**EDA ΑΡΧΕΙΟΥ eda\_cleaned.csv**

**1. Εξερευνητική Ανάλυση Δεδομένων (EDA)**

α. Στατιστική περιγραφή των δεδομένων:

Το σύνολο δεδομένων μας αποτελείται από διάφορους φακέλους και αρχεία. Συγκεκριμένα, έχουμε τα δεδομένα από τον φάκελο "dicom\_dir", ο οποίος περιλαμβάνει πληροφορίες όπως το όνομα του αρχείου (File), το ID του ασθενούς (Patient ID), τον τύπο της εικόνας (Modality), το μέγεθος της εικόνας (Image Size), την ηλικία και το φύλο του ασθενούς (Patient Age, Patient Gender), καθώς και την ημερομηνία της μελέτης (Study Date). Επίσης, έχουμε τον φάκελο "tiff\_images", ο οποίος περιλαμβάνει δεδομένα για το όνομα του αρχείου (Filename), το μέγεθος της εικόνας (Image Size), τον τύπο της εικόνας (Image Type), το εύρος των pixels (Pixel Range), και άλλες πληροφορίες όπως αν η εικόνα είναι αντιστραμμένη (Is Inverted), αν είναι διπλότυπη (Is Duplicate) και αν το μέγεθος είναι έγκυρο (Valid Size). Επιπλέον, υπάρχει το αρχείο "overview", το οποίο παρέχει πληροφορίες για την ηλικία (Age), το Contrast, το ContrastTag, καθώς και άλλες λεπτομέρειες που συνδέουν τα δεδομένα εικόνας με το metadata (raw\_input\_path, id, tiff\_name, dicom\_name). Τέλος, το αρχείο "full\_archive.npz" περιέχει τα ids των εικόνων (idx) και τα arrays των pixels των εικόνων (image), παρέχοντας έτσι πληροφορίες για τις διαστάσεις των εικόνων.

Αν και η εργασία προέβλεπε τουλάχιστον 1000 δείγματα για την ανάλυση, ο υπεύθυνος του προγράμματος μάς ενημέρωσε ότι μπορούμε να προχωρήσουμε με 475 εικόνες (αρχείο full\_archive.npz), καθώς αυτό είναι αποδεκτό για τον σκοπό μας. Από τις 475 εικόνες, οι 100 εικόνες συνδέονται με τα metadata που προέρχονται από τα αρχεία dicom\_dir, tiff\_images και overview.

Για να διευκολύνουμε την ανάλυση, δημιουργήσαμε CSV αρχεία με τις πληροφορίες από τα "dicom\_dir", "tiff\_images" και "overview", ώστε να μπορέσουμε να συγκρίνουμε και να ταυτοποιήσουμε τις εικόνες και τα άτομα στα οποία ανήκουν. Για την εκτέλεση της εξερευνητικής ανάλυσης δεδομένων (EDA), συνδυάσαμε όλες τις πληροφορίες από τις πηγές μας (dicom\_dir, tiff\_images, overview, full\_archive.npz) και δημιουργήσαμε το ενιαίο αρχείο eda\_cleaned.csv, το οποίο περιλαμβάνει τα δεδομένα για τις 100 εικόνες που έχουν αντίστοιχα labels και metadata. Το αρχείο αυτό μας επέτρεψε να προχωρήσουμε στην ανάλυση των χαρακτηριστικών των εικόνων και να εξετάσουμε τυχόν ανισορροπίες ή ακραίες τιμές (outliers) που ενδέχεται να επηρεάσουν τα επόμενα βήματα της διαδικασίας.



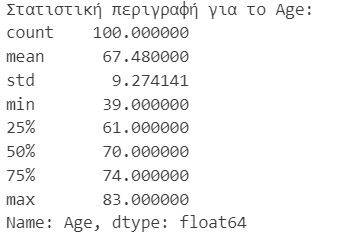
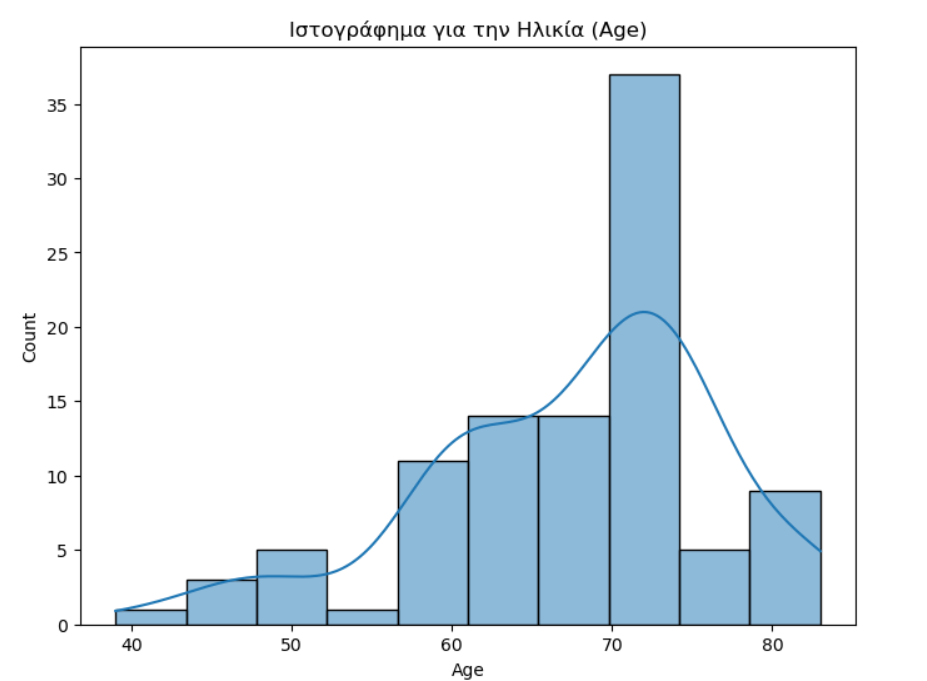
Παράδειγμα δεδομένων στο eda\_cleaned.csv

