Ασκηση Bonus HY-111

CSD-4690

Σκοπος αυτης της ασκησης ειναι η υλοποιηση της γραμμικης παλινδρομησης με τον αλγοριθμο αποτομης καθοδου, σε δεδομενα που αφορουν την σχεση μεταξυ, το μεγεθος, την ηλικια και την τιμη ενοικιασης ενος διαμερισματος.

Γνωριζουμε τις συναρτησεις $f_1(x;\boldsymbol{\Theta}) = \Theta_0 + \theta_1 x \quad f_2(y;\boldsymbol{\Theta}) = \Theta_0 + \theta_1 y$ που εκφραζουν την τιμη ενοικιου σαν συναρτηση των τετραγωνικων και της παλαιοτητας αντιστοιχα.

Στοχος μας και για τις 2 αυτες συναρτησεις ειναι να ελαχιστοποιησουμε την συναρτηση κοστους, καθως με αυτον τον τροπο θα εντοπισουμε τις τιμες για τις μεταβλητες θ_0, θ_1 που θα εκφραζουν με τον καλυτερο δυνατο τροπο τα δεδομενα μας.

Η συναρτηση κοστους δινεται απο τον τυπο,

$$L(\boldsymbol{\Theta}) = \frac{1}{2K}\sum_{k=1}^{K}(f\left(x^{(i)};\boldsymbol{\theta}\right) - y^{(i)})^2.$$

Για να ελαχιστοποιησουμε την συναρτηση αυτη, θα υποθεσουμε αρχικα, τιμες για τις μεταβλητες θ_0, θ_1 που χρησιμοποιοντας τον παραπανω τυπο, θα μας δωσει μια συγκεκριμενη τιμη για το L.

Στη συνεχεια, θα παραγωγισουμε την συναρτηση L ως προς θ_0 και θ_1 ξεχωριστα οπου,

$$\frac{\partial L}{\partial \theta_0} = \frac{1}{K}\sum_{k=1}^{K}(f\left(x^{(\boldsymbol{i})};\boldsymbol{\theta}\right) - y^{(i)}),$$

$$\frac{\partial L}{\partial \theta_1} = \frac{1}{K}\sum_{k=1}^{K}(x(f\left(x^{(\boldsymbol{i})};\boldsymbol{\theta}\right) - y^{(i)})),$$

Και θα αφαιρεσουμε απο τα θ_0, θ_1 τις εξης ποσοτητες,

$$\theta_0' = \theta_0 - a\frac{\partial L}{\partial \theta_0},$$

$$\theta_1' = \theta_1 - a\,\frac{\partial L}{\partial \theta_1},$$

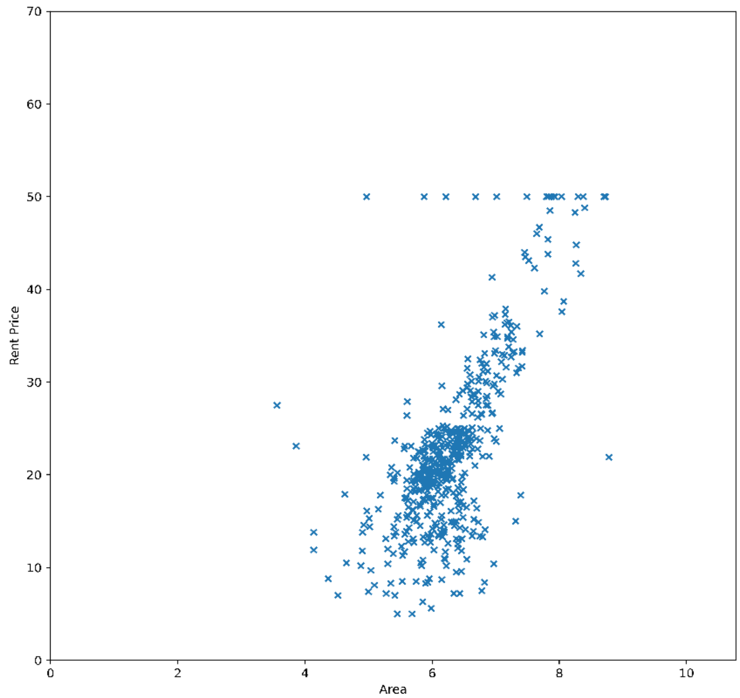
Οπου a το βημα που θα αποφασισουμε εμεις.

Την διαδικασια αυτη, την επαναλαμβανουμε οσες φορες κρινουμε οι ιδιοι οτι ειναι ιδανικο. Αυτο μπορει να ειναι ενας συγκεκριμενος αριθμος επαναληψεων ή μπορει να ειναι οταν τα θ_0, θ_1 τεινουν να συγκλινουν σε συγκεκριμενες τιμες.

Για την υλοποιηση του αλγοριθμου χρησιμοποιηθηκε η γλωσσα “Python” καθως και οι βιβλιοθηκες “NumPy” για διευκολυνση στις πραξεις και η “Matplotlib” για τον σχεδιασμο των γραφικων παραστασεων.

