διασποράς σφαλμάτων),για λεπτομέρειες βλ. σχήμα 10 στο Παράρτημα.

Ωστόσο, η εφαρμογή στατιστικού έλεγχου Levene Το αποτέλεσμα του ελέγχου έδειξε ότι τα κατάλοιπα δεν εμφανίζουν σταθερή διασπορά, γεγονός που αποτελεί ένδειξη ετεροσκοεδαστικότητας (F = 6.0347, p = 0.001129),απορρίπτεται η μηδενικής υπόθεσης,άρα τα κατάλοιπα δεν εμφανίζουν σταθερή διασπορά.Η παρουσία ετεροσκεδαστικότητας μπορεί να οδηγήσει σε αναξιόπιστες εκτιμήσεις.

Για την αντιμετώπιση του ζητήματος, εφαρμόστηκε λογαριθμικός μετασχηματισμός στην εξαρτημένη μεταβλητή PRICE, με σκοπό τη σταθεροποίηση της διασποράς και την ενίσχυση της κανονικότητας των καταλοίπων.

Το νέο μοντέλο εκτιμήθηκε με τη μορφή:

$$\log(\texttt{PRICE}) \sim \texttt{SQFT} + \texttt{FEATS} + \texttt{AREA} + \texttt{AGE} + \texttt{TAX} + \texttt{COR} + \texttt{RESALE}$$

Μετά τον μετασχηματισμό, ο έλεγχος Levene επαναλήφθηκε(F = 2.0173, p = 0.1207), γεγονός που δηλώνει ότι δεν υπάρχει στατιστικά σημαντική διαφορά στις διασπορές των καταλοίπων μεταξύ των ομάδων. Συνεπώς, η υπόθεση ομοιοσκεδαστικότητας **δεν απορρίπτεται** και η υπόθεση ισότητας διασποράς μπορεί πλέον να θεωρηθεί ικανοποιημένη.

Ακολουθεί ο έλεγχος ανεξαρτησίας κατάλοιπων Durbin–Watson (D–W = 1.93 με p = 0.52) Συνεπώς, δεν υπάρχει στατιστικά σημαντική αυτοσυσχέτιση των σφαλμάτων, και η υπόθεση της ανεξαρτησίας ικανοποιείται.

Αφού επιβεβαιώθηκε η εγκυρότητα του μοντέλου ως προς τις βασικές υποθέσεις της γραμμικής παλινδρόμησης (Κανονικότητα, Ομοσκεδαστικότητα, Ανεξαρτησία) προχωρούμε στην επόμενη φάση του εντοπισμού των σημαντικότερων μεταβλητών που επηρεάζουν την τιμή μεταπώλησης, και τη δημιουργία ενός ισχυρού και απλού μοντέλου παλινδρόμησης.

Αρχικά θα επαληθεύσουμε το γεγονός ότι οι μεταβλητές Χαρακτηριστικά, Φόρος και Εσωτερικό μέγεθος σπιτιού όντως μπορούν να αποτελέσουν το μοντέλο πολλαπλής παλινδρόμησης.Ο έλεγχος θα επιτευχθεί με σταδιακή προσθήκη των τριών μεταβλητών. και ταυτόχρονα μελετηθεί η συμπεριφορά του μοντέλου ως προς τις διαθέσιμες μεταβλητές.

Παρατηρούμε ότι το μοντέλο εξηγεί το 86,2% της διακύμανσης στις τιμές μεταπώλησης.

