Εφαρμοσμένη Στατιστική - Προγραμματιστική Άσκηση

Αλλοιμόνου Όλγα 2465

Βουγιαντζής Νικόλαος 2476

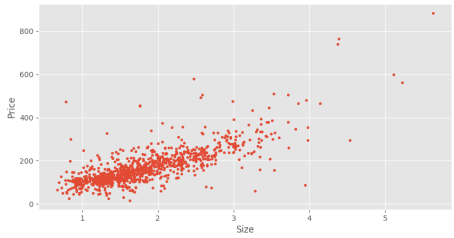
Ευαγγελινού Μαρία 2499

Εισαγωγή

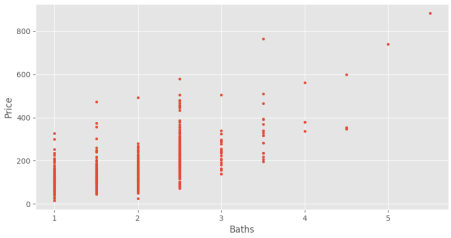
Βασιζόμενοι σε ένα δείγμα από 1063 σπίτια στο προάστιο Saratoga της Νέας Υόρκης, το οποίο περιλάμβανει το μέγεθος του σπιτιού(σε τετραγωνικά πόδια),τον αριθμό μπάνιων,τον αριθμό των υπνοδωματίων,αν το σπίτι έχει τζάκι,τον αριθμό στρεμμάτων του οικοπέδου και την ηλικία του κάθε σπιτιού , θα φτιάξουμε ένα γραμμικό μοντέλο με στόχο να δούμε πως οι ανεξάρτητες μεταβλητές που αναφέραμε παραπάνω σχετίζονται με την τιμή του κάθε σπιτιού.Στη συνέχεια αφού εντοπίσουμε τις σημαντικότερες μεταβλητές για την πρόβλεψη της τιμής των κατοικιών θα εφαρμόσουμε τον αλγόριθμο εξάλειψης προς τα πίσω για να βελτιώσουμε το αρχικό γραμμικό μοντέλο.

Άσκηση 1

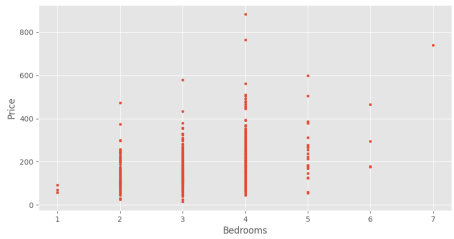
Χρησιμοποιώντας τις βιβλιοθήκες της python κατασκευάζουμε τα διαγράμματα διασποράς που φαίνονται στα Σχήματα 1 εως 6 και υπολογίζουμε τους συντελεστές συσχέτισης ανάνεσα στην εξαρτημένη μεταβλητή Price και κάθε μια από τις ανεξάρτητες μεταβλητές.Κάποια γραμμική συσχέτιση μπορεί να παρατηρηθεί στα πρώτα τρία γραφήματα ενώ στα υπόλοιπα δεν φαίνεται να υπάρχει κάποια.



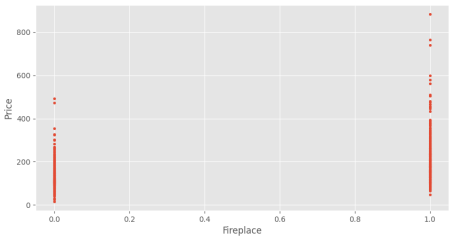
Σχήμα 1



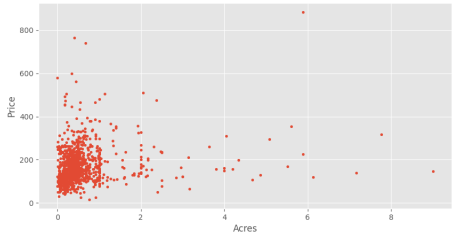
Σχήμα 2



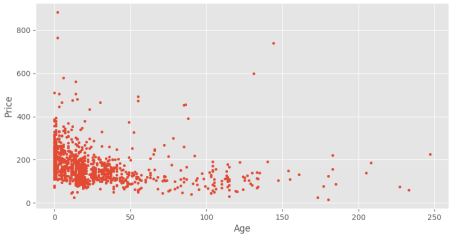
Σχήμα 3



Σχήμα 4



Σχήμα 5



Σχήμα 6

Οι συντελεστές συσχέτισης δίνονται στον παρακάτω πίνακα:

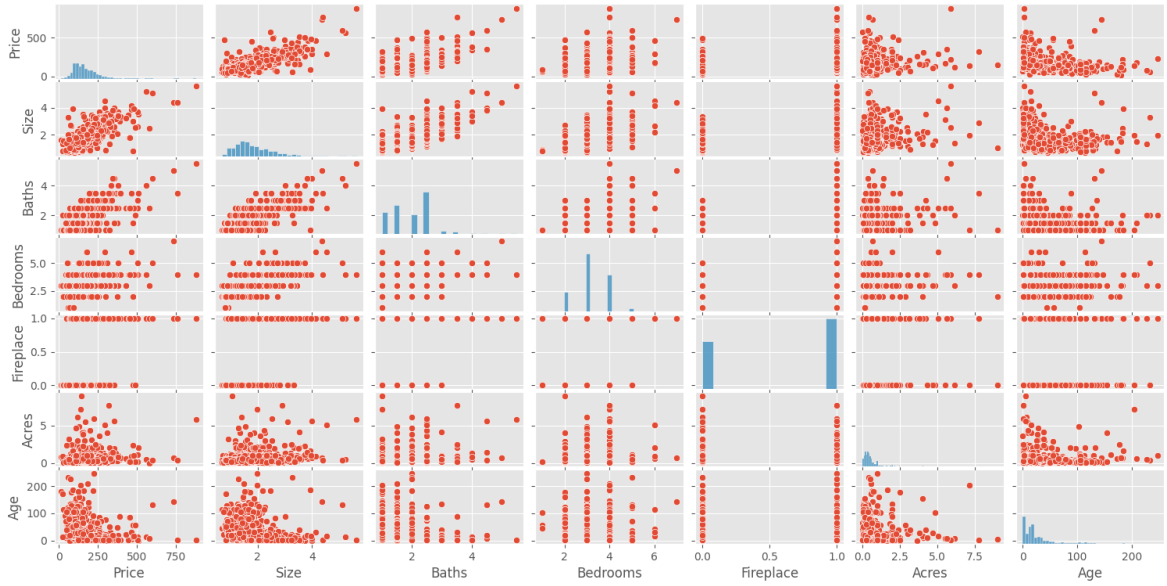
r	Price
Size	0.7710771189077961
Baths	0.6695343237053092
Bedrooms	0.4700213813216685
Fireplace	0.40947927384618976
Acres	0.17958146956009655
Age	-0.26140969022619415

Σύμφωνα με τα δεδομένα του πίνακα αν έπρεπε να περιοριστούμε σε μια μόνο μεταβλητή για την πρόβλεψη της τιμής των κατοικιών αυτή θα ήταν η Size διότι ο συντελεστής συσχέτισης ανάμεσα σε αυτή και την Price, κατά απόλυτη τιμή, είναι πιο κοντά στο 1 σε σχέση με τους υπόλοιπους συντελεστές συσχέτισης.Η εξάρτηση της μεταβλητής Price από την Size είναι ισχυρή.

Άσκηση 2

(α) Θα ελέγξουμε τις προυποθέσεις που πρέπει να τηρούνται για το μοντέλο πολλαπλής γραμμικής παλινδρόμησης.