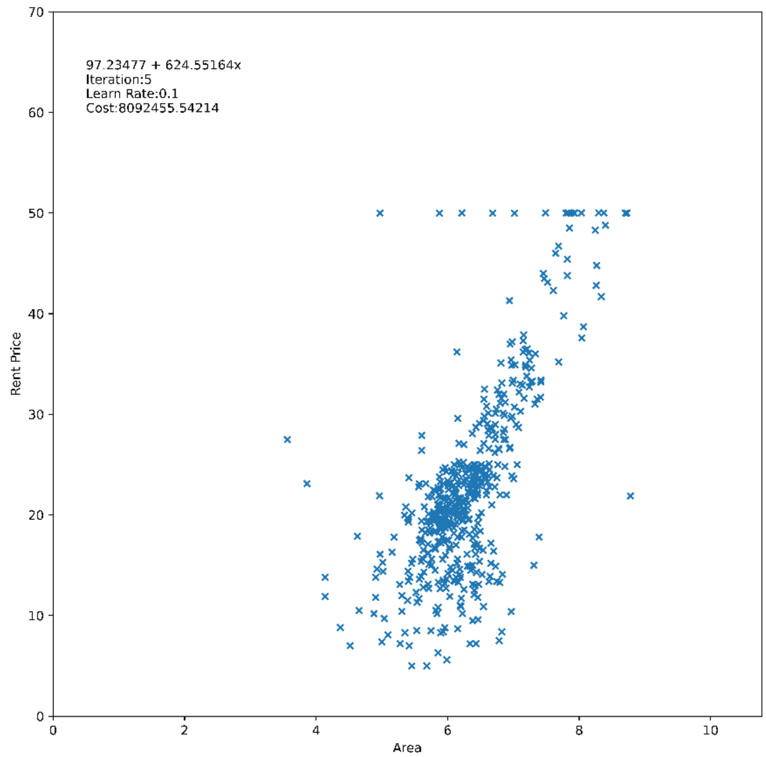
Αρχικα, θα εξετασουμε την συναρτηση f_1 και θα προσπαθησουμε να βρουμε τα καλυτερα θ_0, θ_1 που την περιγραφουν.

Στο παρακατω γραφημα βλεπουμε την σχεση μεταξυ του μεγεθους ενος σπιτιου ως προς την τιμη ενοικιασης του.



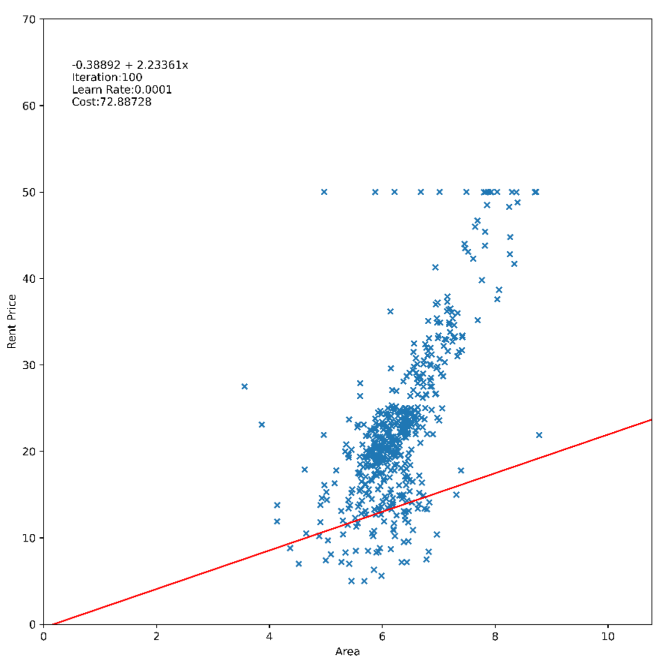
Ξεκινώντας με αρχικες συνθηκες θ_0 = -0.5 , $\theta_1 = 1.5$ και αφηνοντας τον αλγοριθμο να τρεξει 100 φορες για διαφορα a παρατηρουμε τα εξης αποτελεσματα.

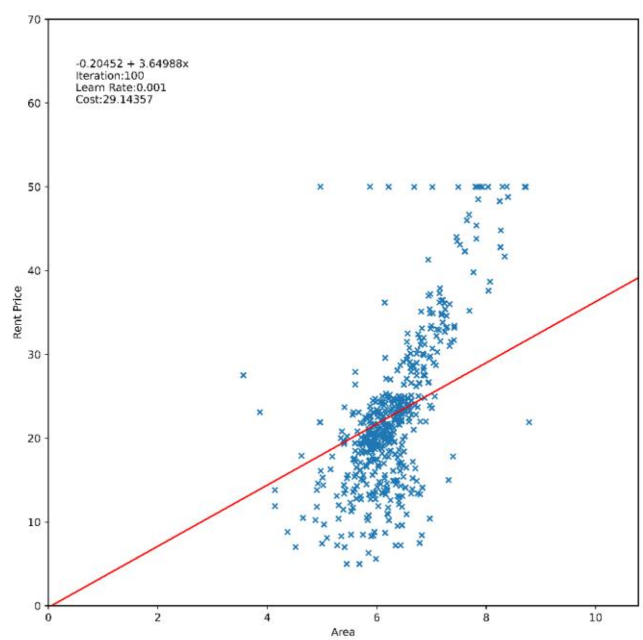
Για a = 0.1, βλεπουμε πως η συναρτηση απο την $5^{η}$ επαναληψη ηδη, εχει ξεφυγει τελειως απο τα δεδομενα μας.