Ο συντελεστής της μεταβλητής `TAX` βρέθηκε στατιστικώς σημαντικός ($\beta = 3.91 \times 10^{-4}$, $p = 0.000372$), υποδεικνύοντας ότι η φορολογία σχετίζεται θετικά και σημαντικά με τη λογαριθμισμένη τιμή του ακινήτου. Επιπλέον, η ανάλυση διακύμανσης τύπου II επιβεβαίωσε τη σημαντική συμβολή της `TAX` στο υπόδειγμα ($F = 14.17$, $p = 0.00037$). Ως αρχικό βήμα, εκτιμήθηκε γραμμικό μοντέλο με λογαριθμισμένη εξαρτημένη μεταβλητή `log(PRICE)` και την`SQFT`.Διαπιστώθηκε ότι η μεταβλητή **SQFT** (τετραγωνικά μέτρα) είναι στατιστικά σημαντική ($p < 2 \times 10^{-16}$) και εξηγεί περίπου το 80% της διακύμανσης στην τιμή ($R^2 = 0.8045$). Με την προσθήκη της μεταβλητής **FEATS**, το R^2 αυξήθηκε ελάχιστα ($R^2 = 0.8065$), επίσης η ίδια η μεταβλητή δεν είναι στατιστικά σημαντική ($p = 0.425$),το τεστ ANOVA τύπου II το επιβεβαίωσε ($F(1, 63) = 261.93$ με $p < 2 \times 10^{-16}$). Αντιθέτως, στο τρίτο μοντέλο, η προσθήκη της μεταβλητής **TAX** οδήγησε σε αισθητή βελτίωση($R^2 = 00.8425$), ενώ η **TAX** παρουσίασε στατιστική σημαντικότητα ($p = 0.000372$), με από ANOVA τύπου II ($F = 14.17$, $p = 0.00037$). Η **FEATS** παρέμεινε μη σημαντική και σε αυτό το μοντέλο. Συνεπώς, καταλήγουμε στο συμπέρασμα ότι οι μεταβλητές **SQFT** και **TAX** αποτελούν τους καταλληλότερους προβλεπτικούς παράγοντες για την τιμή ακινήτου, βάσει της υψηλής προσαρμοστικής ικανότητας του μοντέλου και της στατιστικής σημαντικότητας που επιβεβαιώνεται τόσο από τα αποτελέσματα των συντελεστών όσο και από την ανάλυση διακύμανσης.

Προσθήκη της μεταβλητής AGE στο μοντέλο

Συνεχίζοντας την διαδικασία πρόσθεσης μεταβλητών, εξετάζουμε την επίδραση μεταβλητής **AGE** στο προηγούμενο μοντέλο.Δεν φαίνεται στατιστικά σημαντική, συγκεκριμένα($p = 0.2939$).Παρουσιάζει επίσης μια ελάχιστη αύξηση ($R^2 = 0.843$).Επομένως δεν μπορούμε να συμπεριλάβουμε την συγκεκριμένη μεταβλητή, δεν ενισχύει ουσιαστικά το προβλεπτικό μοντέλο, γεγονός που αποδεικνύεται και από ANOVA τύπου II.

Προσθήκη της μεταβλητής AREA στο μοντέλο

Προσθέτουμε τη μεταβλητή **AREA** στο υπάρχον μοντέλο. Η μεταβλητή αυτή δεν αναδεικνύεται στατιστικά σημαντική, καθώς η τιμή-p είναι αρκετά υψηλή ($p = 0.7567$). Η προσθήκη της συνοδεύεται από μια αμελητέα αύξηση του συντελεστή προσδιορισμού ($R^2 = 0.8446$).Συνεπώς, η μεταβλητή **AREA** δεν ενισχύει ουσιαστικά το προβλεπτικό μοντέλο, όπως επιβεβαιώνεται και από τα αποτελέσματα της ανάλυσης διασποράς τύπου II (ANOVA).

Προσθήκη της μεταβλητής COR στο μοντέλο

Στο επόμενο στάδιο της ανάλυσης προστίθεται η μεταβλητή **COR** στο υπάρχον μοντέλο. Η προσθήκη της **COR** οδήγησε σεμικρή αλλά αξιοσημείωτη αύξηση στον συντελεστή προσδιορισμού ($R^2 = 0.859$).Η **COR** είναι στατιστικά σημαντική ($p = 0.0173$).Συμπεράνουμε ότι η **COR** συνεισφέρει στη βελτίωση του μοντέλου πρόβλεψης, κάτι που επιβεβαιώνεται και από τα αποτελέσματα της ανάλυσης διασποράς τύπου II (ANOVA), όπου παρατηρείται επίσης στατιστικά σημαντική επίδραση ($p = 0.0173$).

