ΕΡΓΑΣΤΗΡΙΑΚΗ ΕΡΓΑΣΙΑ
ΦΑΣΗ 2: ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Καλείστε να κάνετε Διερευνητική Ανάλυση Δεδομένων πάνω στο επιλεγμένο dataset σας. Η ανάλυση θα περιλαμβάνει **1) μονομεταβλητή ανάλυση για όλες τις μεταβλητές** και **2) πολυμεταβλητή ανάλυση κατά την κρίση σας**.

0) Προεπεξεργασία δεδομένων

Διαχείριση ελλειπουσών τιμών: Θα πρέπει στη φάση αυτή να έχετε αποφασίσει πώς να διαχειριστείτε τις ελλείπουσες τιμές στο dataset σας. Αν υπάρχουν ακόμα αμφιβολίες, επικοινωνήστε άμεσα. Αν η απόφαση έχει ληφθεί, υλοποιήστε την: διαγράψτε τις γραμμές ή τις στήλες όπου κρίθηκε προτιμότερο, ή αντικαταστήστε με εκτιμήσεις όπου αυτό αποφασίστηκε.

Διαχείριση κατηγορικών μεταβλητών: κάποιες μέθοδοι της περιγραφικής ανάλυσης απαιτούν one-hot encoding (κωδικοποίηση σε ψευδομεταβλητές / dummy variables). Π.χ. Pearson's Correlation δε μπορούμε να υπολογίσουμε με κατηγορικές μεταβλητές - πρέπει να τις κάνουμε one-hot. Κάποιες μέθοδοι οπτικοποίησης από την άλλη, δουλεύουν καλύτερα με κατηγορίες (π.χ. ποσοστά επιβίωσης για **`embarked=C`** ή **`Q`** ή **`S`** -θα ήταν πιο δύσκολο να τα οπτικοποιήσουμε αν είχαμε κάνει τρεις dummy στήλες **`embarked_C`**, **`embarked_Q`**, **`embarked_S`**). Η λύση είναι να κρατήσουμε και τις δύο εκδοχές στο dataframe, και να χρησιμοποιούμε όποια εκδοχή ταιριάζει σε κάθε μέθοδο.

Να ακολουθεί κείμενο σε Markdown που να περιγράφει τις επιλογές σας και την αρχική και τελική εικόνα του dataset.

1) Μονομεταβλητή ανάλυση

Κατά τη μονομεταβλητή ανάλυση να παραθέσετε:

Α) Μέτρα θέσης και διασποράς της κάθε μεταβλητής χρησιμοποιήστε την **`df.describe()`** και συμπληρώστε με επιπλέον εντολές για να πάρετε όσους δείκτες έχουμε χρησιμοποιήσει.

Β) Ιστόγραμμα ή ραβδόγραμμα (bar plot) αναλόγως του τύπου της μεταβλητής.

Γ) Box plot (όπου έχει νόημα να εφαρμοστεί και συνεισφέρει πληροφορία).

Συνδυάζοντας τις πληροφορίες αυτές, να αναπτυχθεί κείμενο Markdown στο οποίο 1) θα αναφέρονται τυχόν επιλογές και αποφάσεις που πήρατε για την παρουσίαση των μεταβλητών, και θα δίνεται η γενική εικόνα των κατανομών όλων των μεταβλητών και 2) θα αναφέρονται οι ενδιαφέρουσες παρατηρήσεις που μπορούμε να κάνουμε. Σε περιπτώσεις datasets με πάρα πολλές μεταβλητές, για να αποφύγετε το τεράστιο κείμενο

μπορείτε να περιγράψετε περιληπτικά σε ομάδες (π.χ. «οι X1 και X2 έχουν έντονη ασυμμετρία δεξιά, ενώ οι περισσότερες είναι συμμετρικές»).

Hint: *Για να περιγράψετε πολλή πληροφορία σε συμπυκνωμένη μορφή σε Markdown, μπορείτε να κάνετε και πίνακες:* https://stackoverflow.com/questions/48655801/tables-in-markdown-in-jupyter

2) Πολυμεταβλητή ανάλυση

Παραθέστε τον πίνακα συσχετίσεων μεταξύ όλων των μεταβλητών (κατά προτίμηση με heatmap) και περιγράψτε σε Markdown τις πιο σημαντικές παρατηρήσεις.