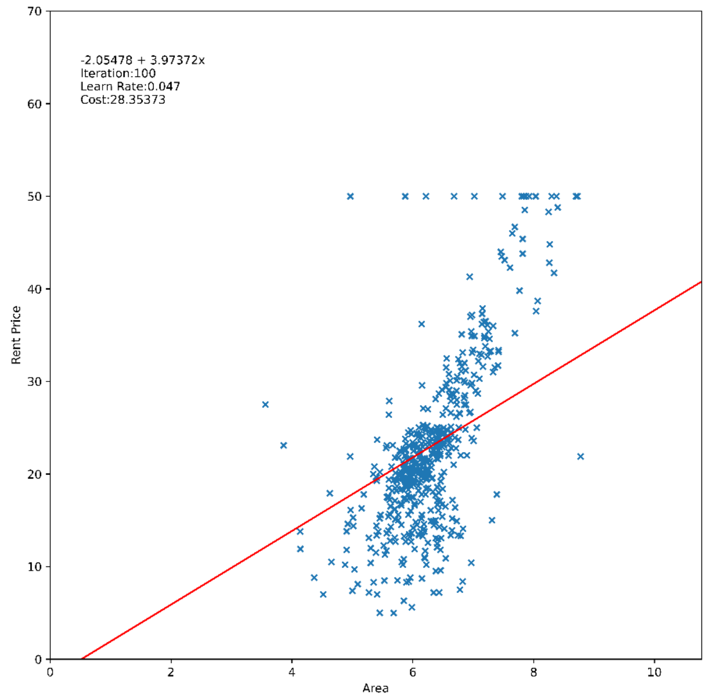


Η μεγαλυτερη επιτρεπτη τιμη για το a που παρατηρηθηκε, ηταν το 0.047.

Συγκρινοντας το αποτελεσμα για a = 0.047, a = 0.001, a = 0.0001, παιρνουμε 3 αρκετα διαφορετικες συναρτησεις ακομα για αυτο το μικρο πληθος επαναληψεων, οπως φαινεται και στα παρακατω γραφηματα.

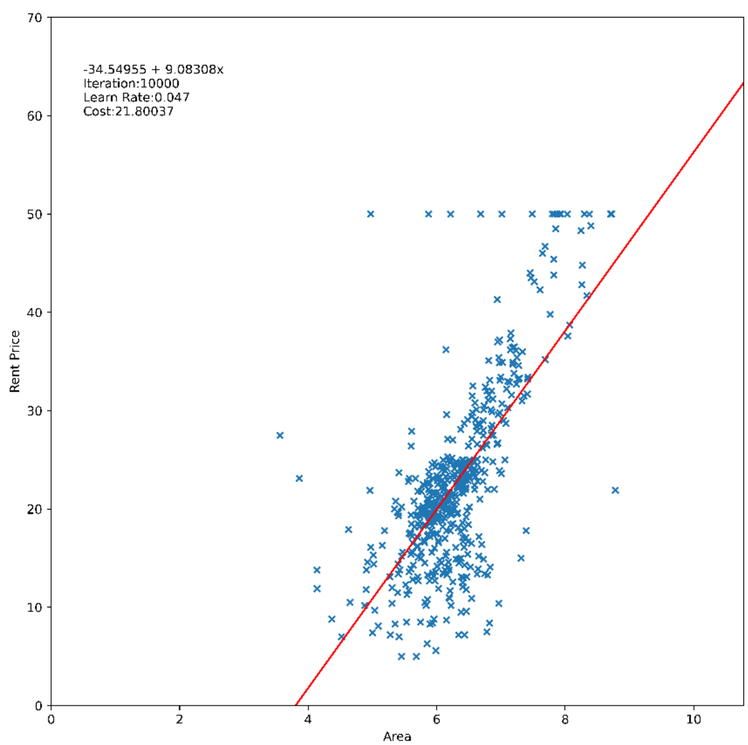




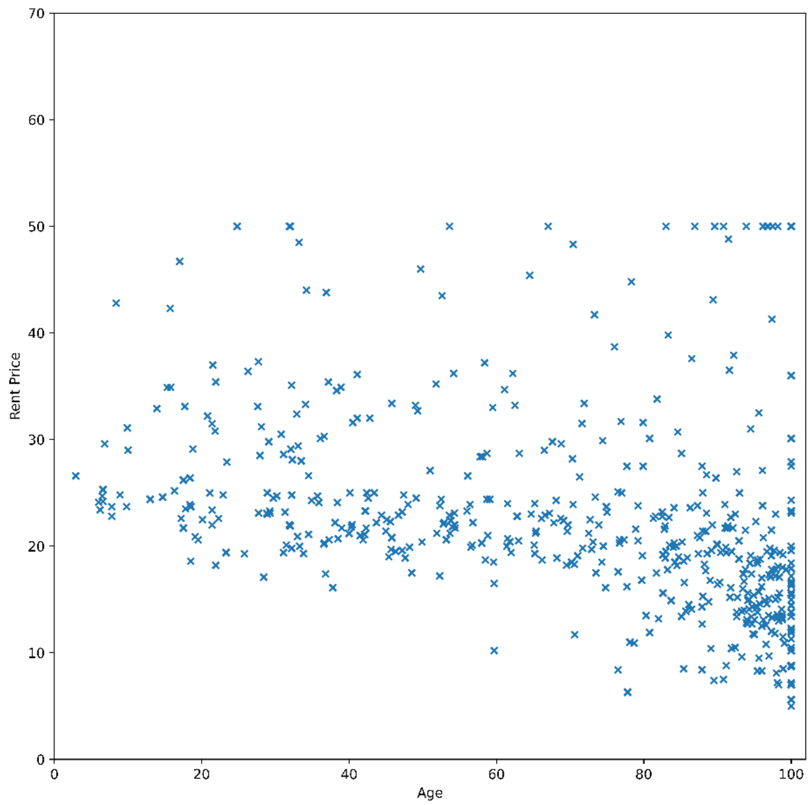


α = 0.047

Η ‘καλυτερη΄ προσεγγιση παρατηρηθηκε μετα απο περιπου 10000 επαναληψεις για το ιδιο a = 0.047.