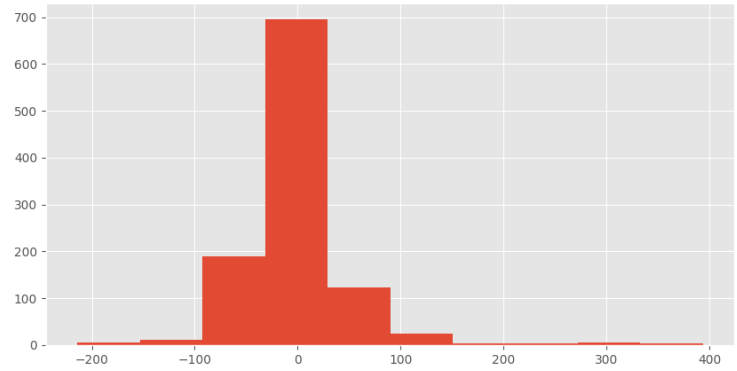
Αρχικά θα πρέπει κάθε ανεξάρτητη μεταβλητή να έχει γραμμική σχέση με την εξαρτημένη (γραμμικότητα),αυτό μπορούμε να το διαπιστώσουμε μέχω των διαγραμμάτων διασποράς



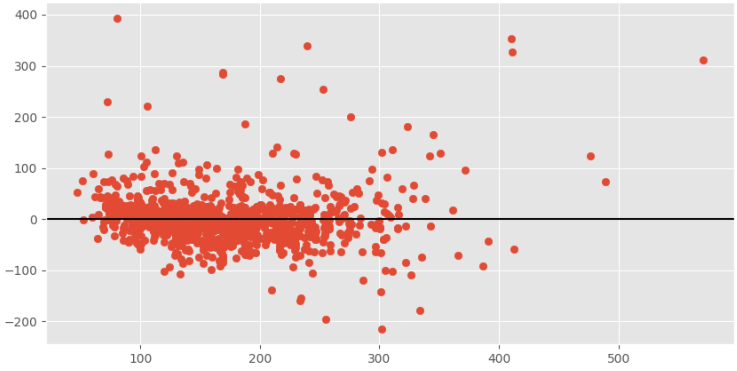
Αυτή η συνθήκη δεν φαίνεται να διατηρείται, κάποια γραμμικότητα διαφαίνεται μόνο ανάμεσα στη μεταβλητή Price και τις Size, Baths, Bedrooms.

Επίσης δεν ικανοποείται ούτε η συνθήκη της απουσίας πολυσυγγραμικότητας καθώς απο τα διαγράματα διασποράς παρατηρούμε γραμμική εξάρτηση ανάμεσα στις Size, Baths και Bedrooms.

Αν θέσουμε y τις πραγματικές τιμές Price που προέρχονται από το δείγμα και $\hat{y}$ τις εκτιμώμενες τιμές της ίδιας μεταβλητής, τότε μια ακόμα προυπόθεση είναι τα σφάλματα $y - \hat{y}$ να ακολουθούν κανονική κατανομή, αυτό μπορούμε το δούμε σχεδιάζοντας το ραβδόγραμμα το οποίο μας δείχνει οτι η κατανομή όντως μοιάζει να είναι κανονική.



Τέλος ελέγχουμε την ομοσκεδαστικότητα των σφαλμάτων με το παρακάτω διάγραμμα διασποράς και βλέπουμε ότι ο συντελεστής συσχέτισης είναι πολυ μικρός και τα σφάλματα έχουν σταθερή διασπορά σε όλο το μήκος του γραφήματος.



(β+γ)

Η βιβλιοθήκη statsmodels μας δίνει τον πίνακα για το μοντέλο

```
                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.619
Dependent Variable: Price            AIC:                11430.0966
Date:               2022-06-27 12:06 BIC:                11464.8785
No. Observations:   1063             Log-Likelihood:     -5708.0
Df Model:           6                F-statistic:        288.5
Df Residuals:       1056             Prob (F-statistic): 1.66e-218
R-squared:          0.621            Scale:              2719.6
------------------------------------------------------------------
                Coef.   Std.Err.     t     P>|t|    [0.025    0.975]
------------------------------------------------------------------
const           0.9498    7.5624   0.1256  0.9001  -13.8893  15.7889
Size           80.0729    4.1696  19.2038  0.0000   71.8912  88.2546
Baths          25.2787    3.8701   6.5318  0.0000   17.6847  32.8727
Bedrooms       -9.0142    2.8523  -3.1603  0.0016  -14.6111  -3.4173
Fireplace       4.8438    3.7762   1.2827  0.1999   -2.5658  12.2535
Acres           1.8874    2.0798   0.9075  0.3644   -2.1936   5.9684
Age            -0.0637    0.0512  -1.2439  0.2138   -0.1642   0.0368
------------------------------------------------------------------
Omnibus:              580.365       Durbin-Watson:          1.584
Prob(Omnibus):        0.000         Jarque-Bera (JB):       8151.571
Skew:                 2.191         Prob(JB):               0.000
Kurtosis:             15.839        Condition No.:          226
==================================================================
```