Προσθήκη της μεταβλητής RESALE στο τελικό μοντέλο}

Στο τελικό στάδιο της ανάλυσης εξετάστηκε η μεταβλητή **RESALE**, η οποία προστέθηκε στο ήδη εκτεταμένο μοντέλο με τις μεταβλητές **SQFT**, **TAX**, **AGE**, **AREA**, **COR** και **FEATS**. Η μεταβλητή αυτή δεν παρουσιάζει στατιστική σημαντικότητα, καθώς η τιμή-p είναι 0.1812, δηλαδή αρκετά υψηλότερη από το αποδεκτό όριο σημαντικότητας. Η προσθήκη της οδηγεί σε οριακή αύξηση στον συντελεστή προσδιορισμού ($R^2 = 0.8633$), η οποία δεν είναι αρκετή ώστε να δικαιολογήσει τη διατήρησή της στο μοντέλο.Συνεπώς, η μεταβλητή αυτή δεν ενισχύει το προβλεπτικό μοντέλο.

Η τελική μορφή του μοντέλου περιλαμβάνει τις μεταβλητές **SQFT** (τετραγωνικά μέτρα), **TAX** (φόρος) και **COR** (ένδειξη γωνιακού οικοπέδου), καθώς αποδείχθηκαν στατιστικά σημαντικές για την πρόβλεψη της τιμής του ακινήτου (*PRICE*). Το τελικό μοντέλο εμφανίζει υψηλή προσαρμοστικότητα, με προσαρμοσμένο συντελεστή προσδιορισμού $R^2_{\text{adj}} = 0.8483$, γεγονός που σημαίνει ότι εξηγεί το 84.8% της διακύμανσης στο μοντέλο.Όλοι οι συντελεστές είναι στατιστικά σημαντικοί.

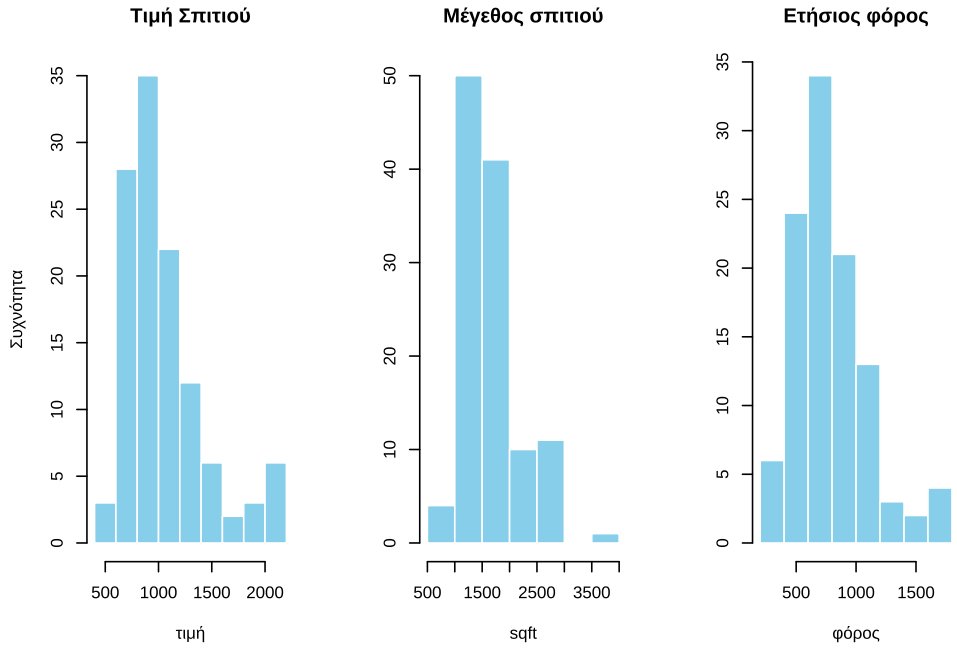
Εξίσωση Τελικού Μοντέλου

$$\log(\text{PRICE}) = 6.124 + 0.0003011 \cdot \text{SQFT} + 0.0003852 \cdot \text{TAX} + 0.1030 \cdot \text{COR} \tag{1}$$

