ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ ΑΠΑΛΛΑΚΤΙΚΗ ΕΡΓΑΣΙΑ 2022
ΑΝΑΣΤΑΣΙΟΣ ΜΟΥΣΤΑΚΑΣ Π16194

ΤΕΧΝΙΚΑ ΕΡΓΑΣΙΑΣ

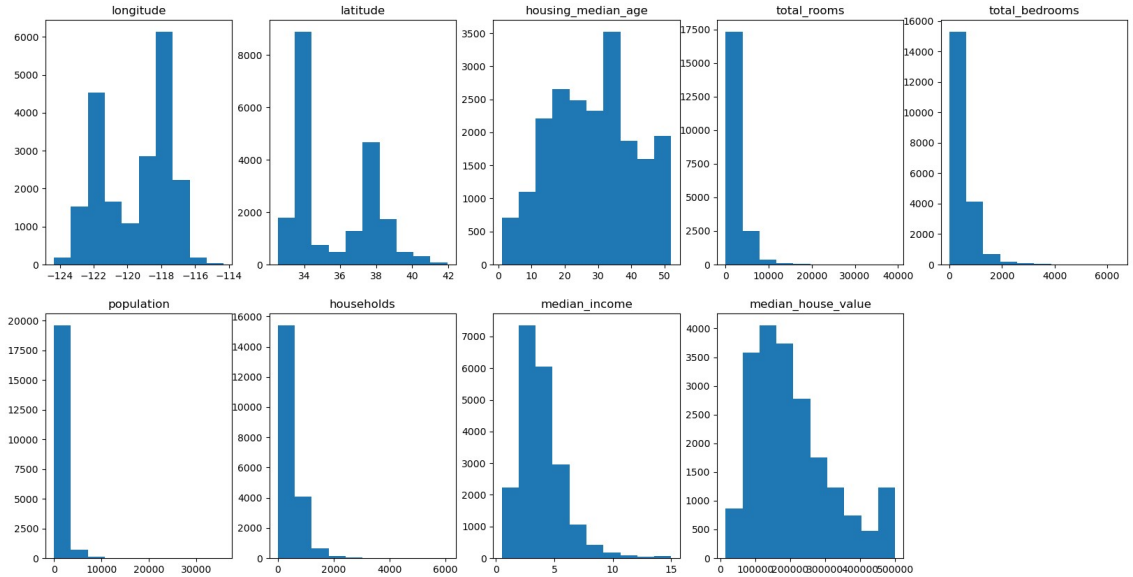
Η εργασία υλοποιήθηκε σε python και χρησιμοποιήθηκε το πακέτο sklearn για το μεγαλύτερο μέρος των υπολογισμών.Το αρχείο pattern_rec.py περιέχει την κλάση housing_price η οποία υλοποιεί τις απαντήσεις στα ερωτήματα της εργασίας.Η εκτέλεση ολόκληρης της εργασίας γίνεται μέσα από το αρχείο main.py.Για την εκτέλεση δεν χρειάζεται κάποιο εξωτερικό αρχείο πέραν του dataset.

ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Μετά την εισαγωγή του δωσμένου dataset σε ένα DataFrame του pandas,αφαιρέσαμε τις εγγραφές που περιέχουν κάποιο μηδενικό στοιχείο με την ενσωματωμένη εντολή dropna().Έπειτα για τις στήλες που περιέχουν αριθμητικά δεδομένα προχωρήσαμε σε κλιμάκωση τους με την συνάρτηση MinMaxScaler του sklearn.Αυτή η συνάρτηση αναθέτει την τιμή 1 στην μέγιστη τιμή του δωσμένου συνόλου και το 0 στην ελάχιστη τιμή,με όλες τις ενδιάμεσες τιμές να έχουν κάποια ενδιάμεση τιμή αναλογική με το πόσο κοντά είναι σε κάποιο από τα δύο άκρα.Έπειτα για τα δεδομένα σε μορφή κατηγορήματος χρησιμοποιήσαμε την συνάρτηση OneHotEncode η οποία δημιουργει έναν πίνακα με την μορφή πίνακα αληθείας όπου κάθε στήλη αντιστοιχεί σε κάποια εγγραφή αλφαριθμητικού και πλέον οι στήλες έχουν την τιμή 1,εάν στην αρχική στήλη συναντόταν το αλφαριθμητικό,ενώ 0 εάν δεν ήταν αυτό στην αρχική στήλη.

