

Καθαρισμός δεδομένων



Τα δεδομένα που συλλέγουμε από μια έρευνα χρειάζονται καθαρισμό, όταν υπάρχουν σε αυτά κάποιες προβληματικές τιμές όπως: ελλιπείς τιμές (missing values), τιμές εκτός αποδεκτών ορίων (outliers) ή υπάρχουν διπλοεγγραφές. Είτε με αυτόματο τρόπο είτε χειρωνακτικά, αυτά τα προβλήματα μπορούν να επιλυθούν. Οι ελλιπείς τιμές μπορούν να συμπληρωθούν. Ο πιο απλός από τους τρόπους συμπλήρωσης είναι να συμπληρωθούν αυτά που λείπουν με το μέσο όρο των υπολοίπων τιμών (υπάρχουν πολλοί περισσότεροι και πολύ σύνθετοι τρόποι συμπλήρωσης που δεν θα συζητήσουμε εδώ). Όταν εντοπίζονται δεδομένα των οποίων οι τιμές μας προβληματίζουν (με τυπογραφικά σφάλματα, λάθος μετρήσεις κ.λπ.), γίνονται προσπάθειες αυτά να διορθωθούν. Τέλος, η αφαίρεση διπλοεγγραφών μπορεί να γίνει και αυτοματοποιημένα, με τη χρήση κάποιου εργαλείου.