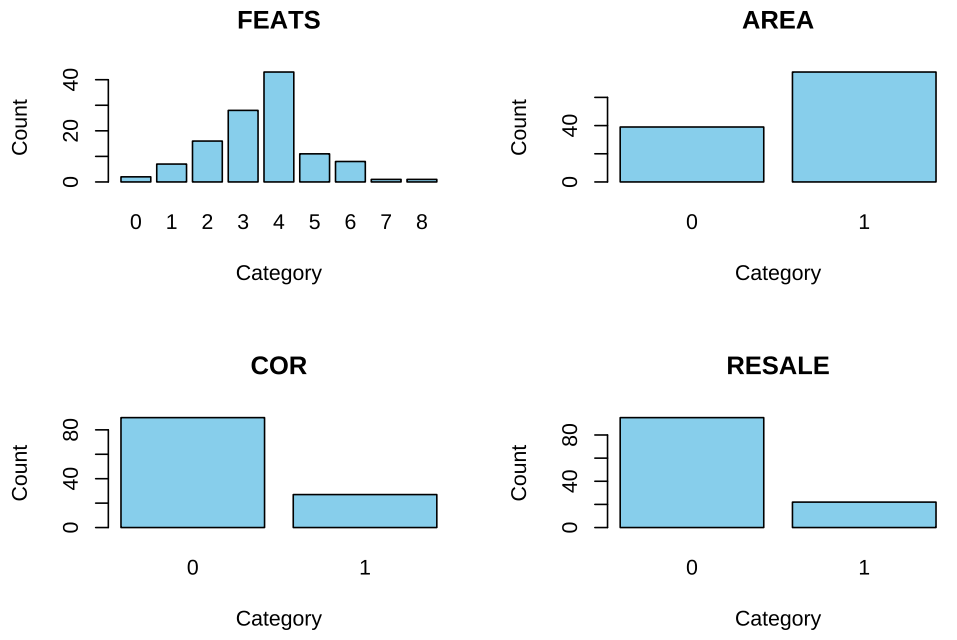
Συμπεράσματα και συζήτηση

Καταλήξαμε στην ανάπτυξη ενός ικανοποιητικού μοντέλου και σημαντικό είναι το γεγονός ότι έχει ισχυρή προβλεπτική ικανότητα.Ιδιαίτερα σημαντικό είναι ότι το μοντέλο διατηρεί απλότητα, κάνοντας χρήση μόλις τριών από τις έξι διαθέσιμων μεταβλητών, γεγονός που το καθιστά εύχρηστο και ερμηνεύσιμο. Υπάρχει περιθώριο περαιτέρω διερεύνησης και βελτιστοποίησης. Μελλοντικές επεκτάσεις θα μπορούσαν να εξετάσουν την ενσωμάτωση πρόσθετων μεταβλητών, τη χρήση μη γραμμικών μοντέλων ή τεχνικών μηχανικής μάθησης, προκειμένου να ενισχυθεί η ακρίβεια και η γενικευσιμότητα του μοντέλου σε ευρύτερα δείγματα δεδομένων.