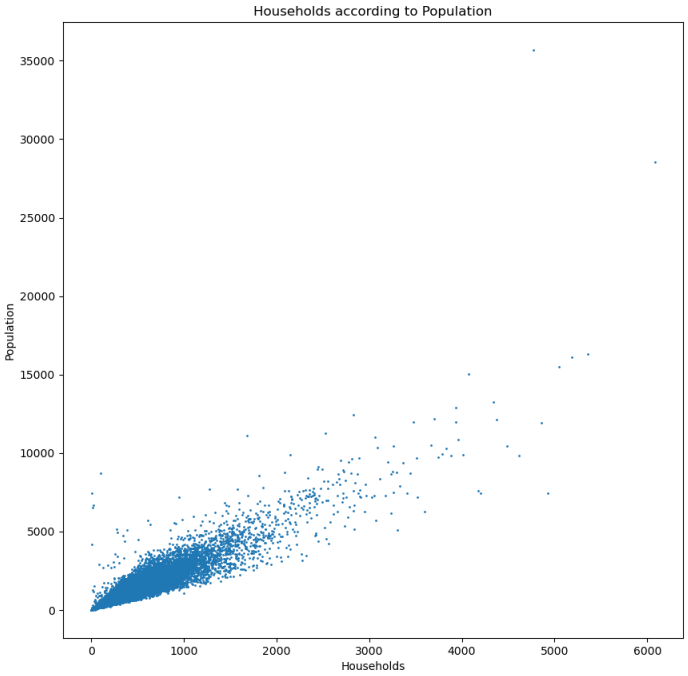
ΟΠΤΙΚΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

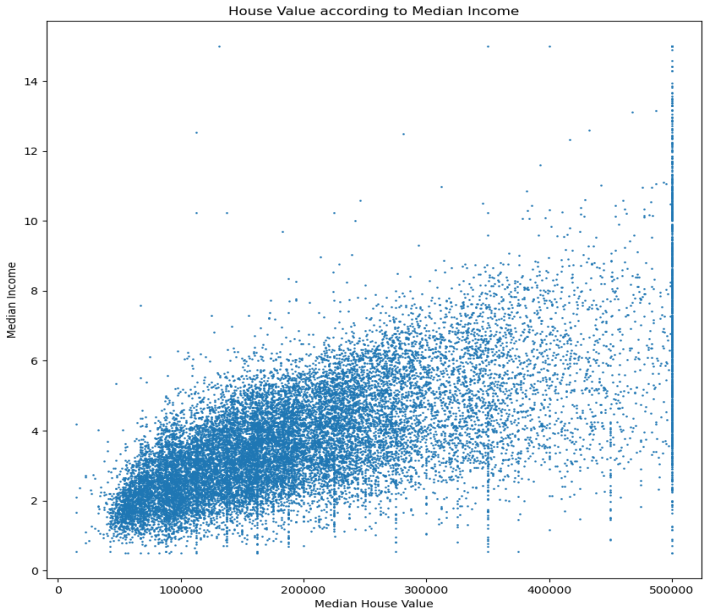
Για την οπτικοποίηση χρησιμοποιήθηκε το πακέτο matplotlib και πιο συγκεκριμένα η pyplot. Αρχικά εμφανίζεται το διάγραμμα που αποτελείται απο τα ιστογράμματα συχνότητας των μεταβλήτων του προβλήματος.