το οποίο τελικά είναι :

$$\hat{y} = 0.9498 + 80.0729 \cdot \text{Size} + 25.2787 \cdot \text{Baths} - 9.0142 \cdot \text{Bedrooms} + 4.8438 \cdot \text{Fireplace} + 1.8874 \cdot \text{Acres} - 0.0637 \cdot \text{Age}$$

Στην πρώτη άσκηση είδαμε ότι η μεταβλητή με την μεγαλύτερη σημαντικότητα για την πρόβλεψη της τιμής ήταν η Size.

Ο συντελεστής της Size στο μοντέλο είναι 80.0729 ,αυτό μας δείχνει ότι αν η Size αυξηθεί κατα μία μονάδα και όλες οι άλλες παράμετροι παραμείνουν σταθερές τότε η $\hat{y}$ θα αυξηθεί κατα 80.0729 μονάδες.

Επιπλέον ο συντελεστής της Size κατά απόλυτη τιμή διαφέρει αρκετά από οποιονδήποτε άλλο και συνεπώς επηρεάζει σε μεγαλύτερο βαθμό τις τιμές της $\hat{y}$.Με την ίδια λογική οι αμέσως πιο σημαντικές μεταβλητές για το μοντέλο μας θα είναι οι Baths και Bedrooms και αυτό επιβεβαιώνεται από τις τιμές των p-values αυτών οι οποίες είναι όλες μικρότερες του 0.01.

Άσκηση 3

Αρχικά θεωρούμε το πλήρες μοντέλο και σε αυτό εφαρμόζουμε τον αλγόριθμο της εξάλειψης προς τα πίσω (backward elimination).

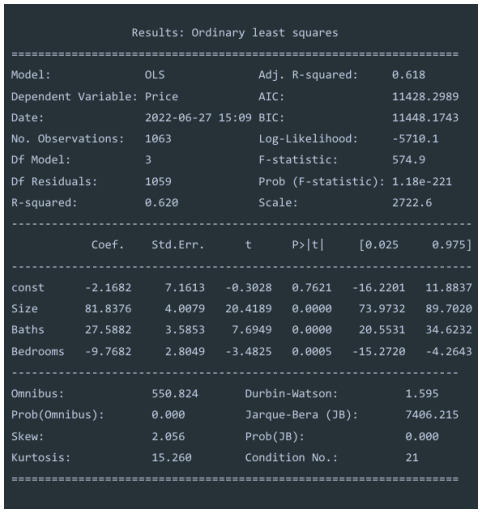
Αν πάρουμε α=0.01 το μεγαλύτερο P-value το έχει η μεταβλητή Acres , οπότε και την αφαιρούμε.

```
                Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.619
Dependent Variable: Price            AIC:                11428.9252
Date:               2022-06-27 15:07 BIC:                11458.7383
No. Observations:   1063             Log-Likelihood:     -5708.5
Df Model:           5                F-statistic:        346.0
Df Residuals:       1057             Prob (F-statistic): 1.28e-219
R-squared:          0.621            Scale:              2719.1
------------------------------------------------------------------
               Coef.   Std.Err.     t     P>|t|    [0.025    0.975]
------------------------------------------------------------------
const           1.1650    7.5581   0.1541  0.8775  -13.6656  15.9955
Size           80.7267    4.1066  19.6579  0.0000   72.6688  88.7847
Baths          25.1755    3.8681   6.5084  0.0000   17.5855  32.7656
Bedrooms       -9.0440    2.8519  -3.1712  0.0016  -14.6401  -3.4480
Fireplace       4.6969    3.7724   1.2451  0.2134   -2.7053  12.0991
Age            -0.0618    0.0512  -1.2075  0.2275   -0.1622   0.0386
------------------------------------------------------------------
Omnibus:             580.909       Durbin-Watson:          1.590
Prob(Omnibus):       0.000         Jarque-Bera (JB):       8132.656
Skew:                2.195         Prob(JB):               0.000
Kurtosis:            15.819        Condition No.:          226
==================================================================
```