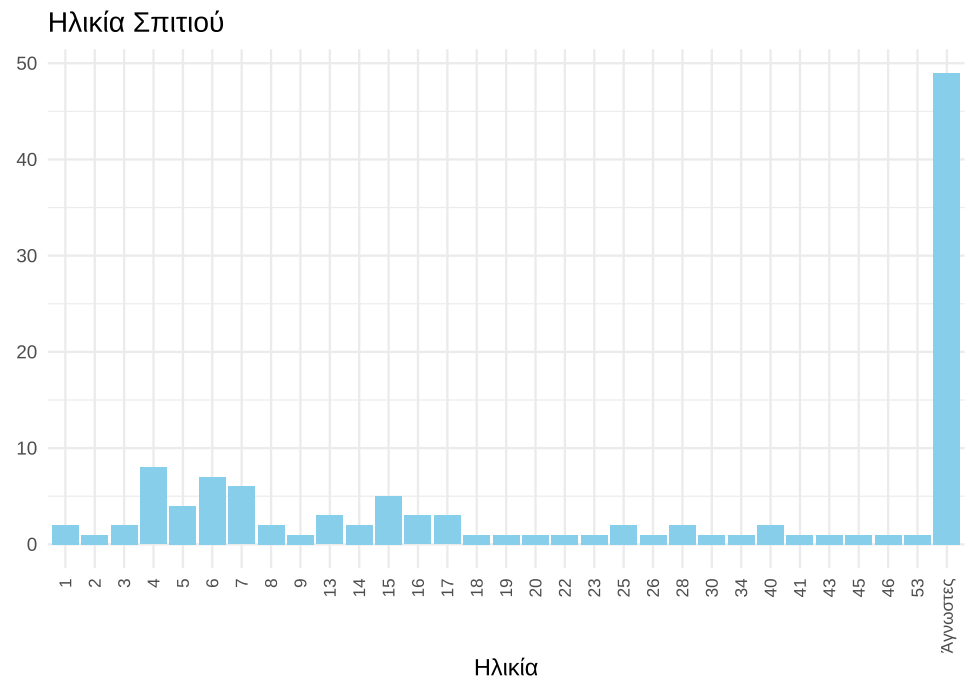
Ευρετήριο Πινάκων και διαγραμμάτων



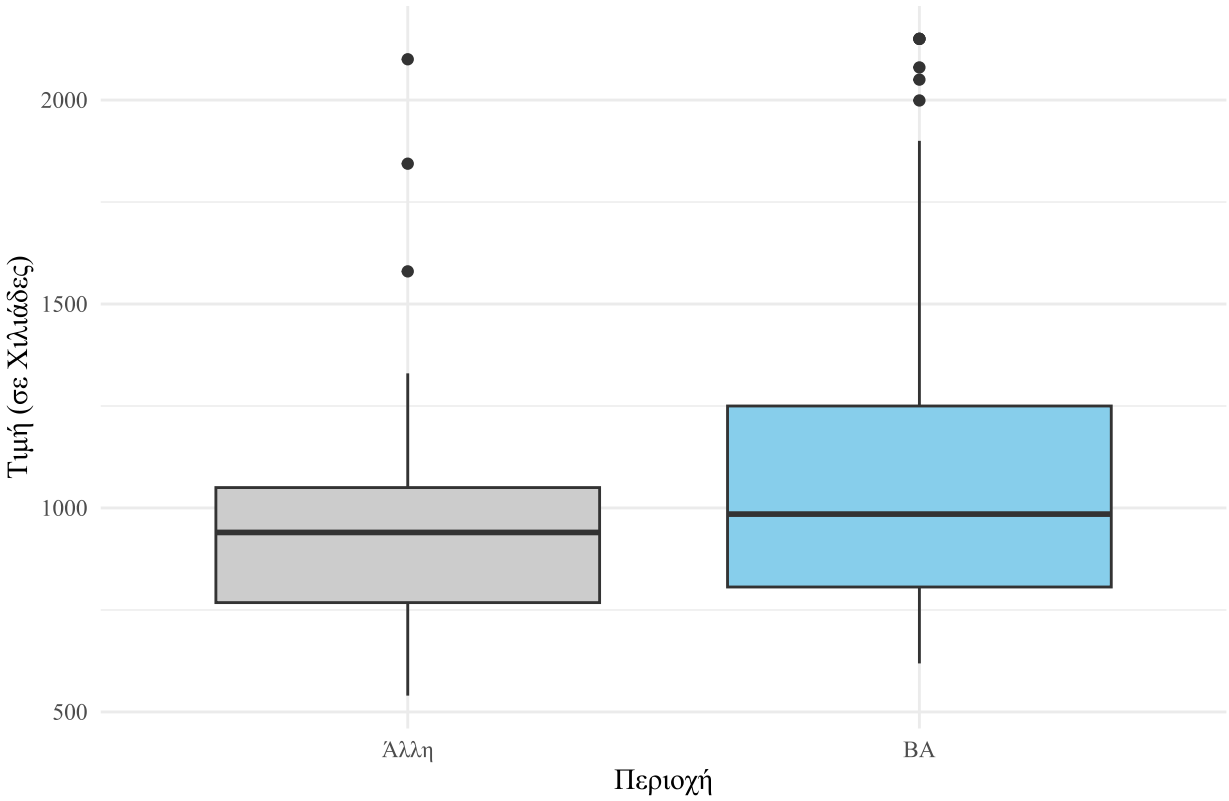
Σχήμα 1: Ιστογράμματα Ποσοτικών Μεταβλητών



