Αλέξανδρος Ξιάρχος
st1059619@ceid.upatras.gr
1059619

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ · ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΜΑΘΗΣΗΣ

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ · 2023–2024

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΡΩΤΗΜΑ 1

Το σύνολο δεδομένων περιλαμβάνει 22 `.csv` αρχεία που αντιστοιχούν σε 22 συμμετέχοντες. Σύμφωνα με την περιγραφή του dataset, περιλαμβάνεται η στήλη `timestamp`, με την ημερομηνία και ώρα, οι στήλες $\texttt{back}_{x,y,z}$ και $\texttt{thigh}_{x,y,z}$ με τις τιμές του κάθε αισθητήρα για κάθε διάσταση, και η στήλη `label`, η οποία προσδιορίζει τη δραστηριότητα του συμμετέχοντα τη δεδομένη στιγμή.

Η στήλη `label` παίρνει τις εξής τιμές:

```
1 - Walking
2 - Running
3 - Shuffling
4 - Stairs (ascending)
5 - Stairs (descending)
6 - Standing
8: lying
13 - Cycling (sit)
14 - Cycling (stand)
130 - Cycling (sit, inactive)
140 - Cycling (stand, inactive)
```

Για την εισαγωγή και την προεπεξεργασία των αρχείων, θα χρησιμοποιήσουμε τη βιβλιοθήκη `pandas` ενώ για την οπτικοποίηση τις βιβλιοθήκες `matplotlib` και `seaborn` της Python.

1.1 ΑΝΑΛΥΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Εισάγουμε τα `.csv` αρχεία μέσω της `os` βιβλιοθήκης και της `read_csv()` συνάρτησης. Καταρχάς, χρησιμοποιώντας τη `head()` μπορούμε να δούμε τις πρώτες εγγραφές του dataset μας. Για παράδειγμα για το πρώτο αρχείο του συνόλου δεδομένων `S006.csv`:

	timestamp	back_x	back_y	back_z	thigh_x	thigh_y	thigh_z	label
0	2019-01-12 00:00:00.000	-0.760242	0.299570	0.468570	-5.092732	-0.298644	0.709439	6
1	2019-01-12 00:00:00.010	-0.530138	0.281880	0.319987	0.900547	0.286944	0.340309	6
2	2019-01-12 00:00:00.020	-1.170922	0.186353	-0.167010	-0.035442	-0.078423	-0.515212	6
3	2019-01-12 00:00:00.030	-0.648772	0.016579	-0.054284	-1.554248	-0.950978	-0.221140	6
4	2019-01-12 00:00:00.040	-0.355071	-0.051831	-0.113419	-0.547471	0.140903	-0.653782	6

Μέσω της `info()` εξάγουμε το συμπέρασμα πώς για κάθε χρονική στιγμή δίνονται οι τιμές των αισθητήρων, αποθηκευμένες ως `float24`, στις τρεις διαστάσεις (x, y, z) για τις περιοχές της πλάτης και του μηρού, καθώς και ένα `int64` για το `label`. Η ίδια μορφολογία παρατηρείται σε όλα τα `.csv` του συνόλου δεδομένων, με κάποιες διαφοροποιήσεις που θα αναλυθούν στη συνέχεια.

Παρόλο που στην περιγραφή αναφέρεται πως δεν υπάρχουν missing values, για να ελέγξουμε την ακεραιότητα του dataset, μέσω της συνάρτησης `concat()` ενώνουμε όλα τα 22 αρχεία σε ένα ενιαίο dataframe. Τρέχοντας την `isnull().sum()`, έχουμε:

	sum
timestamp	0
$\texttt{back}_x$	0
$\texttt{back}_y$	0
$\texttt{back}_z$	0
$\texttt{thigh}_x$	0
$\texttt{thigh}_y$	0
$\texttt{thigh}_z$	0
label	0
index	5740689
Unnamed: 0	6323682

Παρατηρούμε πως στις στήλες $\texttt{back}_{x,y,z}$ και $\texttt{thigh}_{x,y,z}$, οι οποίες είναι και αυτές που μας ενδιαφέρουν, όντως δεν παρατηρούνται missing values. Όμως, έχουν εμφανιστεί `NaN` τιμές στις στήλες `"index"` και `"Unnamed: 0"`, οι οποίες στήλες μάλλον θα εμφανίζονται επιπλέον σε κάποια αρχεία.