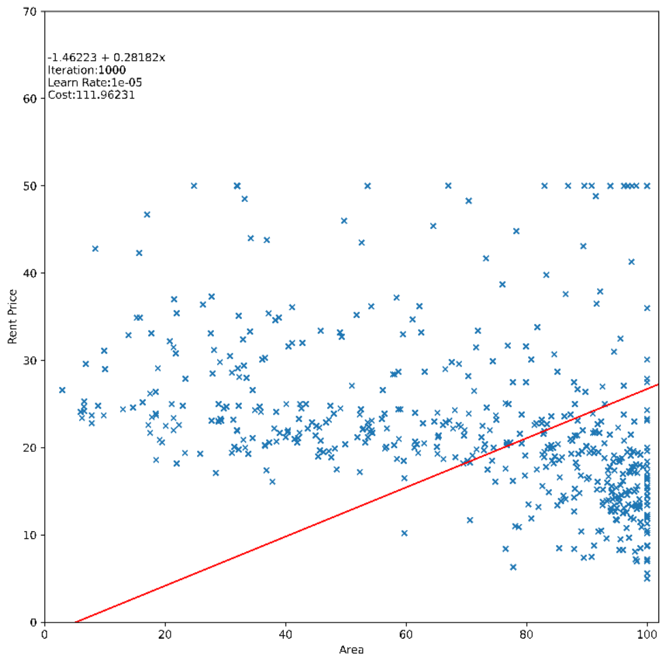


Στη συνεχεια θα αναλυσουμε την f_2, δηλαδη, την τιμη του ενοικιου σαν συναρτηση της παλαιοτητας. Στην παρακατω γραφικη παρασταση βλεπουμε ακριβως αυτην την σχεση.

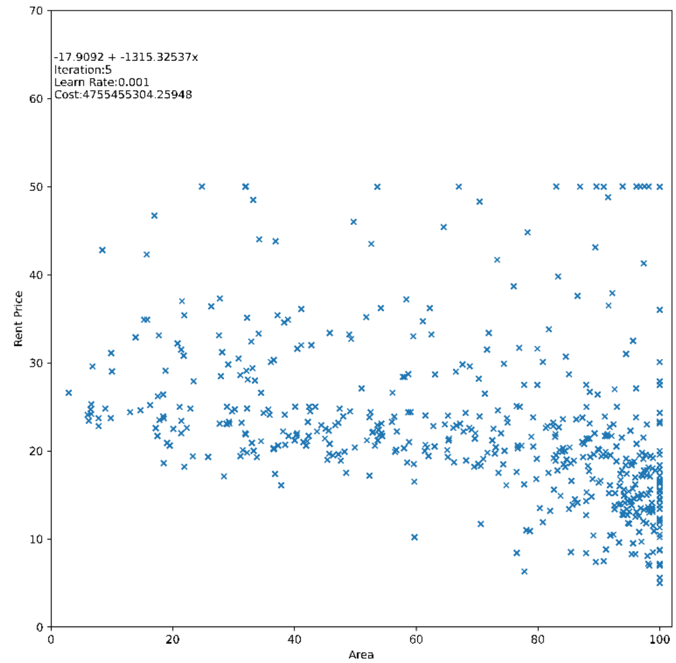


Για αρχικες συνθηκες θ_0 = -1.5 , $\theta_1 = 1$ και με αρχικο βημα a = 0.00001, μετα απο 1000 επαναληψεις, παρατηρουμε οτι οι μεταβλητες μας εχουν αλλαξει τιμες, αλλα οχι σε ικανοποιητικο βαθμο.

Στο αλλο ακρο, αν θεσουμε a = 0.001, μετα απο μονο 5 επαναληψεις παρατηρουμε οτι το βημα ειναι υπερβολικα μεγαλο και ξεφευγει τελειως απο τα δεδομενα μας.

