

## Πρώτη Σειρά Ασκήσεων – Μέρος Β

Σε αυτή την άσκηση θα κάνετε διερευνητική ανάλυση (exploratory analysis) δεδομένων σχετικά με ταινίες. Ο στόχος είναι να κάνετε κάποιες μετρήσεις πάνω στα δεδομένα, να βρείτε ενδιαφέρουσες συσχετίσεις και να ερευνήσετε κάποιες υποθέσεις. Επίσης, να εξασκηθείτε με την χρήση των Pandas για ανάλυση δεδομένων.

Η άσκηση αποτελείται από βήματα. Ο στόχος είναι να τα υλοποιήσετε χρησιμοποιώντας κατά κύριο λόγο μεθόδους της βιβλιοθήκης Pandas (συν δικές σας συναρτήσεις που θα εφαρμόσετε με `apply`), και τις συναρτήσεις που έχουμε δει για plotting και στατιστική ανάλυση. Για κάποια βήματα ενδεχομένως να χρειαστείτε κάποιες επιπλέον βιβλιοθήκες ή εντολές από αυτές που έχετε δει στο φροντιστήριο. Μπορείτε να κάνετε αναζήτηση online ή στο Kaggle και να χρησιμοποιήσετε την βοήθεια από AI εργαλεία όπως το ChatGPT για να βρείτε τις κατάλληλες μεθόδους ή βιβλιοθήκες, αλλά όχι για να πάρετε έτοιμο κώδικα για όλο το ερώτημα. Τα συνήθη ερωτήματα θα είναι της μορφής: «Πως μπορώ με τη βιβλιοθήκη Pandas να κάνω το Χ?». Στα παραδοτέα (δείτε στο τέλος) θα πρέπει να παραδώσετε και ένα αρχείο με τα ερωτήματα που κάνατε και τις απαντήσεις που πήρατε, καθώς και άλλες πηγές κώδικα που χρησιμοποιήσατε.

Για την άσκηση θα χρησιμοποιήσετε δεδομένα από το Kaggle, τα οποία μπορείτε να κατεβάσετε από [εδώ](#). Τα δεδομένα αφορούν ταινίες και έχουν συγκεντρωθεί από το IMDB και το TMDb (The Movie DataBase). Από τα διαφορετικά αρχεία, στην άσκηση αυτή θα χρησιμοποιήσουμε στα αρχεία **movies\_metadata.csv** και **credits.csv**. Μπορείτε να δείτε τα πεδία των αρχείων στο Kaggle. Το κλειδί και για τα δύο αρχεία είναι το `id` το οποίο μοναδικό για κάθε ταινία. Αν θέλετε μπορείτε να χρησιμοποιήσετε τα άλλα αρχεία για το Βήμα 8.

**Βήμα 1.** Φορτώστε πρώτα το αρχείο με τα metadata σε Dataframe. Για τα βήματα 1-7 θα χρειαστείτε τα πεδία: `id`, `title`, `vote_counts`, `runtime`, `vote_average`, `revenue`, `budget`, `release_date`, `genres`. Τα δεδομένα χρειάζονται καθάρισμα, η στήλη με την περιγραφή περιέχει χαρακτήρες αλλαγής γραμμής που προκαλούν πρόβλημα στην φόρτωση του αρχείου (αναφέρεται και στο discussion), και κάποιες στήλες έχουν λάθος τιμές. Βεβαιωθείτε ότι οι στήλες που μας ενδιαφέρουν είναι σωστές (π.χ., έχουν αριθμητικές τιμές). Πετάξτε ότι δεν είναι στη σωστή μορφή, καθώς και γραμμές με παράλογες τιμές. Εκτυπώστε στο τέλος το μέγεθος των δεδομένων.

Χρησιμοποιήστε λειτουργίες της βιβλιοθήκης Pandas για την επεξεργασία και το καθάρισμα των δεδομένων.

**Βήμα 2.** Στο κομμάτι αυτό μας ενδιαφέρει να καταλάβουμε την κατανομή που ακολουθεί ο αριθμός των ψήφων (`vote_counts`). Κάνετε τις εξής γραφικές παραστάσεις (plots):

1. Ένα ιστόγραμμα του `vote_counts` με 100 κάδους (bins) χρησιμοποιώντας έτοιμη συνάρτηση της βιβλιοθήκης Pandas
2. Δημιουργείτε εσείς ένα ιστόγραμμα με κάδους που το μέγεθος τους αυξάνεται εκθετικά (διπλασιάζονται). Ο πρώτος κάδος θα έχει μέγεθος 100 (αριθμός ψήφων μεταξύ [1,100]), ο δεύτερος 200 (αριθμός ψήφων μεταξύ [100,300]), κ.ο.κ. Κάνετε plot σε λογαριθμική κλίμακα και στους δύο

άξονες. Εναλλακτικά, μπορείτε να κάνετε point plot τον αριθμό των εγγραφών σε κάθε κάδο ως προς το δεξί άκρο του κάδου, σε λογαριθμική κλίμακα και στους δύο άξονες.