Επαναλαμβάνουμε τη διαδικασία και παρατηρούμε πως το αμέσως μεγάλυτερο p-value το έχει η Age την οποία επίσης αφαιρούμε.

```
                Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.619
Dependent Variable: Price            AIC:                11428.3905
Date:               2022-06-27 15:08 BIC:                11453.2347
No. Observations:   1063             Log-Likelihood:     -5709.2
Df Model:           4                F-statistic:        432.0
Df Residuals:       1058             Prob (F-statistic): 1.19e-220
R-squared:          0.620            Scale:              2720.3
------------------------------------------------------------------
               Coef.   Std.Err.     t     P>|t|    [0.025    0.975]
------------------------------------------------------------------
const          -1.7461    7.1648  -0.2437  0.8075  -15.8049  12.3127
Size           80.5968    4.1061  19.6288  0.0000   72.5399  88.6538
Baths          26.7915    3.6300   7.3805  0.0000   19.6686  33.9144
Bedrooms       -9.6727    2.8046  -3.4489  0.0006  -15.1759  -4.1695
Fireplace       5.1739    3.7525   1.3788  0.1682   -2.1892  12.5370
------------------------------------------------------------------
Omnibus:             562.105       Durbin-Watson:          1.596
Prob(Omnibus):       0.000         Jarque-Bera (JB):       7698.772
Skew:                2.107         Prob(JB):               0.000
Kurtosis:            15.493        Condition No.:          21
==================================================================
```

Τέλος, εφαρμόζοντας τον αλγόριθμο ακόμα μία φορά βλέπουμε πως η Fireplace έχει p-value μεγαλύτερο του 0.01 άρα την αφαιρούμε.



```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.618
Dependent Variable: Price            AIC:                11428.2989
Date:               2022-06-27 15:09 BIC:                11448.1743
No. Observations:   1063             Log-Likelihood:     -5710.1
Df Model:           3                F-statistic:        574.9
Df Residuals:       1059             Prob (F-statistic): 1.18e-221
R-squared:          0.620            Scale:              2722.6
----------------------------------------------------------------
            Coef.    Std.Err.      t      P>|t|     [0.025     0.975]
----------------------------------------------------------------
const      -2.1682     7.1613   -0.3028   0.7621   -16.2201   11.8837
Size       81.8376     4.0079   20.4189   0.0000    73.9732   89.7020
Baths      27.5882     3.5853    7.6949   0.0000    20.5531   34.6232
Bedrooms   -9.7682     2.8049   -3.4825   0.0005   -15.2720   -4.2643
----------------------------------------------------------------
Omnibus:              550.824       Durbin-Watson:          1.595
Prob(Omnibus):        0.000         Jarque-Bera (JB):       7406.215
Skew:                 2.056         Prob(JB):               0.000
Kurtosis:             15.260        Condition No.:          21
================================================================
```

Όλες οι μεταβλητές έχουν P-value μικρότερο του 0.01 οπότε σταματάμε και παρατηρούμε οτι το R^2 έχει μειωθεί ελαφρώς άρα οι μεταβλητές Acres, Age και Fireplace δεν ήταν καθοριστικοί παράγοντες στον υπολογισμό της $\hat{y}$.

Το τελικό μοντέλο είναι:

$$\hat{y} = -2.1682 + 81.8376 \cdot \text{Size} + 27.5882 \cdot \text{Baths} - 9.7682 \cdot \text{Bedrooms}$$