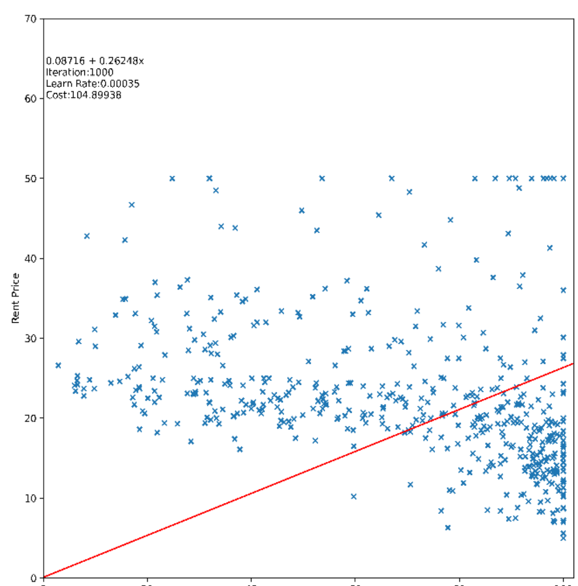


a = 0.00001

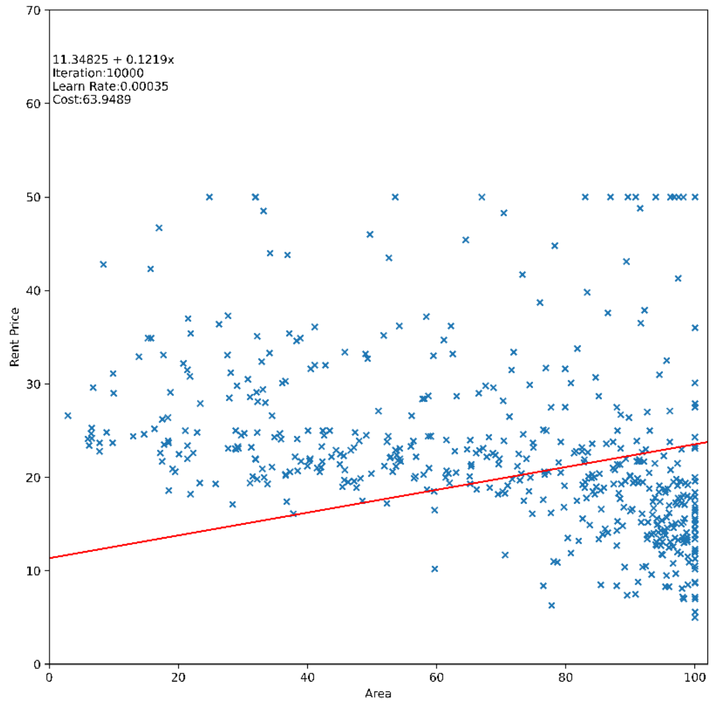


a = 0.001

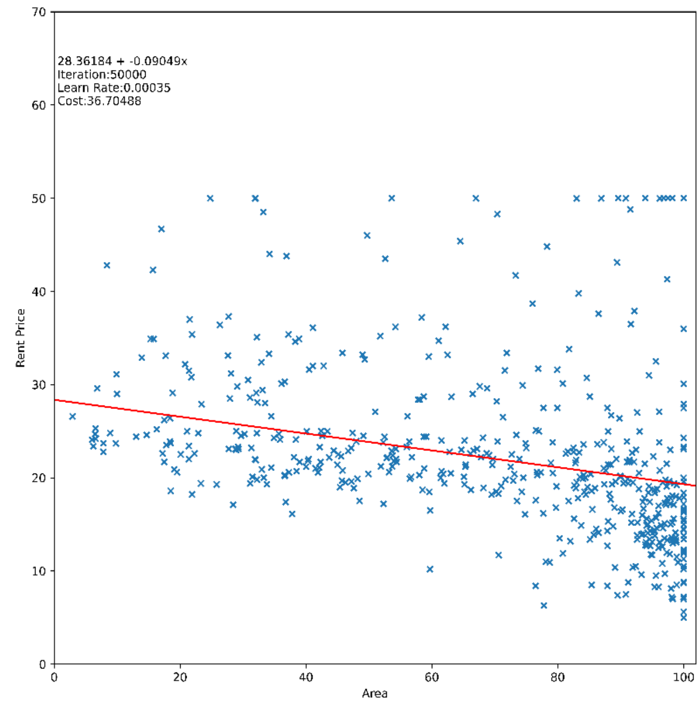
Η ιδανικη τιμη για την μεταβλητη a που εντοπιστηκε, ηταν το a = 0.00035. Για αυτο το a ο αλγοριθμος ετρεξε 1000, 10000 και 50000 επαναληψεις. Παρατηρουμε, πως λογο του μικροτερου βηματος, για να σταθεροποιηθει σε ενα τοπικο ελαχιστο, ο αλγοριθμος επρεπε να τρεξει περισσοτερες επαναληψεις απο την f_1.



Iteration = 1000



Iteration = 10000



Iteration = 50000

Τωρα θα εξετασουμε την τιμη ενοικιου ως συναρτηση των τετραγωνικων και ηλικιας μαζι, την συναρτηση f_3 δηλαδη. Το μονο πραγμα που αλλαζει στον αλγοριθμο απο την απλη περιπτωση της f_1, f_2, ειναι οτι θα πρεπει πλεον να παραγωγιζουμε 3 μεταβλητες, τις $\theta_0, \theta_1, \theta_2$.

Η συναρτηση f_3, δινεται απο τον τυπο:

$$f_3(x; y; \boldsymbol{\Theta}) = \Theta_0 + \theta_1 x + \theta_2 y,$$

Ενω οι μεταβλητες θ που πρεπει να παραγωγισουμε ειναι οι:

$$\frac{\partial L}{\partial \theta_0} = \frac{1}{K} \sum_{k=1}^{K} (f(x^{(\boldsymbol{i})}; y^{(i)}; \boldsymbol{\theta}) - z^{(i)}),$$

$$\frac{\partial L}{\partial \theta_1} = \frac{1}{K} \sum_{k=1}^{K} (x(f(x^{(\boldsymbol{i})}; y^{(i)}; \boldsymbol{\theta}) - z^{(i)})),$$