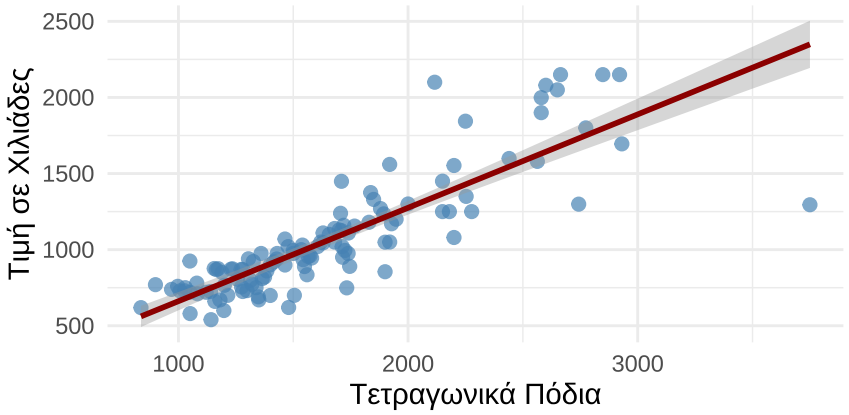
Σχήμα 2: Bar chart ποιοτικών μεταβλητών



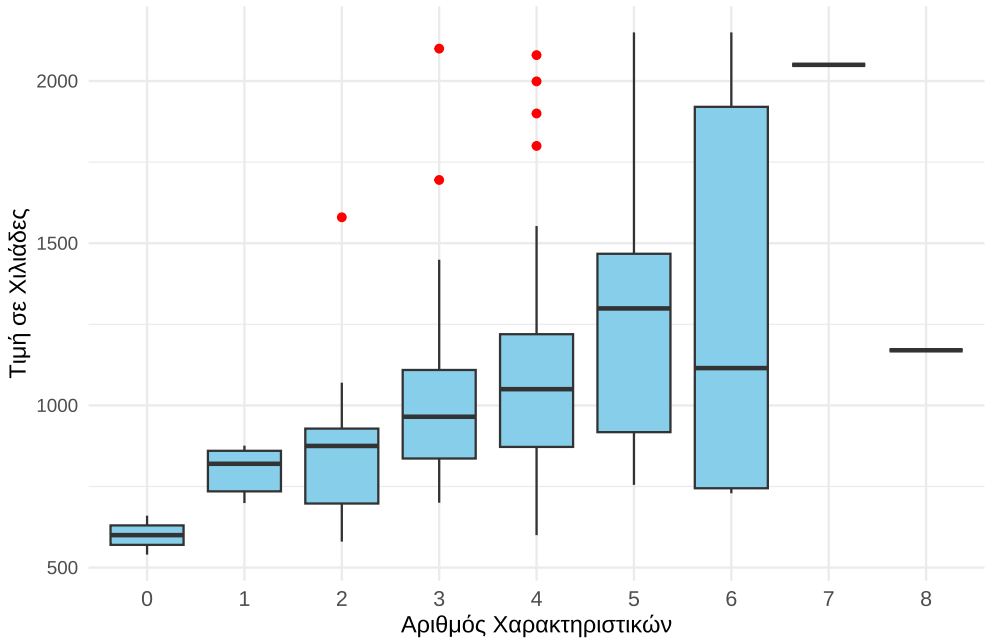
Σχήμα 3: Διάγραμμα Κατηγορίας Ηλικίας



