HW3 – Univariate and Multivariate Analysis and Visualization – Missing values and Outliers

Σας δίνεται το dataset diabetes.csv που αποτελείται από 403 εγγραφές με 22 μεταβλητές, εκ των οποίων οι 7 είναι computed (total_chol_hdl_ratio, BMI, waist_hip_ratio, BMI_cat_5, age_cat, waist_hip_cat, BMI_cat_3). Η φύση των μεταβλητών είναι κατανοητή από το όνομά τους. Δίνονται διευκρινίσεις για τις εξής ονομασίες: id (μοναδικός κωδικός ασθενή), hdl (καλή χοληστερίνη), glycosoloatedhgb (γλυκοζυλιωμένη αιμοσφαιρίνη), bp.1s (συστολική πίεση), bp.1d (διαστολική πίεση).

Χωρίς να προβείτε σε διαχείριση (removal, imputation) των NA, αλλά μόνο των πιθανών outliers, χρησιμοποιείστε τεχνικές των κεφαλαίων 9-13 του edav.info για να αναλύσετε και οπτικοποιήσετε τις μεταβλητές.

Σας δίνεται ένα αντίγραφο του diabetes.csv με το όνομα diabetes2.csv στο οποίο έχουν κρατηθεί μόνο οι στήλες total_cholesterol, age, gender, weight και με τεχνητό τρόπο έχει αλλαχτεί το 30% των περιεχομένων σε NA. Χρησιμοποιήστε τεχνικές διαχείρισης missing values για να “επισκευάσετε” το dataset σας και αξιολογείστε το αποτέλεσμα που πετυχαίνουν οι διαφορετικές τεχνικές (δηλαδή πόσο κοντά στις πραγματικές τιμές είναι οι προβλεπόμενες).

Παραδώστε ένα αρχείο Rmd και το αντίστοιχο knitted αρχείο html.
Εννοείται ότι θα πρέπει να συμπεριλάβετε εκτενή σχολιασμό σε ο,τιδήποτε παρουσιάζετε.

Μερικές βοηθητικές πηγές σχετικές με missing values και outliers:

1. Discussion paper: An introduction to data cleaning with R
2. Πολύ καλό βιβλίο της Wiley από του συγγραφείς του παραπάνω paper που είναι διαθέσιμο μέσω του Healink: Statistical Data Cleaning with Applications in R
3. Άρθρο από το RPubs: Data Cleansing (ιδιαίτερο ενδιαφέρον παρουσιάζουν τα Automated Reports με το dlookr)