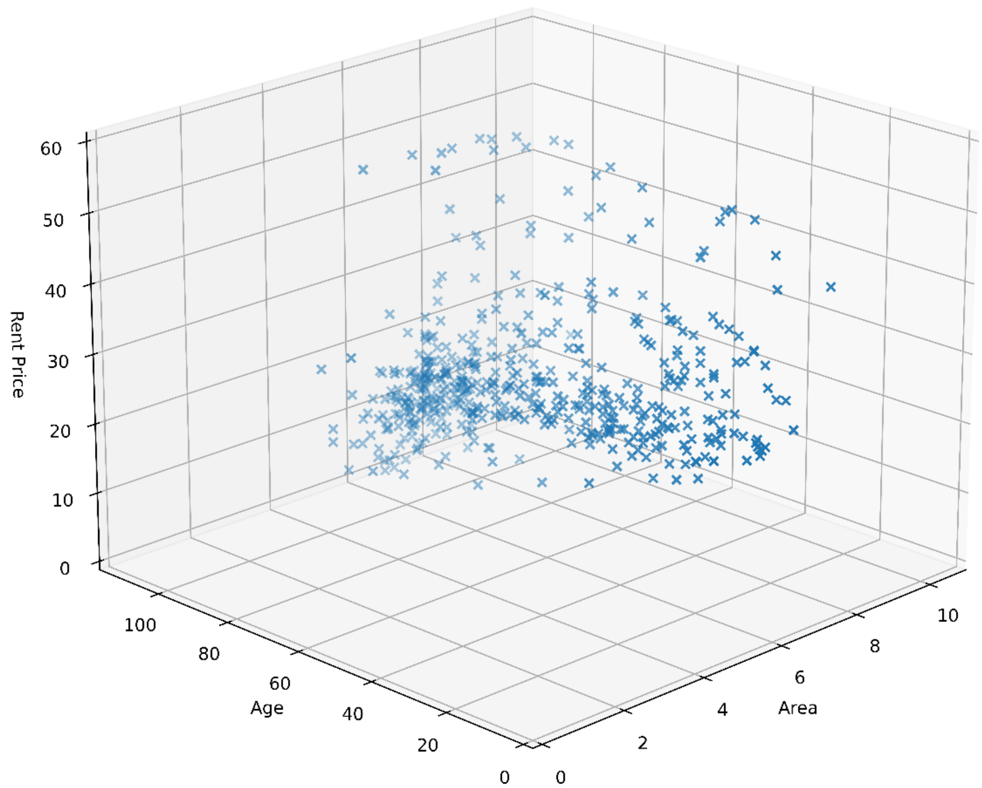
$$\frac{\partial L}{\partial \theta_2} = \frac{1}{K} \sum_{k=1}^{K} (y(f(x^{(\boldsymbol{i})}; y^{(i)}; \boldsymbol{\theta}) - z^{(i)})),$$

$$\theta_0' = \theta_0 - a \frac{\partial L}{\partial \theta_0},$$

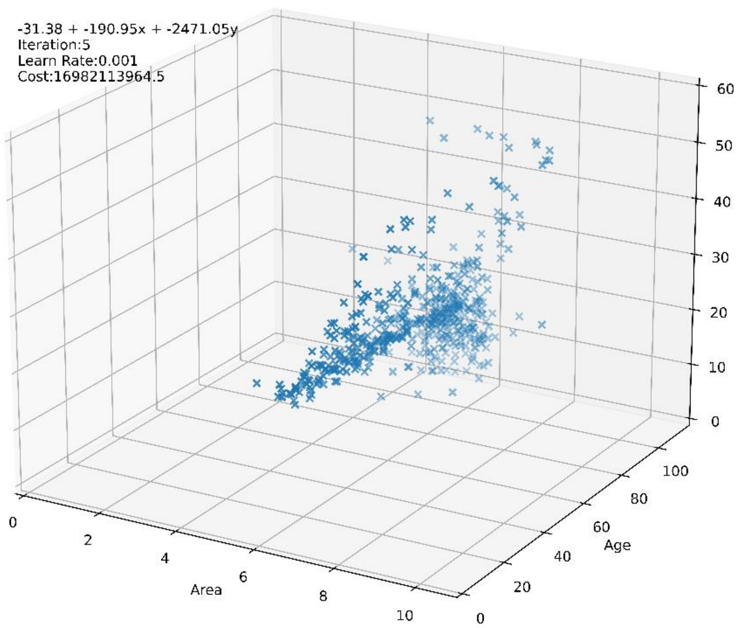
$$\theta_1' = \theta_1 - a \frac{\partial L}{\partial \theta_1},$$