- Υπολογίστε το cumulative frequency vector. Το διάνυσμα, για κάθε αριθμό από ψήφους (από το μικρότερο προς το μεγαλύτερο) κρατάει τον αριθμό των εγγραφών που έχουν **τουλάχιστον** τόσες ψήφους (π.χ., για το μικρότερο αριθμό είναι όλες οι εγγραφές). Κάνετε μια γραφική παράσταση του cumulative frequency ως συνάρτηση του αριθμού των ψήφων σε λογαριθμική κλίμακα και στους δύο άξονες.
- Κάνετε το Zipf plot για το vote\_counts. Το Zipf plot κατασκευάζεται έχοντας στον Y άξονα τις τιμές (τον αριθμό των ψήφων στην περίπτωση μας) και στο X την τάξη (rank) των τιμών. Για παράδειγμα ο μέγιστος αριθμός έχει rank 1, ο δεύτερος μεγαλύτερο 2, κ.ο.κ. Το plot θα είναι σε λογαριθμική κλίμακα και τους δύο άξονες.
- Δημιουργείτε μία νέα στήλη log\_votes με το **λογάριθμο** του vote\_counts και κάνετε ένα ιστόγραμμα με 100 κάδους χρησιμοποιώντας πάλι την μέθοδο της βιβλιοθήκης Pandas

Παρουσιάστε τα γραφήματα σας σε ένα grid και **σχολιάστε την κατανομή**.

Σημείωση: Για τα βήματα 2-4 μπορείτε αν θέλετε να τα υλοποιήσετε μεταφέροντας τα δεδομένα σε λίστες.

**Βήμα 3.** Στο βήμα αυτό θα εξετάσετε αν υπάρχει συσχέτιση μεταξύ της διάρκειας της ταινίας (runtime) και τριών άλλων πεδίων: vote\_average, budget, και revenue. Η υπόθεση που θέλουμε να εξετάσουμε είναι αν οι ταινίες με μεγάλη διάρκεια έχουν μεγαλύτερο προϋπολογισμό, αλλά είναι λιγότερο δημοφιλείς (μικρότερο vote\_average και revenue). Για το βήμα αυτό θα κρατήσετε ταινίες που έχουν runtime > 30, vote\_count > 10 budget>20000, και revenue > 20000.

Κάνετε scatter plots των τριών αυτών πεδίων και του runtime, για να οπτικοποιήσετε την συσχέτιση τους. Παρουσιάστε τα γραφήματα σας σε ένα grid 1 × 3. Μετά για κάθε ζευγάρι (runtime και ένα άλλο πεδίο) υπολογίστε το Pearson Correlation Coefficient και το αντίστοιχο p-value. Σχολιάστε τα αποτελέσματα.

**Βήμα 4.** Πολύ κόσμος υποστηρίζει ότι η ποιότητα των ταινιών χειροτερεύει με το χρόνο και οι παλιές ταινίες ήταν καλύτερες. Εξετάστε με δεδομένα αυτή την υπόθεση. Χρησιμοποιείτε το πεδίο release\_date για να πάρετε την δεκαετία στην οποία βγήκε η ταινία, και το vote\_average για να πάρετε την ποιότητα της ταινίας. Αγνοήστε ταινίες με λιγότερα από 10 votes, και δεκαετίες με λιγότερες από 50 ταινίες. Χρησιμοποιείτε την συνάρτηση lineplot της seaborn και δημιουργείτε ένα γράφημα με την μέση ποιότητα των ταινιών ανά δεκαετία και το 95% confidence interval. Τι παρατηρείτε?

**Bonus:** Μια ένσταση στην παραπάνω μέτρηση είναι ότι στα δεδομένα έχουμε λιγότερες ταινίες από παλαιότερες δεκαετίες, και ενδεχομένως αυτές είναι οι καλύτερες και πιο δημοφιλείς της δεκαετίας. Προτείνετε και υλοποιείτε ένα διαφορετικό τρόπο να κατασκευάσετε το παραπάνω γράφημα το οποίο θα μπορούσε να αντιμετωπίσει αυτή την ένσταση.