Αν το πλήθος και οι τύποι των μεταβλητών το επιτρέπουν, παρουσιάστε pair plot για όσες μεταβλητές μπορείτε -ιδανικά χρωματισμένο με βάση τη μεταβλητή-στόχο. Σε περίπτωση υπερβολικά μεγάλου πλήθους μεταβλητών, μπορείτε να επιλέξτε ένα ενδεικτικό υποσύνολο. *Μπορείτε ακόμα αν το επιθυμείτε να δημιουργήσετε νέες μεταβλητές, ως συνάρτηση των υπαρχόντων, που να δείχνουν πιο συμπυκνωμένα την πληροφορία. Στην περίπτωση αυτή μην παραλείψετε να τις συμπεριλάβετε στη μονομεταβλητή ανάλυση.*

Παρουσιάστε ζεύγη ή τριάδες μεταβλητών των οποίων οι από κοινού κατανομές παρουσιάζουν ενδιαφέρον. Μελετήστε τη συμπεριφορά διάφορων μεταβλητών μαζί με τη μεταβλητή-στόχο, αλλά μην περιοριστείτε απαραίτητα σε αυτήν. Για την απεικόνιση των ζευγών/τριάδων χρησιμοποιήστε τα κατάλληλα γραφήματα αναλόγως των τύπων και των τιμών των μεταβλητών: scatterplots, ραβδογράμματα/ιστογράμματα ανά κατηγορία, box plots ανά κατήγορία.

Κριτήρια βαθμολόγησης:

Α. **Πληρότητα**: Υποβολή .ipynb μαζί με .csv ώστε να μπορεί να εκτελεστεί ο κώδικας. Κώδικας χωρίς σφάλματα. Συμπερίληψη όλων των αποτελεσμάτων, δεικτών, και γραφημάτων που ορίζει η εκφώνηση.

Β. **Επιστημονική ορθότητα**: Σωστή χρήση της ορολογίας. Τα συμπεράσματα που διατυπώνετε να δικαιολογούνται από τη μεθοδολογία που χρησιμοποιείτε. Εφαρμογή των κατάλληλων μεθόδων ανά περίπτωση.

Γ. **Παρουσίαση**: Μόνο χρήσιμος κώδικας, όχι περιττά μπλοκ ή περιττές εντολές. Σχολιασμός κώδικα σε περιπτώσεις περίπλοκων διεργασιών (όχι σε μεμονωμένες απλές εντολές). Επεξηγηματικά μηνύματα στα print() που να επιτρέπουν κατανόηση των outputs από τρίτους χωρίς να πρέπει να ανατρέχουμε στον κώδικα. Όλη η ουσιαστική πληροφορία να παρουσιάζεται και **αυτοτελώς** στα Markdown blocks, τα οποία θα περιλαμβάνουν και τις σημαντικές αριθμητικές τιμές και παρατηρήσεις από τα outputs (π.χ. «το μεγαλύτερο correlation εμφανίζεται μεταξύ X1 και X2 και είναι 0.78»). Καλή χρήση Markdown (ίδιο μέγεθος τίτλων/υποτίτλων παντού, χρήση

bold/italics/bullets για ευανάγνωστο κείμενο). Ο στόχος είναι να μπορεί κανείς να παρακολουθήσει τη δουλειά σας βλέποντας μόνο τη Markdown και τα γραφήματα, και *αν θέλει να μπει βαθύτερα* να βλέπει τον κώδικα και τα print outputs.

Δ. **Πρωτοβουλία και κριτική σκέψη**: επιλογή ενδιαφερουσών μεταβλητών για ανά δύο ή ανά τρεις απεικόνιση. Ερμηνεία των outputs με κριτήρια του πραγματικού κόσμου, εξαγωγή ενδιαφερόντων συμπερασμάτων για τα παρατηρούμενα φαινόμενα.