Ελέγχοντας όλα τα αρχεία, η στήλη `"index"` εμφανίζεται στα αρχεία `S015.csv` και `S021.csv` και η στήλη `"Unnamed: 0"` στο αρχείο `S023.csv`. Έπειτα από έλεγχο, φαίνεται πως πρόκειται για δείκτες αύξουσας αρίθμησης που δεν προσφέρουν κάποια επιπλέον πληροφορία. Επομένως, μπορούμε να τις αφαιρέσουμε χρησιμοποιώντας τη συνάρτηση `drop('όνομα', axis=1)`.

Χρησιμοποιώντας τη συνάρτηση `describe()` μπορούμε να υπολογίσουμε βασικές στατιστικές μετρικές για τα δεδομένα μας. Η συνάρτηση επιστρέφει ένα dataframe με τις ακόλουθες στήλες:

- **`count`**: ο συνολικός αριθμός των μη-μηδενικών τιμών για κάθε στήλη.
- **`mean`**: ο μέσος όρος των τιμών για κάθε στήλη.
- **`min`**: η ελάχιστη τιμή για κάθε στήλη.
- **`25%`**: η τιμή του 25ου εκατοστημορίου για κάθε στήλη.
- **`50%`**: η τιμή του 50ου εκατοστημορίου για κάθε στήλη.
- **`75%`**: η τιμή του 75ου εκατοστημορίου για κάθε στήλη.
- **`max`**: η μέγιστη τιμή για κάθε στήλη.

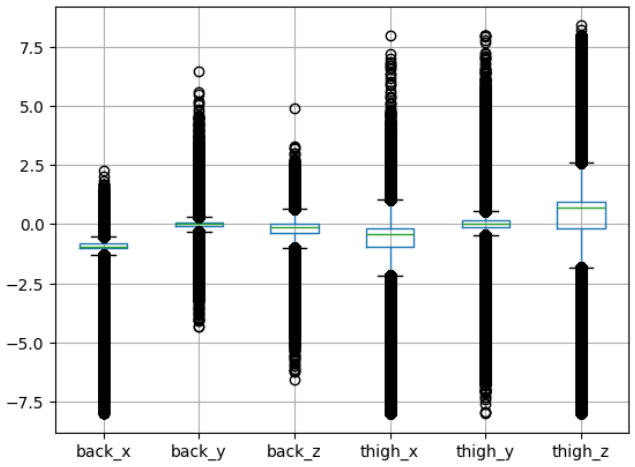
Ενώνοντας συγκεντρωτικά τις μετρήσεις όλων των συμμετεχόντων στο `df_combined` μέσω της `concat()`, αυτά είναι τα **βασικά συγκεντρωτικά στατιστικά μεγέθη** όπως προκύπτουν από το `describe()` για όλες τις μετρήσεις από τους συμμετέχοντες, έχοντας αφαιρέσει την ετικέτα `label` μιας και αποτελείται από κατηγορικές τιμές:

	back_x	back_y	back_z	thigh_x	thigh_y	thigh_z
count	6461328	6461328	6461328	6461328	6461328	6461328
mean	-0.884957	-0.013261	-0.169378	-0.594888	0.020877	0.374916
std	0.377592	0.231171	0.364738	0.626347	0.388451	0.736098
min	-8.000000	-4.307617	-6.574463	-8.000000	-7.997314	-8.000000
25%	-1.002393	-0.083129	-0.372070	-0.974211	-0.100087	-0.155714
50%	-0.974900	0.002594	-0.137451	-0.421731	0.032629	0.700439
75%	-0.812303	0.072510	0.046473	-0.167876	0.154951	0.948675
max	2.291708	6.491943	4.909483	7.999756	7.999756	8.406235

