DATA MINING ACADEMY 2019

Final project - June 2019

Η εργασία αφορά τα δεδομένα που βρίσκονται στο αρχείο **`real_estate.xls`**. Συγκεκριμένα, το αρχείο περιέχει πληροφορίες από ένα τυχαίο δείγμα 100 μονοκατοικιών με κήπο. Τα ακίνητα βρίσκονται στην Κομητεία Γουέικ, της πολιτείας της Βόρειας Καρολίνας των Η.Π.Α.. Οι 6 μεταβλητές του αρχείου είναι:

TotalUSD: Η συνολική αντικειμενική αξία του ακινήτου σε δολάρια Η.Π.Α. (**συνεχής**)
SqM_Building: Τα τετραγωνικά μέτρα της οικίας (**συνεχής**)
SqM_Land: Τα τετραγωνικά μέτρα του κήπου (**συνεχής**)
Year: Το έτος κατασκευής της οικίας (**διακριτή**)
Baths: Ο αριθμός μπάνιων της οικίας (**διακριτή**)
Fireplace: Η ύπαρξη τζακιού στην οικία (**κατηγορική**)

Να δώσετε ένα αρχείο ***`.R`*** *με τους κώδικες και ένα αρχείο με τα output, τα γραφήματα και συνοπτικές απαντήσεις.*

Α. Να εισάγετε τα δεδομένα στην **`RStudio`**. Έπειτα, να ελέγξετε αν υπάρχουν NAs, πόσα είναι και σε ποια σειρά.

Β. Να δώστε τα περιγραφικά στοιχεία και τα κατάλληλα γραφήματα για τις ποσοτικές μεταβλητές και να παρέχετε ένα πολλαπλό γράφημα για αυτές με την εντολή **`chart.Correlation()`**.

Γ. Να κάνετε ένα έλεγχο κανονικότητας για τις συνεχείς μεταβλητές. Να λογαριθμίσετε τα δεδομένα, αν απορρίψετε την κανονικότητα.

Για τα μετασχηματισμένα δεδομένα:

Δ. Να παρέχετε ένα boxplot της μεταβλητής **TotalUSD** σε σχέση με τη μεταβλητή **Fireplace**. Επιπλέον να ελέγξετε αν η **Fireplace** επιδρά στις τιμές της **TotalUSD**.

Ε. Να κατασκευάσετε ένα μοντέλο κύριων επιδράσεων με εξαρτημένη μεταβλητή την **TotalUSD** και να καταλήξετε σε ένα τελικό μοντέλο με την stepwise process.

Για το τελικό μοντέλο:

ΣΤ. Να ελέγξετε την καλή προσαρμογή του μέσα από τα κατάλληλα γραφήματα και τους κατάλληλους ελέγχους και να δώσετε την ερμηνεία των παραμέτρων.

Ζ. Να εκτιμήσετε την τιμή του σπιτιού που βρίσκεται στην $5^{η}$ σειρά του πίνακα.