$$\theta_2' = \theta_2 - a \frac{\partial L}{\partial \theta_2}.$$

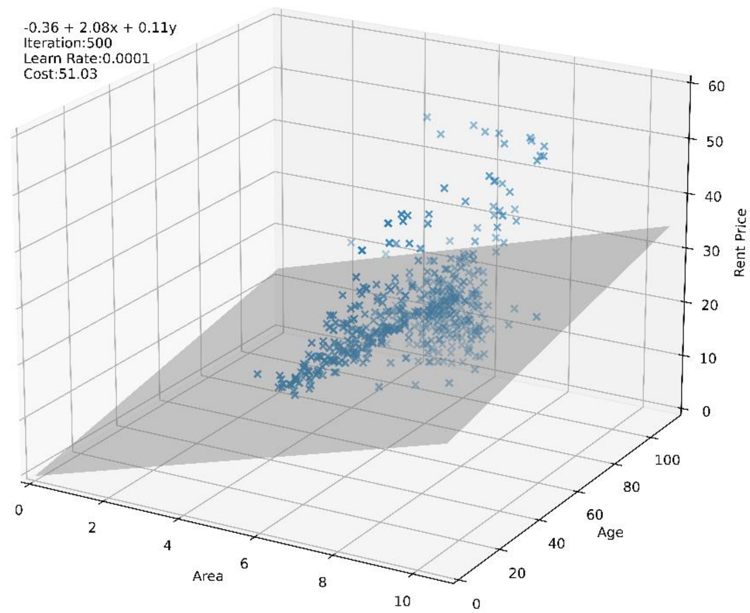
Στο παρακατω γραφημα, παρουσιαζεται το διαγραμμα διασπορας της f_3.



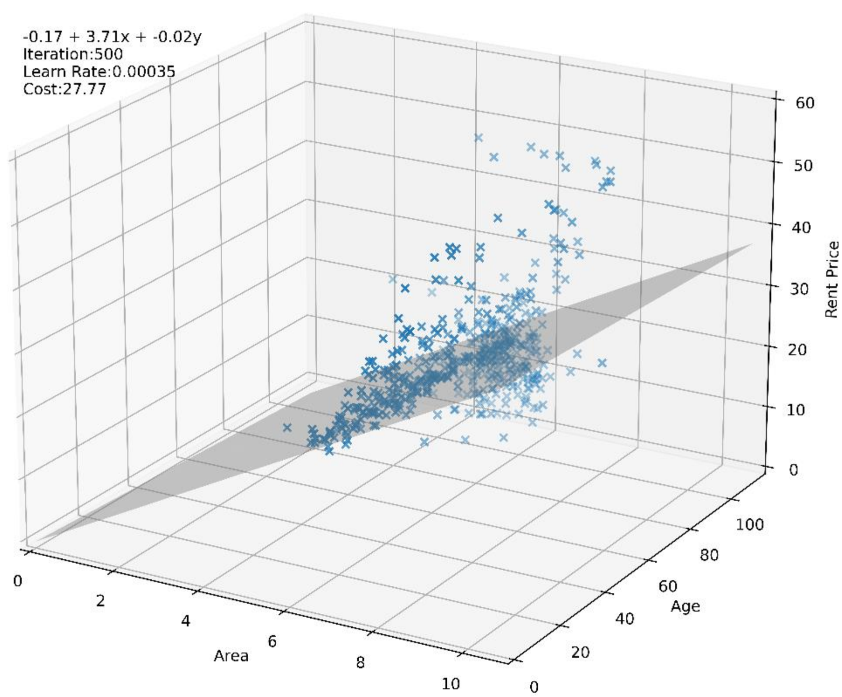
Ξεκινωντας με αρχικες συνθηκες, $\theta_0 = -.05$, $\theta_1 = 1$, $\theta_2 = 1.5$, και αφηνοντας τον αλγοριθμο να τρεξει για 500 επαναληψεις παρατηρουμε τις διαφορετικες τιμες των μεταβλητων μας για a = 0.001, a = 0.0001, a = 0.00035.