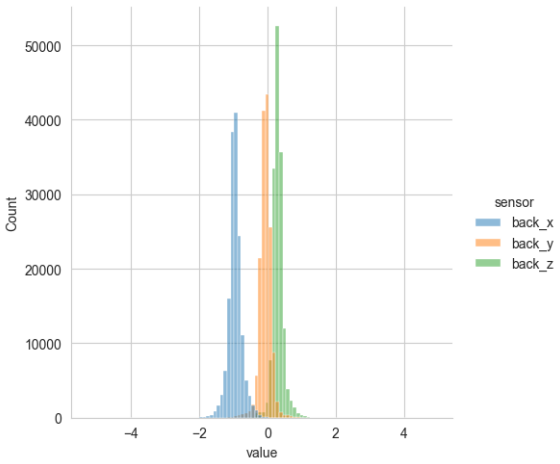
Ως αρχικές παρατηρήσεις, βλέπουμε πως οι τιμές βρίσκονται στο διάστημα $[-8, 8]$, ενώ η τυπική τους απόκλιση είναι μικρή, το οποίο δείχνει ότι οι μετρήσεις είναι αρκετά συμπυκνωμένες γύρω από τον μέσο όρο που είναι κοντά στο μηδέν. Προφανώς ελέγχοντας κάθε συμμετέχοντα ξεχωριστά, μπορεί να διεξαχθούν αντίστοιχα συμπεράσματα.

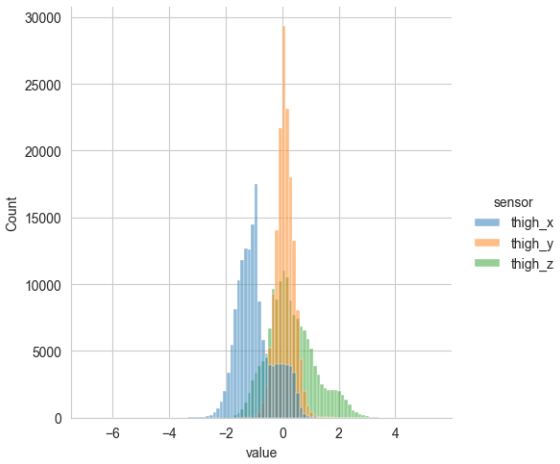
1.2 ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

Μέσω της `plotbox()` της `Matplotlib`, μπορούμε να δημιουργήσουμε το διάγραμμα των τιμών της `df_combined` για μια πρώτη οπτικοποίηση των δεδομένων:

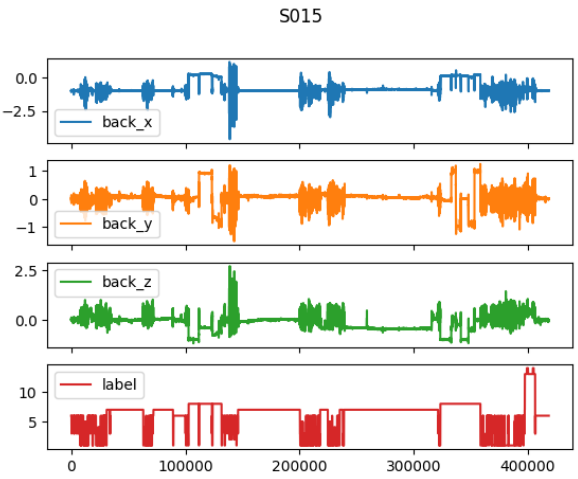


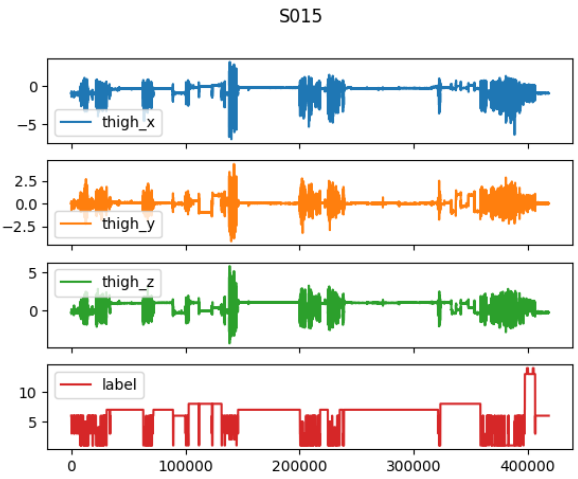
Πέρα από τις προηγούμενες παρατηρήσεις που επιβεβαιώνονται, επιπλέον παρατηρούμε μια συμμετρικότητα κοντά στο μηδέν για κάθε διάσταση. Επίσης, χρησιμοποιώντας την `displot()`, μπορούμε να οπτικοποιήσουμε το πώς κατανέμονται οι τιμές. Ενδεικτικά για τον `S009`:





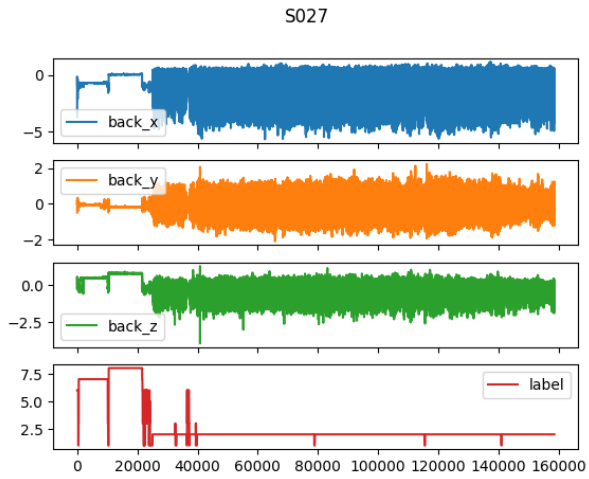
Χρησιμοποιώντας την `plot()`, μπορούμε να δημιουργήσουμε subplots για τις στήλες $\texttt{back}_{x,y,z}$ και $\texttt{thigh}_{x,y,z}$ ενός τυχαίου συμμετέχοντα, `S015`:

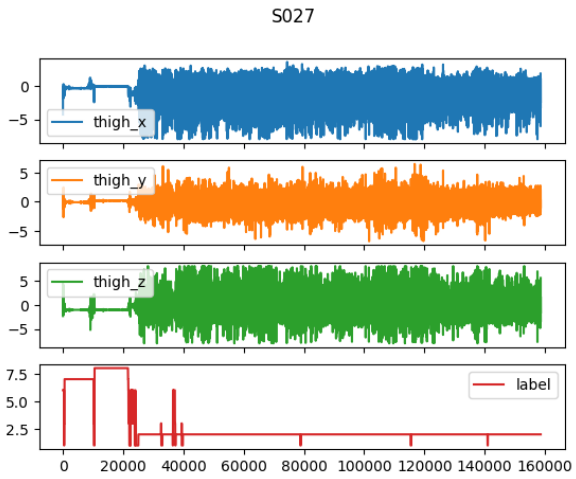




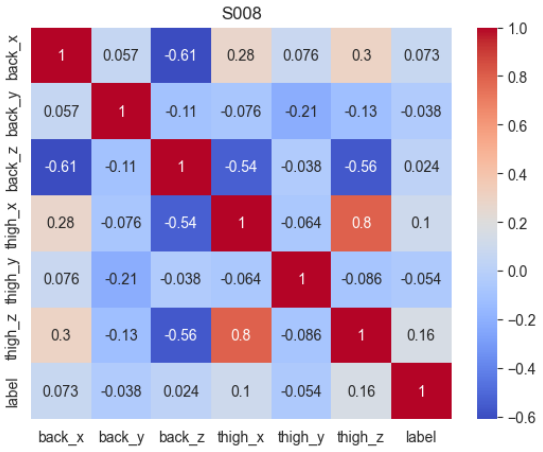
Καταρχάς είναι προφανές ότι υπάρχει μια αντιστοιχία μεταξύ των κινήσεων της πλάτης και του μηρού, καθώς παρατηρούμε ομοιότητες στα διαγράμματα. Φαίνεται ότι ο συμμετέχοντας κατά τη διάρκεια της μέτρησης μεταβάλλει τη φυσική του δραστηριότητα, καθώς υπάρχουν στιγμές που δεν υπάρχουν έντονες διακυμάνσεις των τιμών των μετρήσεων των αισθητήρων (κάτι που αποτυπώνεται και στη τιμή του `label`, καθώς φαίνεται να είναι `8 - Standing`) και άλλες όπου είναι πιο ενεργός, με την τιμή του `label` να αλλάζει και αυτή. Μάλιστα, δεδομένου της έντονης μεταβολής του `label`, μπορούμε να υποθέσουμε πως ο συμμετέχοντας ανεβοκατεβαίνει σκάλες.

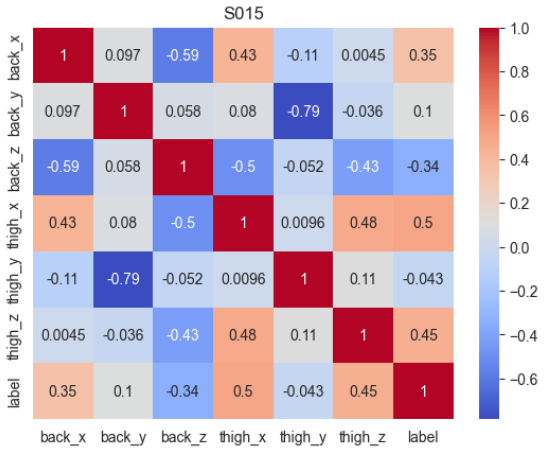
Από την άλλη, για παράδειγμα στον συμμετέχοντα `S027`, η `label` παραμένει σταθερή καθ'όλη τη διάρκεια της έντονης φυσικής δραστηριότητας, με τιμή κοντά στο 2.5, κάτι από το οποίο μπορούμε να συμπεράνουμε πως ο συμμετέχοντας τρέχει:



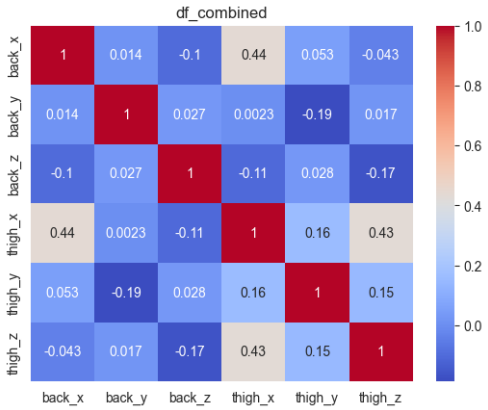


Τέλος, για τον εντοπισμό συσχετίσεων, μπορούμε να δημιουργήσουμε ένα `heatmap()` μέσω της `seaborn`. Για παράδειγμα, για τον συμμετέχοντα `S008` φαίνεται πως υπάρχει μια σαφής συσχέτιση ανάμεσα στις στήλες `back_x` + `back_z`, `thigh_x` + `thigh_z` και `back_z` + `thigh_x` ενώ στον `S015` ανάμεσα στις στήλες `back_x` + `back_z` και `back_y` + `thigh_y`.





Δημιουργώντας heatmap και για το συγκεντρωτικό dataframe `df_combined`, παρατηρούμε μια αμυδρή συσχέτιση ανάμεσα στα `thigh_x` + `back_x` και `thigh_x` + `thigh_z`.



Στο παράρτημα παρατίθονται τα plots και τα heatmaps για όλους τους συμμετέχοντες.