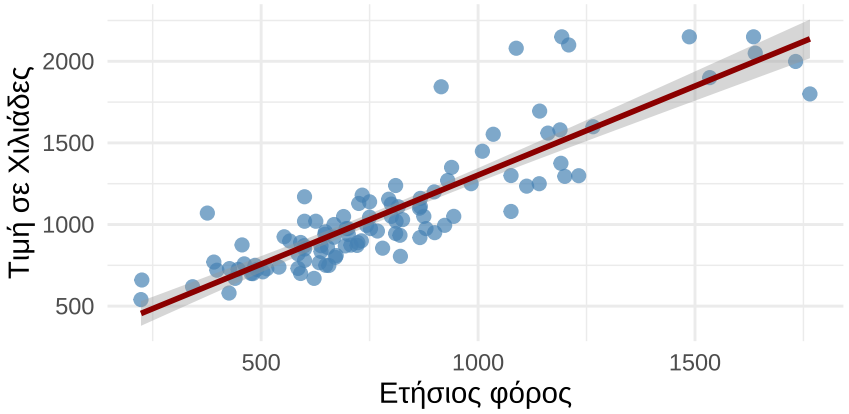
Σχήμα 4: Boxplot τιμή - τοποθεσία



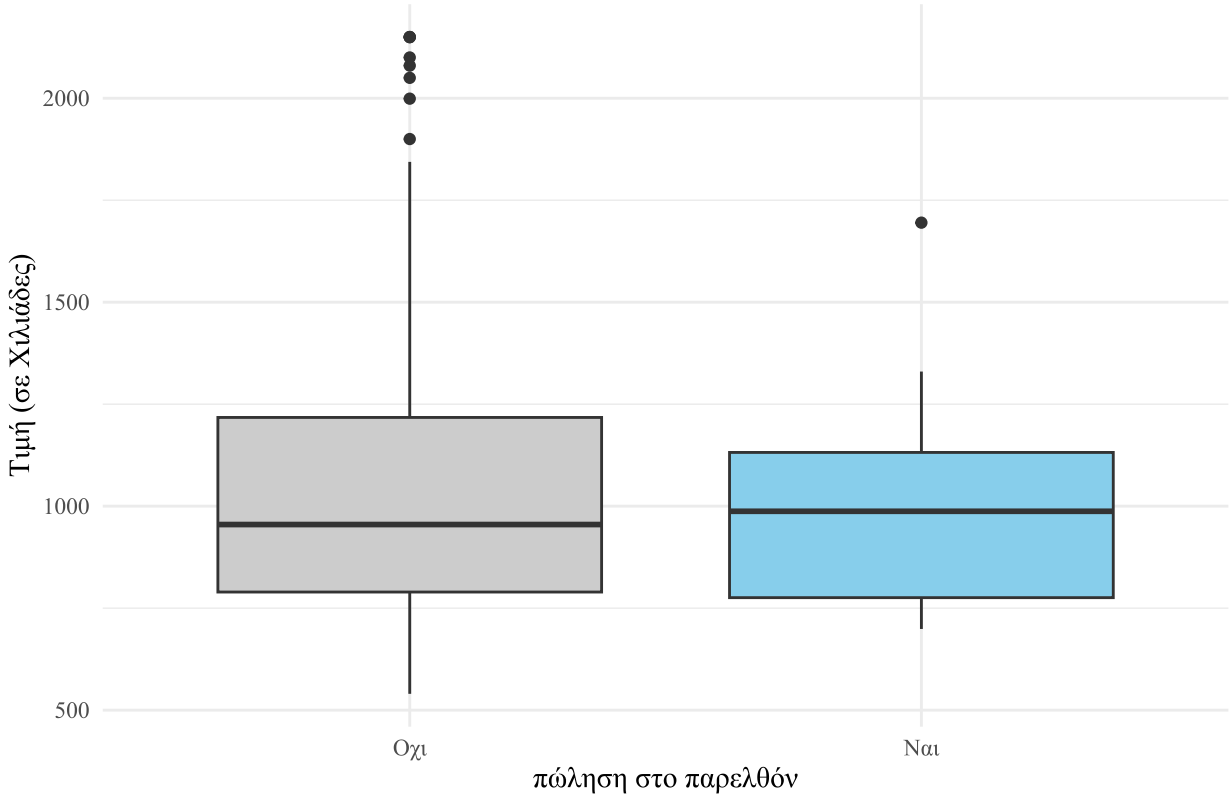
Σχήμα 5: Scatterplot τιμή - εσωτερικό μέγεθος σπιτιού