a = 0.001

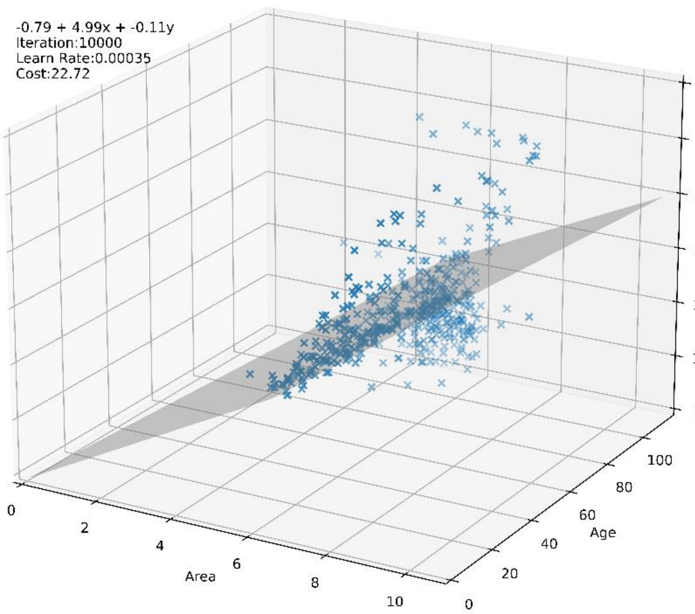


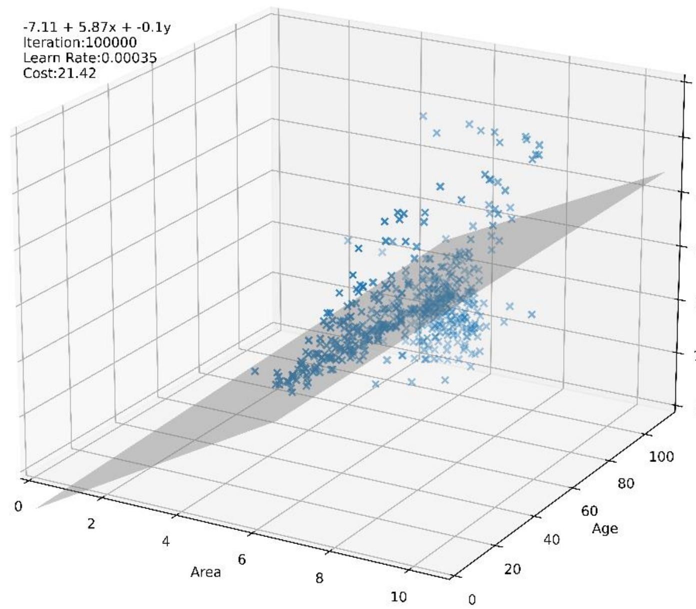
a =0.0001



a = 0.00035

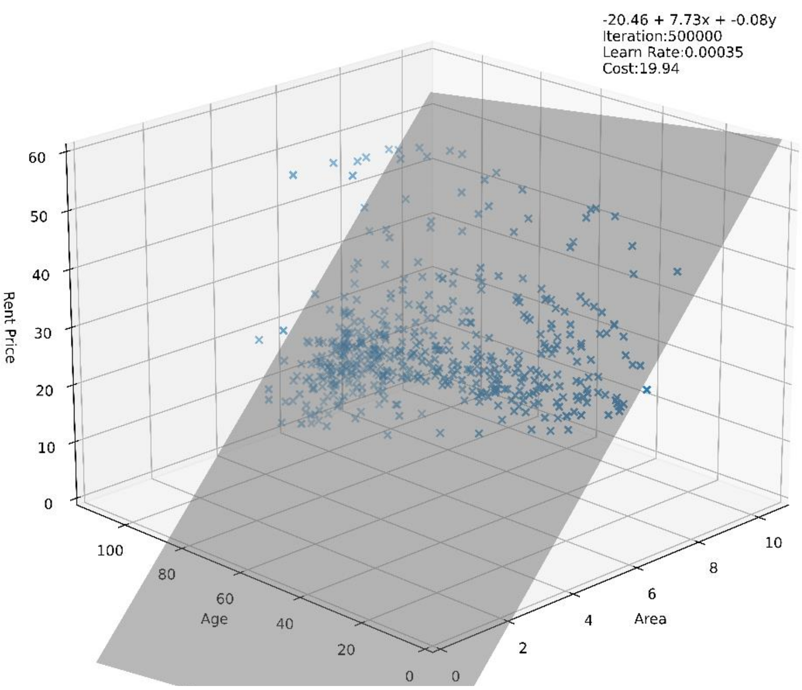
Για a = 0.00035, ο αλγοριθμος ετρεξε 10000, 100000, 500000 φορες, με σκοπο να βρεθει η καλυτερη δυνατη συναρτηση που εκφραζει τα δεδομενα μας.

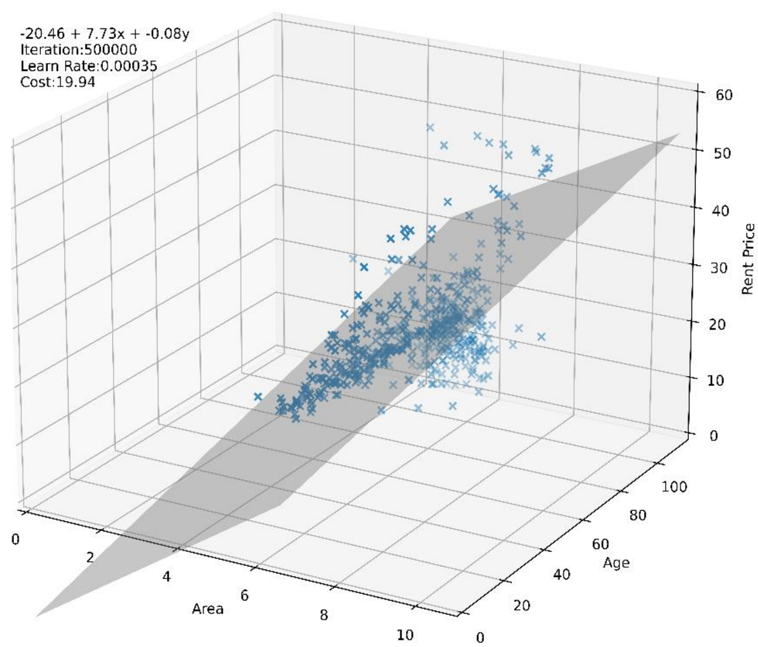




Iteration = 10000

Iteration = 100000





Iteration = 500000(L)

Iteration = 500000(R)

Πολυ ευκολα μπορουμε να διαπιστωσουμε, πλεον, πως οταν εξεταζουμε τις μεταβλητες εισοδου ξεχωριστα, δεν μπορουμε να εχουμε μια καθαρη εικονα για την συναρτηση συνολικα, ειδικα οταν οι μεταβλητες εισοδου ειναι παραπανω απο 2.

Απο την αλλη μερια, οταν εξεταζουμε μαζι ολες τις μεταβλητες μπορουμε να προσεγγισουμε το δεδομενα μας με πολυ μεγαλυτερη ακριβεια.

Σε αυτο το link ειναι αναρτημενος ενας φακελος με τις υλοποιησεις των αλγοριθμων για 2 και 3 μεταβλητες, καθως και τα screenshot της αναφορας στην κανονικη αναλυση.

Ευχαριστω για τον χρονο σας,

Νικολας Μπαρμπαρουσης, csd-4690.