Η εξερευνητική ανάλυση δεδομένων (EDA) πραγματοποιήθηκε για τις 100 εικόνες που συνδέονται με τα metadata του συνόλου δεδομένων, διότι περιλαμβάνουν labels (σημειώσεις που περιγράφουν τη σχέση της κάθε εικόνας με την κατηγορία στόχο, όπως το χαρακτηριστικό Contrast). Αυτές οι εικόνες είναι κατάλληλες για ανάλυση και εκπαίδευση μοντέλων μηχανικής μάθησης. Αντίθετα, οι υπόλοιπες 375 εικόνες δεν περιλαμβάνουν labels και επομένως δεν μπορούμε να τις χρησιμοποιήσουμε άμεσα για ανάλυση που αφορά την κατηγοριοποίηση ή την εκπαίδευση.

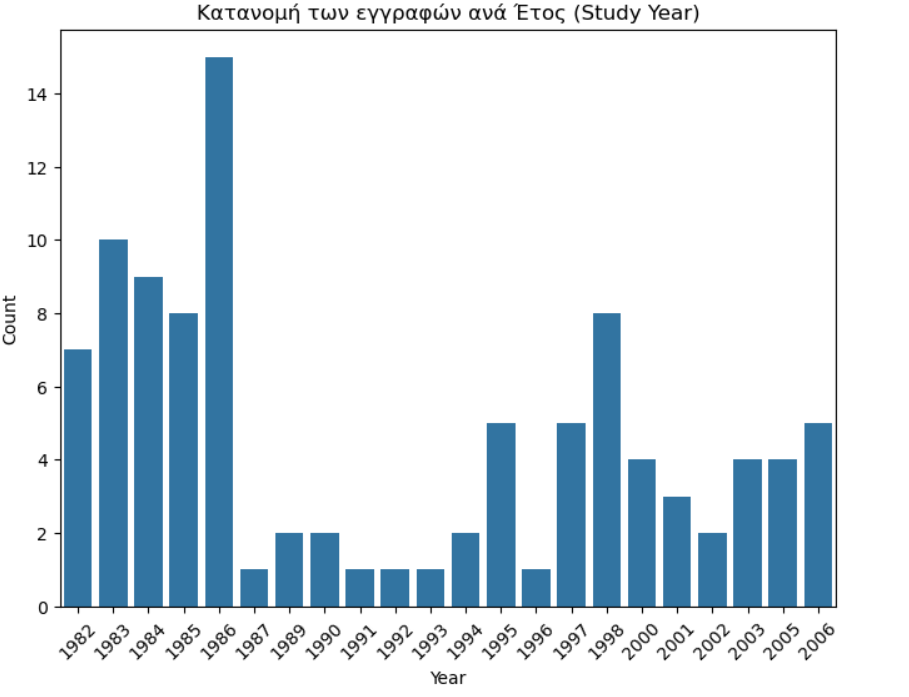
Η ανάλυση των 100 εικόνων μας βοήθησε να κατανοήσουμε καλύτερα τη δομή και την κατανομή των δεδομένων. Εντοπίσαμε πιθανές ανισορροπίες (όπως ανισόρροπη κατανομή στις κλάσεις στόχους) και αναγνωρίσαμε ακραίες τιμές ή σφάλματα στα δεδομένα, ώστε να προχωρήσουμε στη σωστή προεπεξεργασία και καθαρισμό τους.

Πιο συγκεκριμένα:

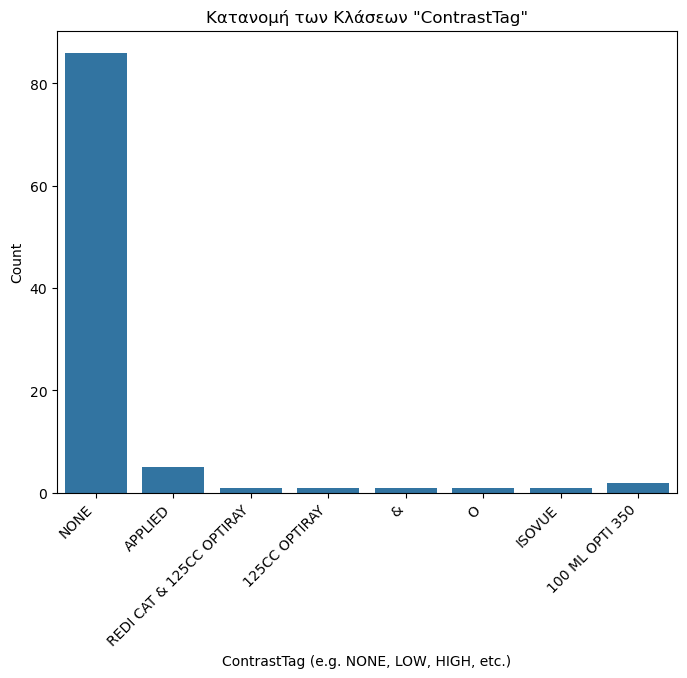
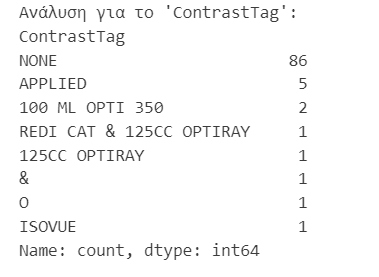
Για το χαρακτηριστικό "Ηλικία" (Age), ο αριθμός των παρατηρήσεων είναι 100, με τη μέση ηλικία των ατόμων να είναι 67.48 έτη και τυπική απόκλιση 9.27. Η ηλικία κυμαίνεται από 39 έως 83 έτη, με το 25ο εκατοστημόριο να είναι 61 έτη, το 50ο (διάμεσος) να είναι 70 έτη και το 75ο εκατοστημόριο να είναι 74 έτη. Η πιο νεαρή ηλικία είναι 39 και η μεγαλύτερη είναι 83 έτη.