Ένα ιστόγραμμα είναι ένα γράφημα στηλών(bar graph),όπου κάθε στήλη είναι κάδος ορισμένου μήκους,το ύψος της οποίας αναπαριστά το πλύθος των εμφανίσεων των τιμών που είναι μέσα στα όρια του κάδου στο σύνολο των δεδομένων.

Στην συνέχεια θα παραθέσουμε γραφήματα που συνδυάζουν 2 μεταβλητές ώστε να παρουσιάσουν κάποιο συμπέρασμα επί των δεδομένων.Το παράδειγμα εμφανίζει 4 διαγράμματα,παρακάτω θα παρατεθούν ενδεικτικά 2 από αυτά.





ΠΑΛΛΙΝΔΡΟΜΗΣΗ

1)Perceptron.Σε αυτό το ερώτημα υλοποιήσαμε έναν απλό γραμμικό ταξινομητή της μορφής perceptron. Το σύνολο των δεδομένων χωρίστηκε σε δύο υποκλάσεις με βάση την τιμή(median_house_value).Στην πρώτη κατηγορία με flag=-1 ανήκουν όσα δείγματα έχουν τιμή κάτω απο την μέση(mean) τιμή του συνόλου,ενώ στην δεύτερη με flag=1 ανήκουν όσα δείγματα έχουν τιμή πάνω απο την μέση τιμή.Το σχήμα αυτό εκτελείται για 10 εποχές εκπαίδευσης με ρυθμό μάθησης 0.05 και χωρισμό train/test 90/10.Το παράδειγμα εκτελείται με την μέθοδο perceptron της κλάσης housing_price,βασισμένο στην αλγόριθμο της σελίδας 102 του βιβλίου αναγνώριση προτύπων και επιστρέφει το εκπαιδευμένο διάνυσμα βαρών,το μέσο τετραγωνικό σφάλμα(MSE) και το μέσο απόλυτο σφάλμα(MAE).

```
==PERCEPTRON==
Mean Absolute Error: 0.6467710371819961
Mean Squared Error: 1.2935420743639923
Weight Vector: [   7.89811441   34.78796855   82.4927397    41.08962178   27.07672463
    2.70565613   27.93831718  279.19001577  155.21671641 -175.88965524
    2.3347938    71.48469603   84.02566495]
```

2)Least Squares.Σε αυτό το ερώτημα υλοποιήσαμε τον αλγόριθμο ελαχίστων τετραγώνων με βάση τον αλγόριθμο της σελίδας 118 του βιβλίου αναγνώριση προτύπων.Όπως και παραπάνω ο χωρισμός train/test είναι 90/10 και το μοντέλο επιστρέφει το μέσο και το απόλυτο τετραγωνικό σφάλμα καθώς και το διάνυσμα βαρών.Η εκτέλεση γίνεται με κλήση της συνάρτησης least_squares της κλάσης housing_price.

```
==Least Squares==
Mean Absolute Error: 0.30724070450097846
Mean Squared Error: 0.6144814090019569
Weight Vector: [ -2.72709823  -2.24980725   0.29786253   1.34068714   0.91850462
 -14.32489995   5.65749308   3.3655733    0.97083544   0.63987351
   1.95865147   0.81895873   0.9546216 ]
```

3)Παλλινδρόμηση.Σε αυτό το βήμα χρησιμοποιήσαμε την ενσωματωμένη συνάρτηση του sklearn για παλλινδρόμηση δεδομένων με στόχο την πρόβλεψη της τιμής των ακινήτων.Η λειτουργία αυτή υλοποιείται με την κλήση της μεθόδου regression() και επιστρέφει την εκτίμηση του μοντέλου καθώς και τα μέσο και το απόλυτο τετραγωνικό σφάλμα.

```
==Regression==
Regression Score: 0.6413830780531804
Mean Absolute Error: 0.10170732757665885
Mean Squared Error: 0.01945709942147531
```