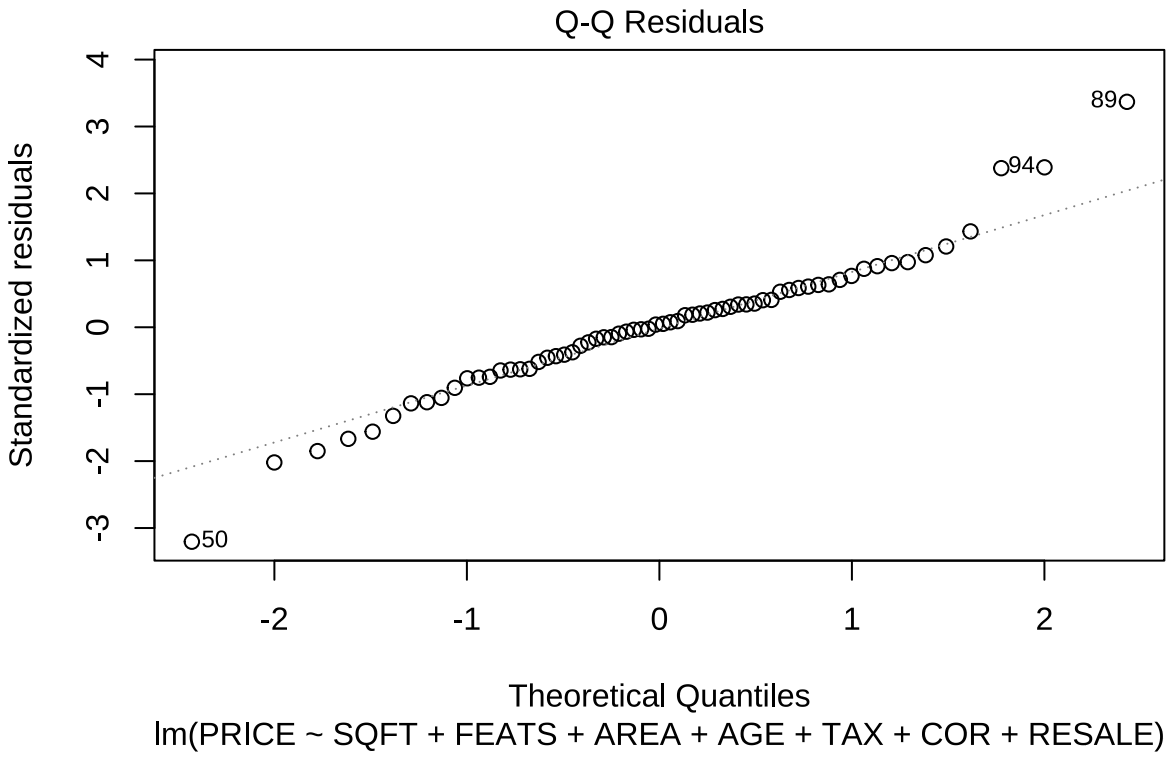
Σχήμα 6: Σχέση τιμής και αριθμού χαρακτηριστικών



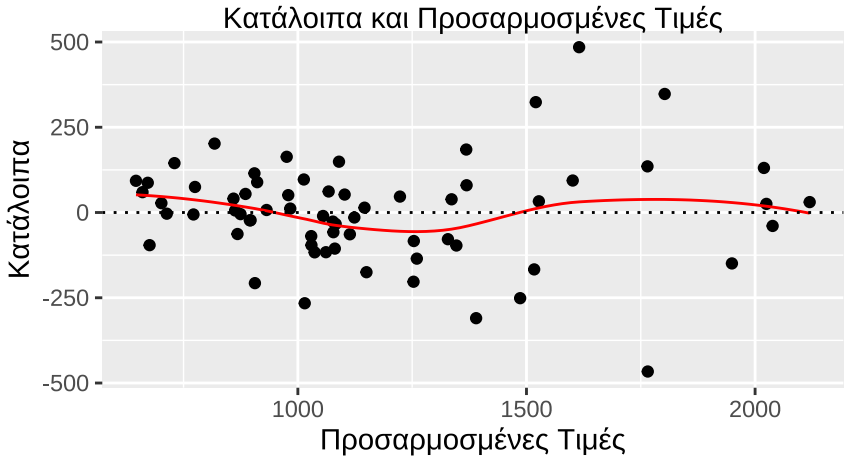
Σχήμα 7: Scatterplot ττιμή - φόρος



Σχήμα 8: Boxplot τιμή - πώληση στο παρελθόν



Σχήμα 9: Κανονικότητα κατάλοιπων



Σχήμα 10: Κατάλοιπα και Προσαρμοσμένες Τιμές