Για την "Ημερομηνία Μελέτης" (Study Date), παρατηρούμε ότι οι ημερομηνίες κυμαίνονται από το 1982 έως το 2006.

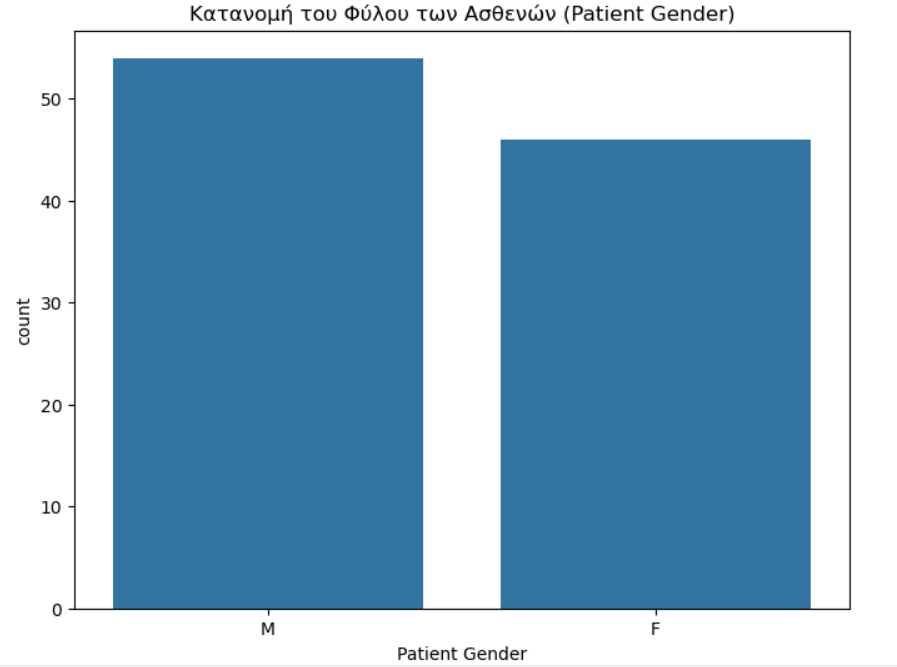


Για τις κατηγορικές στήλες, παρατηρήσαμε ότι το χαρακτηριστικό "ContrastTag" περιέχει 8 διαφορετικές τιμές, με τη μεγαλύτερη συχνότητα να είναι το "NONE" (86 εμφανίσεις). Οι υπόλοιπες τιμές εμφανίζονται λιγότερο συχνά, με την τιμή "APPLIED" να έχει 5 εμφανίσεις και κάποιες άλλες τιμές να εμφανίζονται μόνο 1 ή 2 φορές.

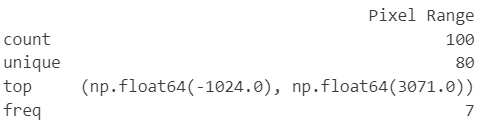


Επειδή το dataset είναι μικρό, δεν θα αφαιρέσουμε τις σπάνιες κατηγορίες, καθώς αυτές ενδέχεται να παρέχουν χρήσιμη πληροφορία. Αντίθετα, για να χειριστούμε την ανισορροπία στις κατηγορίες, θα χρησιμοποιήσουμε τεχνικές κατά τη διάρκεια της εκπαίδευσης του μοντέλου, όπως oversampling (π.χ., τεχνικές όπως SMOTE για τη δημιουργία νέων παρατηρήσεων από τις λιγότερο αντιπροσωπευτικές κατηγορίες), ή undersampling (μείωση του μεγέθους των κατηγοριών με τις περισσότερες εμφανίσεις), ή βάρη για τις κατηγορίες (δημιουργία βαρών για τις κατηγορίες, έτσι ώστε οι λιγότερο συχνές κατηγορίες να λάβουν περισσότερη σημασία κατά την εκπαίδευση του μοντέλου).