**Βήμα 5.** Στο βήμα αυτό θα εξετάσετε αν υπάρχει συσχέτιση μεταξύ του φύλου του σκηνοθέτη και του genre των ταινιών τις οποίες γυρίζει.

Για το βήμα αυτό θα πρέπει να εξάγετε το genre των ταινιών. Μπορείτε να πάρετε τα genres μιας ταινίας (συνήθως είναι πάνω από ένα) από το attribute genres. Δημιουργήστε ένα Dataframe με ζευγάρια το id της ταινίας και το genre. Μια ταινία μπορεί να εμφανίζεται πολλές φορές. Τυπώστε στατιστικά για τα διαφορετικά genres, και κρατήστε μόνο τα genres που έχουν περισσότερες από 100 ταινίες. (Υπόδειξη: Μπορεί να σας φανεί χρήσιμη η εντολή explode).

Στη συνέχεια από το αρχείο credits.csv εξάγετε για κάθε ταινία τον σκηνοθέτη ('job' : 'Director'), και το φύλο του σκηνοθέτη. Το 1 αντιστοιχεί σε γυναίκες, το 2 σε άντρες και το 0 σημαίνει unknown. Απορρίψτε τις ταινίες που δεν μπορείτε να προσδιορίσετε τον σκηνοθέτη ή το φύλο του. Τυπώστε στατιστικά για τις διαφορετικές τιμές, και κρατήστε μόνο τα gender 1 και 2.

Κάνετε merge τα dataframes με τα genres και τα φύλα, και δημιουργήστε το contingency table μεταξύ του φύλου και του genre. Μπορείτε να το αναπαραστήσετε με ένα heatmap, αλλά θα πρέπει να φαίνονται και οι τιμές. Κάνετε το  $\chi^2$ -test για να διαπιστώσετε αν υπάρχει κάποια στατιστικά σημαντική συσχέτιση. Αξιολογήστε τα αποτελέσματα κοιτώντας τα p-value. Κοιτάξτε τα ζευγάρια από τιμές με το μεγαλύτερο και το μικρότερο lift (λόγος της τιμής του ζευγαριού προς την αναμενόμενη τιμή – σας την δίνει το  $\chi^2$ -τεστ). Σχολιάστε τα αποτελέσματα.

**Βήμα 6.** Στο κομμάτι αυτό θα εξετάσετε αν υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές του revenue, budget και vote\_average μεταξύ των genres 'Action' και 'Drama'. Για κάθε attribute δημιουργήστε ένα bar plot που θα έχετε την μέση τιμή για κάθε genre και το 95% confidence interval. Παρουσιάστε τα γραφήματα σε ένα 1 × 3 grid. Προσπαθήστε να εξάγετε κάποιο συμπέρασμα αρχικά από το plot. Στη συνέχεια για κάθε attribute χρησιμοποιήστε το t-test για να εξετάσετε αν υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές μεταξύ των διαφορετικών genres. Περιγράψτε τα αποτελέσματα και τα συμπεράσματα σας.

**Βήμα 7.** “Οι γυναίκες σκηνοθέτες έχουν μικρότερο προϋπολογισμό από τους άνδρες σκηνοθέτες”. Χρησιμοποιήστε τα δεδομένα και στατιστικά τεστ για να εξερευνήσετε την παραπάνω υπόθεση.

**Βήμα 8.** Διατυπώστε μια δική σας υπόθεση και εξετάστε την χρησιμοποιώντας τα δεδομένα. Η υπόθεση σας θα πρέπει να είναι κάτι μη τετριμμένο και να την εξετάσετε χρησιμοποιώντας (και) κάποιο στατιστικό τεστ.

**Παράδοση:** Παραδώστε ένα Notebook το οποίο θα περιέχει τον κώδικα για την επεξεργασία των δεδομένων, τις γραφικές παραστάσεις και τους υπολογισμούς που κάνατε, καθώς και τις παρατηρήσεις και τα συμπεράσματα σας. Βάλτε headers ώστε να ξεχωρίζουν τα διαφορετικά κομμάτια της άσκησης. Μαζί με το Notebook παραδώστε και ένα word ή txt αρχείο στο οποίο θα έχετε τις αναζητήσεις που κάνατε online, τις πηγές που χρησιμοποιήσατε και την βοήθεια που πήρατε από AI εργαλεία.