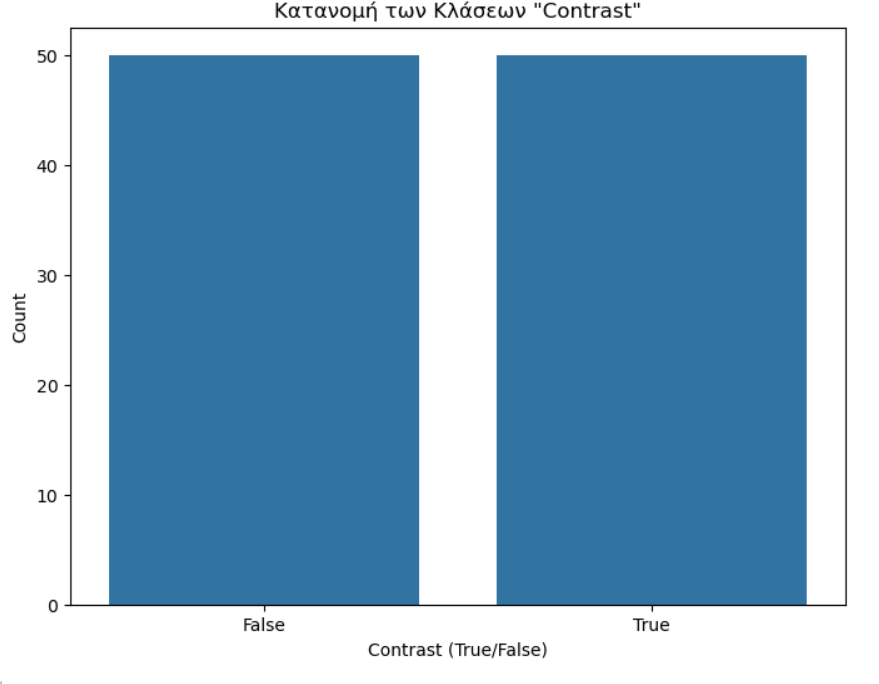
Η στήλη "Image Size\_x" περιέχει μόνο μία τιμή, δηλαδή τις διαστάσεις (512, 512) για όλες τις εικόνες. Όσον αφορά το "Image Type", όλες οι εικόνες είναι τύπου "Grayscale". Για το χαρακτηριστικό "Modality", όλες οι εικόνες προέρχονται από αξονική τομογραφία (CT). Το χαρακτηριστικό "Patient Gender" είναι σχεδόν ισομερές, με 54 άνδρες (M) και 46 γυναίκες (F).



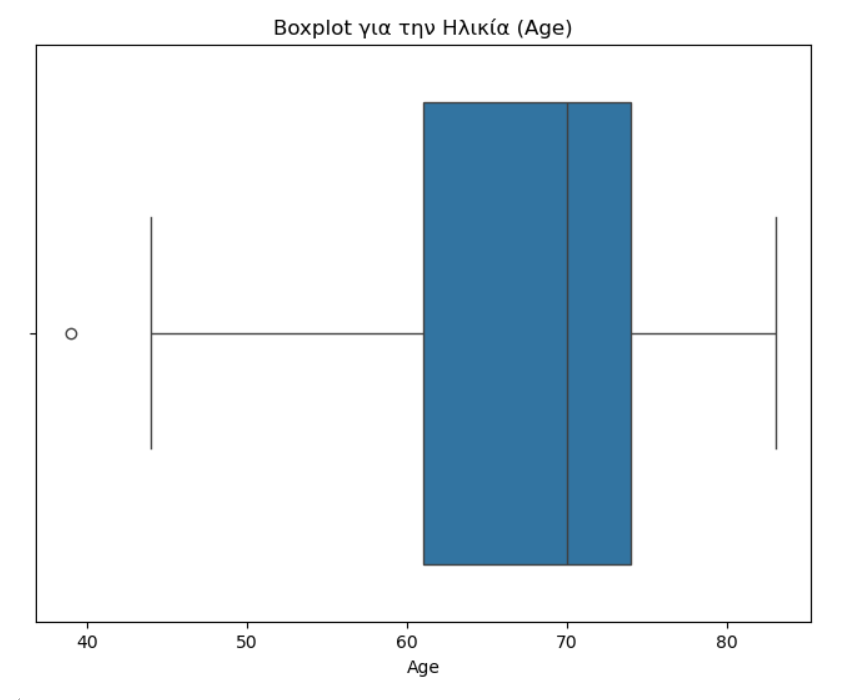
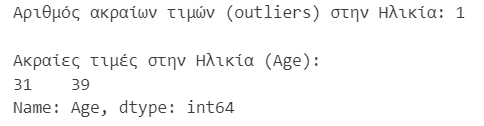
Στην στήλη "Pixel Range", παρατηρούμε διάφορες τιμές, με την πιο συχνή να είναι το (np.float64(-1024.0), np.float64(3071.0)) που εμφανίζεται 7 φορές. Αυτή η ποικιλία των τιμών στις κλίμακες των pixel υποδηλώνει ότι θα χρειαστεί κανονικοποίηση των τιμών των pixels πριν από οποιαδήποτε ανάλυση ή εκπαίδευση μοντέλων, προκειμένου να μειωθούν οι πιθανές διακυμάνσεις και να βελτιωθεί η απόδοση των αλγορίθμων μηχανικής μάθησης.

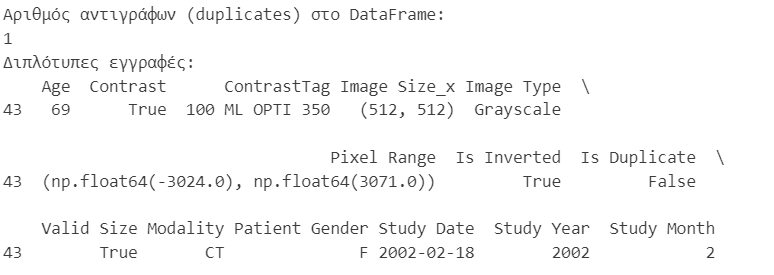


Όσον αφορά την κατανομή των κλάσεων στόχου για το χαρακτηριστικό "Contrast", οι 50 εικόνες έχουν την τιμή "True" και οι 50 την τιμή "False". Δηλαδή, οι κλάσεις είναι ισοκατανεμημένες, πράγμα που σημαίνει ότι δεν υπάρχει ανισορροπία στις κλάσεις.

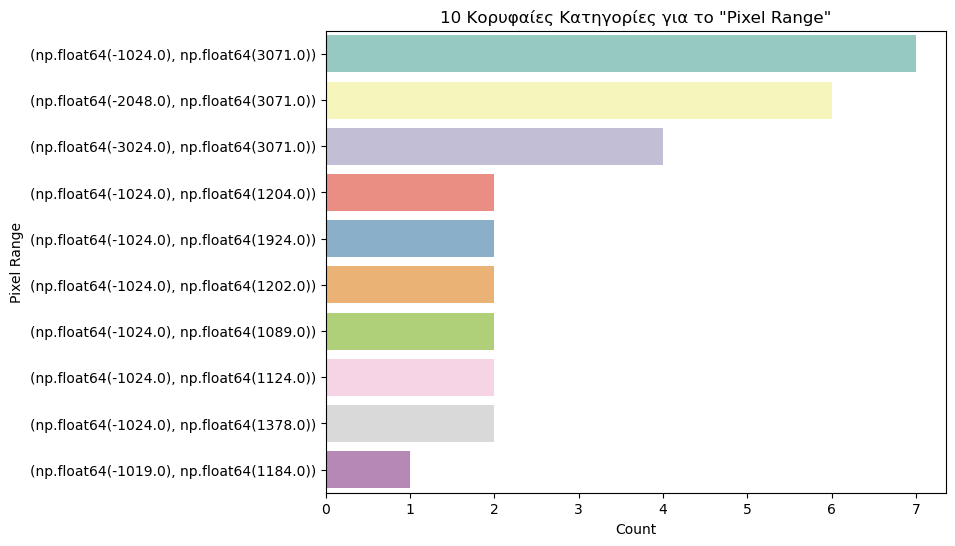


Επιπλέον, εντοπίσαμε μια ακραία τιμή στην "Ηλικία" (Age). Η τιμή 39 θεωρείται ακραία, καθώς είναι η μικρότερη καταγεγραμμένη τιμή ηλικίας. Η αντιμετώπιση αυτής της ακραίας τιμής θα πρέπει να εξεταστεί για να αποφευχθούν σφάλματα στην ανάλυση. Αν και αυτή η τιμή θεωρείται ακραία, επιλέγουμε να τη διατηρήσουμε στο dataset, καθώς το σύνολο δεδομένων είναι μικρό (μόνο 100 παρατηρήσεις), και η αφαίρεση αυτής της τιμής θα μπορούσε να προκαλέσει απώλεια σημαντικών πληροφοριών. Θα την αντικαταστήσουμε με τη μέση τιμή της ηλικίας, προκειμένου να εξισορροπήσουμε τα δεδομένα και να μειώσουμε την επίδραση αυτής της ακραίας τιμής στην ανάλυση.

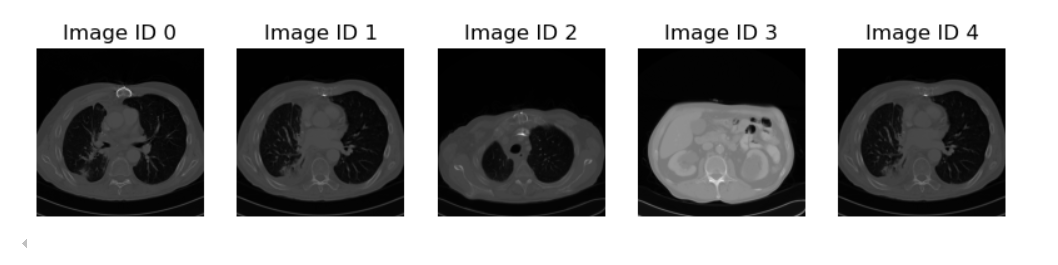
 

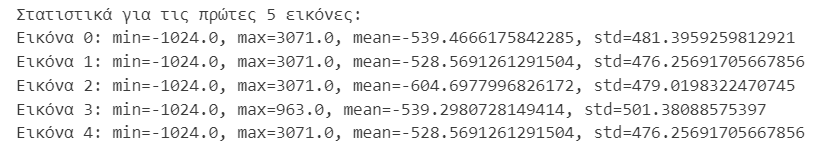
Ένα άλλο σημείο που παρατηρήθηκε είναι ότι υπάρχει 1 διπλότυπη εγγραφή στο σύνολο των δεδομένων. Η ύπαρξή της ενδέχεται να επηρεάσει τα αποτελέσματα της ανάλυσης, καθώς τα διπλότυπα δεδομένα μπορεί να προκαλέσουν παραμόρφωση στην κατανομή ή τη μέση τιμή των χαρακτηριστικών. Ωστόσο, δεδομένου ότι το dataset είναι ήδη μικρό και αποτελείται από μόνο 100 εικόνες, αποφασίσαμε να μην αφαιρέσουμε τη διπλότυπη εγγραφή. Η αφαίρεση θα είχε ως αποτέλεσμα τη μείωση του μεγέθους του dataset και ενδεχομένως την απώλεια πολύτιμων πληροφοριών, κάτι που δεν είναι επιθυμητό σε αυτήν τη φάση της ανάλυσης.

Η ανάλυση των ποσοστών για το χαρακτηριστικό Pixel Range αποκαλύπτει ότι υπάρχουν αρκετές διαφορετικές τιμές για αυτό το χαρακτηριστικό, οι οποίες αντιστοιχούν σε διαφορετικά εύρη τιμών για τα pixels. Οι τιμές τέτοιου τύπου χρησιμοποιούνται σε ιατρικές εικόνες (όπως οι εικόνες CT) και είναι του τύπου δεδομένων NumPy. Οι τιμές αυτές περιλαμβάνουν τα εξής εύρη: Το πιο συχνό εύρος είναι το (np.float64(-1024.0), np.float64(3071.0)), το οποίο εμφανίζεται στο 7% των καταχωρήσεων. Ακολουθούν και άλλα εύρη με λιγότερη συχνότητα εμφάνισης, όπως το (np.float64(-2048.0), np.float64(3071.0)), με ποσοστό 6%, και το (np.float64(-3024.0), np.float64(3071.0)), με ποσοστό 4%. Παρατηρείται ότι υπάρχουν πολλές διαφορετικές τιμές για το Pixel Range, αλλά οι συχνότερες εμφανίσεις έχουν να κάνουν με το εύρος των τιμών από -1024 έως 3071, το οποίο μπορεί να σχετίζεται με τις διαφορές στην επεξεργασία ή αποθήκευση των εικόνων.



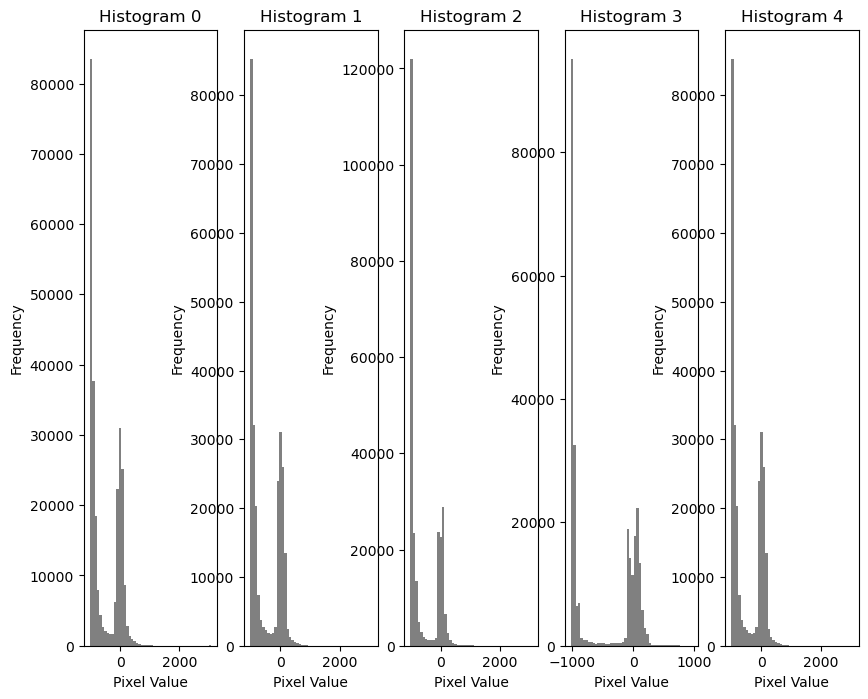
Στην εξερευνητική ανάλυση δεδομένων (EDA) για τις 100 εικόνες (για τις οποίες έχουμε metadata), αρχικά εμφανίσαμε τις 5 πρώτες για καλύτερη κατανόηση του θέματος και στατιστικά τους.





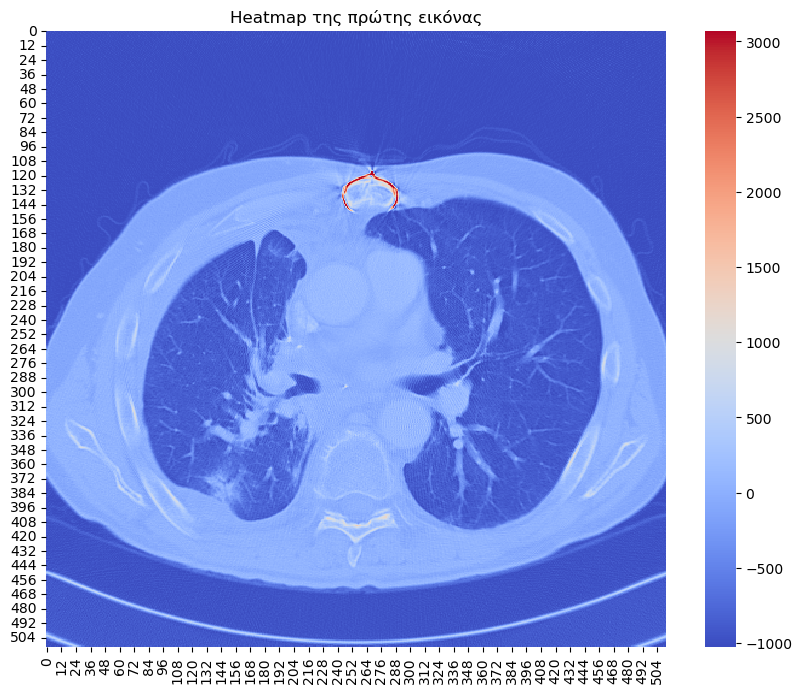
Το πρώτο που παρατηρούμε είναι ότι οι τιμές των pixel κυμαίνονται σε ένα ευρύ εύρος, με αρνητικές τιμές να είναι αρκετά συχνές. Συγκεκριμένα, για τις πρώτες 5 εικόνες, οι ελάχιστες τιμές των pixel είναι -1024 και οι μέγιστες τιμές φτάνουν τα 3071, ενώ η μέση τιμή και η τυπική απόκλιση διαφέρουν για κάθε εικόνα. Αυτό υποδεικνύει ότι οι εικόνες δεν έχουν κανονικοποιηθεί σωστά, κάτι που θα μπορούσε να προκαλέσει προβλήματα κατά την εκπαίδευση μοντέλων μηχανικής μάθησης. Η κανονικοποίηση των τιμών των pixel είναι επομένως απαραίτητη, προκειμένου να μετατραπούν οι τιμές σε ένα πιο διαχειρίσιμο εύρος, όπως [0, 255] ή [0, 1].

Τα ιστογράμματα των pixel για τις πρώτες 5 εικόνες δείχνουν ότι οι τιμές των pixel είναι διασκορπισμένες σε όλο το εύρος, με τις περισσότερες εικόνες να περιέχουν αρκετές αρνητικές τιμές. Αυτό ενισχύει την ανάγκη για κανονικοποίηση, καθώς οι αρνητικές τιμές ενδέχεται να δημιουργήσουν δυσκολίες στη διαδικασία εκμάθησης. Παρά τη διακύμανση των τιμών, η πυκνότητα των pixel φαίνεται να είναι σχετικά ομοιόμορφη για τις πρώτες 5 εικόνες, κάτι που υποδηλώνει ότι οι εικόνες ενδέχεται να έχουν παρόμοια χαρακτηριστικά όσον αφορά τις τιμές των pixel τους.

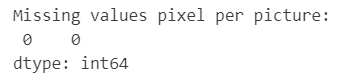


Αναφορικά με τα outliers, παρατηρήσαμε ότι ο αριθμός των ακραίων τιμών διαφέρει σημαντικά μεταξύ των εικόνων. Για παράδειγμα, η πρώτη εικόνα έχει 190 outliers, ενώ άλλες εικόνες (όπως η τέταρτη) δεν έχουν καθόλου. Αυτή η ποικιλία στις ακραίες τιμές υποδεικνύει ότι οι εικόνες δεν είναι ομοιογενείς και ότι σε μερικές περιπτώσεις η παρουσία outliers μπορεί να επηρεάζει την ποιότητα των δεδομένων. Αυτός ο αριθμός των outliers θα μπορούσε να μειωθεί με τη χρήση τεχνικών επεξεργασίας, όπως η απομάκρυνση ή η αντικατάσταση των outliers με πιο αντιπροσωπευτικές τιμές. Ένα τμήμα τους παρουσιάζεται εδώ: 

Ένα ενδιαφέρον εργαλείο που χρησιμοποιήσαμε για την ανάλυση ήταν η απεικόνιση της πρώτης εικόνας μέσω ενός heatmap. Ο χάρτης θερμότητας για την πρώτη εικόνα παρουσιάζει κυρίως μπλε περιοχές, που αντιστοιχούν σε χαμηλές τιμές pixel, ενώ μια μικρή περιοχή στην κορυφή είναι κόκκινη, υποδεικνύοντας υψηλότερες τιμές pixel. Αυτή η κόκκινη περιοχή μπορεί να αναπαριστά μια περιοχή υψηλής έντασης ή ιδιαίτερης σημασίας στην εικόνα, κάτι που θα μπορούσε να είναι χρήσιμο σε εφαρμογές ανάλυσης εικόνας ή ανίχνευσης χαρακτηριστικών.



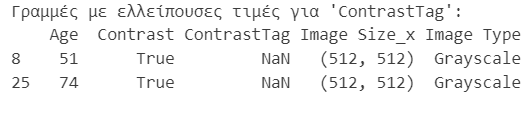
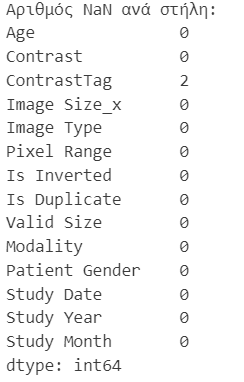
Κατά τον έλεγχο των ελλειπουσών τιμών στα pixels των εικόνων, διαπιστώθηκε ότι καμία από τις εικόνες δεν περιέχει ελλείπουσες τιμές στα pixels της.

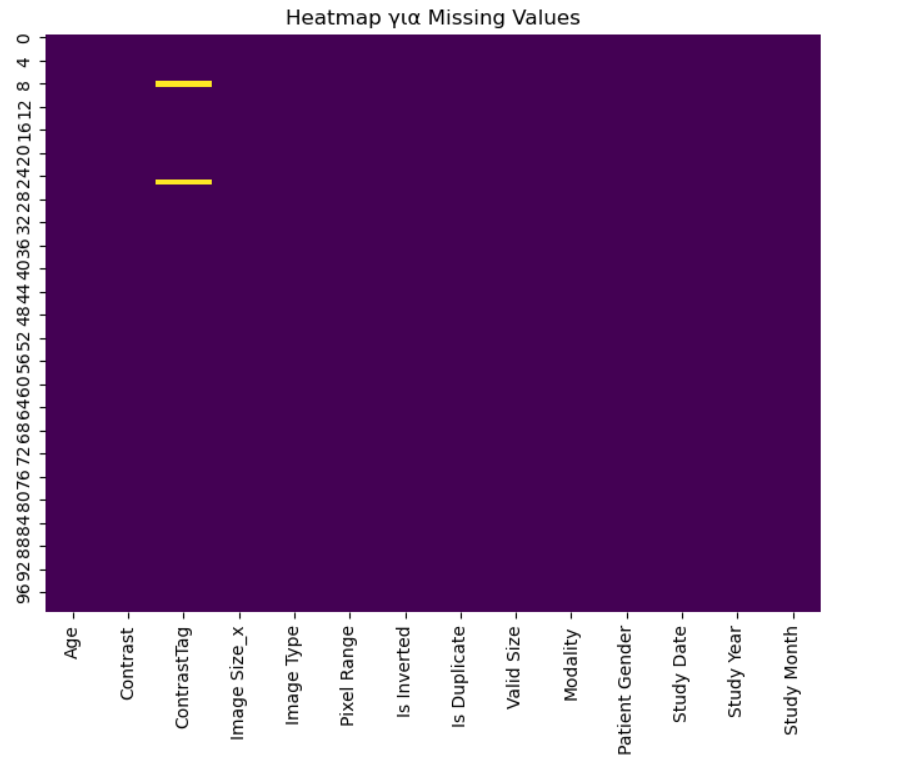


**2. Προεπεξεργασία Δεδομένων**

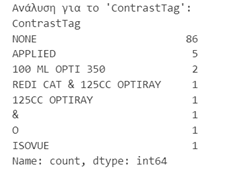
Για τον έλεγχο των ελλειπουσών τιμών στη στήλη ContrastTag, φορτώσαμε το dataset και εντοπίσαμε τις γραμμές που περιέχουν ελλείπουσες τιμές.

Παρατηρούμε ότι μόνο η στήλη "ContrastTag" περιέχει 2 ελλείπουσες τιμές. Αυτό σημαίνει ότι σε δύο εγγραφές δεν έχει καταχωρηθεί η τιμή για το χαρακτηριστικό "ContrastTag".





Για τη στήλη ContrastTag, όπως έχουμε ήδη αναφέρει, με μεγαλύτερη συχνότητα είναι το "NONE" (86 εμφανίσεις). Οι υπόλοιπες τιμές εμφανίζονται λιγότερο συχνά, με την τιμή "APPLIED" να έχει 5 εμφανίσεις και κάποιες άλλες τιμές να εμφανίζονται μόνο 1 ή 2 φορές



Αυτό μας δείχνει ότι οι ελλείψεις στη στήλη ContrastTag είναι αρκετά σπάνιες και δεν ακολουθούν κάποιο σαφές μοτίβο.

Με σκοπό να διερευνήσουμε αν οι ελλείψεις τιμών στη στήλη ContrastTag σχετίζονται με άλλες αριθμητικές στήλες του dataset, υπολογίσαμε τη συσχέτιση των missing τιμών αυτής της στήλης με τις αριθμητικές στήλες (π.χ., Age, Study Date). Τα αποτελέσματα της ανάλυσης δείχνουν ότι δεν παρατηρείται καμία ουσιαστική εξάρτηση των missing τιμών από άλλες παραμέτρους του dataset, όπως η ηλικία ή η ημερομηνία μελέτης.

Πιο συγκεκριμένα, η συσχέτιση των missing τιμών της στήλης ContrastTag με τις αριθμητικές στήλες Age και Study Date είναι 1.0. Ωστόσο, αυτή η υψηλή συσχέτιση δεν υποδηλώνει πραγματική εξάρτηση μεταξύ των missing τιμών και αυτών των παραμέτρων. Στην πραγματικότητα, η τιμή 1.0 προκύπτει λόγω του τρόπου με τον οποίο οι missing τιμές (NaN) υπολογίζονται στην ανάλυση δεδομένων. Συγκεκριμένα, η διαδικασία υπολογισμού της συσχέτισης αντιμετωπίζει τις missing τιμές με τον ίδιο τρόπο που υπολογίζονται οι άλλες τιμές, και επομένως το αποτέλεσμα είναι 1.0, κάτι που δεν αντικατοπτρίζει πραγματική σχέση μεταξύ των μεταβλητών.

Αυτό σημαίνει ότι η εμφάνιση των missing τιμών στη στήλη ContrastTag είναι τυχαία και ανεξάρτητη από τις υπόλοιπες τιμές του dataset, καταλήγοντας στο συμπέρασμα ότι οι missing τιμές ακολουθούν το πρότυπο των Missing Completely at Random (MCAR). Δηλαδή, οι missing τιμές δεν εξαρτώνται από τις άλλες τιμές ή χαρακτηριστικά του dataset και η εμφάνισή τους είναι καθαρά τυχαία.

Με βάση τη φύση των δεδομένων μας και την ανάγκη για μια απλή και αποτελεσματική προσέγγιση για την αντικατάσταση των missing τιμών, επιλέγουμε τη μέθοδο της αντικατάστασης με την επικρατούσα τιμή (NONE). Η μέθοδος αντικατάστασης των missing τιμών με την επικρατούσα τιμή (mode) φαίνεται να έχει μικρή επίδραση στην κατανομή των δεδομένων, αφού μόνο η κατηγορία NONE επηρεάστηκε, και μάλιστα με μικρή αύξηση. Η κατανομή των υπόλοιπων κατηγοριών παρέμεινε σταθερή. Συνολικά, η μέθοδος αυτή μπορεί να θεωρηθεί αποτελεσματική και δεν προκαλεί σημαντική αλλοίωση της αρχικής κατανομής των δεδομένων. Η αντικατάσταση αποθηκεύτηκε στο eda\_cleaned\_updated.csv

