Πρώτη Σειρά Ασκήσεων

Αυτή είναι η πρώτη σειρά ασκήσεων. Η προθεσμία για την παράδοση είναι στις 15 Δεκεμβρίου 11:59 μ.μ. Παραδώστε Notebooks με τον κώδικα και τις αναφορές. Για την ερώτηση 1, μπορείτε να παραδώσετε και pdf με την απόδειξη, ή φωτογραφίες από χειρόγραφα. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Η παράδοση θα γίνει μέσω του ecourse. Λεπτομέρειες στη σελίδα Ασκήσεις του μαθήματος. Η άσκηση είναι **ατομική**.

Ερώτηση 1

Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο Reservoir Sampling που περιγράψαμε στο μάθημα, ώστε να κάνει δειγματοληψία $\boldsymbol{K}$ αντικειμένων ομοιόμορφα τυχαία από ένα ρεύμα N αντικειμένων. Το κάθε αντικείμενο να έχει πιθανότητα K/N να εμφανιστεί στο δείγμα. Ο αλγόριθμος σας θα πρέπει να δουλεύει με ένα μόνο πέρασμα στα δεδομένα διαβάζοντας τα αντικείμενα ένα-ένα, χωρίς προηγούμενη γνώση του μεγέθους του ρεύματος (το μέγεθος N), και να χρησιμοποιεί $O(K)$ μνήμη (υποθέστε ότι το μέγεθος του κάθε αντικειμένου είναι σταθερό). Αυτό σημαίνει ότι δεν μπορείτε να αποθηκεύσετε όλο το ρεύμα των δεδομένων στη μνήμη.

1. Περιγράψετε τον αλγόριθμο που διαλέγει ένα ομοιόμορφο δείγμα K αντικειμένων από ένα ρεύμα N αντικειμένων. Η περιγραφή του αλγορίθμου **δεν** πρέπει να είναι σε κώδικα ή ψευδοκώδικα, ούτε η περιγραφή του κώδικα σε φυσική γλώσσα. Στην περίπτωση αυτή η απάντηση σας μηδενίζεται. Η περιγραφή θα πρέπει να εξηγεί τη λογική του αλγορίθμου σε απλά Ελληνικά.
   Για παράδειγμα, αυτή είναι μία περιγραφή σε απλά Ελληνικά του αλγορίθμου για τη δειγματοληψία ενός αντικειμένου που είδαμε στην τάξη:
   Ο αλγόριθμος κρατάει χώρο για ένα αντικείμενο και ένα μετρητή με τον αριθμό των αντικειμένων που έχει δει. Διατρέχει τα αντικείμενα ένα-ένα όπως έρχονται από το ρεύμα. Όταν βλέπει το n-οστό αντικείμενο, το επιλέγει με πιθανότητα $\frac{1}{n}$ και το αποθηκεύει, αντικαθιστώντας το υπάρχον αντικείμενο (αν δεν είναι το πρώτο). Ενημερώνει τον μετρητή. Όταν ολοκληρωθεί το ρεύμα, επιστρέφει το αντικείμενο που έχει αποθηκεύσει.
2. Αποδείξτε ότι ο αλγόριθμος σας παράγει ένα ομοιόμορφα τυχαίο δείγμα, δηλαδή, για κάθε $i, 1 \leq i \leq N$, το i -οστό στοιχείο έχει πιθανότητα K/N να εμφανιστεί στο δείγμα.
3. Γράψτε μία συνάρτηση **sample σε Python** που υλοποιεί τον αλγόριθμο σας. Η συνάρτηση σας θα πρέπει να παίρνει σαν όρισμα το όνομα ενός αρχείου, και τον αριθμό K και να επιστρέφει μια λίστα που κρατάει ένα δείγμα με K τυχαίες γραμμές από το αρχείο. Θα πρέπει να διαβάζετε το αρχείο γραμμή-γραμμή και να μην το φορτώσετε στη μνήμη. Χρησιμοποιήστε την συνάρτηση σας μέσα σε ένα πρόγραμμα για να πάρετε 10 τυχαίες γραμμές από το αρχείο *input.txt* που θα σας δίνεται. Εκτυπώστε τις γραμμές στο δείγμα.

Εξηγείστε την αντιστοίχιση μεταξύ της περιγραφής που δώσατε στο 1ο βήμα και του κώδικα σας. Για παράδειγμα, για τον αλγόριθμο που επιλέγει ένα μόνο αντικείμενο, θα μπορούσατε να γράψετε κάτι της μορφής: «*Στις γραμμές 2-3 γίνεται η επιλογή του n-οστού αντικειμένου με πιθανότητα* $\frac{1}{n}$».

Δημιουργείστε ένα Notebook με δύο κελιά κώδικα. Ένα με τα imports και τον ορισμό της συνάρτησης sample και ένα στο οποίο θα χρησιμοποιείτε την συνάρτηση σας στο αρχείο input.txt, θα αποθηκεύετε το αποτέλεσμα και θα το εκτυπώνετε. Μπορείτε να κατεβάσετε το αρχείο input.txt από την σελίδα Ασκήσεις. Προσθέστε και δύο κελιά κειμένου, ένα με την περιγραφή του αλγορίθμου (Βήμα 1), και ένα με την αντιστοίχιση μεταξύ κώδικά και περιγραφής. Μπορείτε να προσθέσετε την απόδειξη (Βήμα 2) σε ένα ξεχωριστό κελί, ή να την γράψετε ξεχωριστά και να παραδώσετε ένα pdf με το κείμενο (ή φωτογραφίες αν είναι χειρόγραφη). Παραδώστε το Notebook και το αρχείο input.txt, και το pdf (ή φωτογραφίες) με την απόδειξη αν υπάρχει.

Ερώτηση 2

Σας δίνεται το αρχείο "data.csv" το οποίο μπορείτε να κατεβάσετε από τη σελίδα Ασκήσεις. Το αρχείο έχει τρεις στήλες χωρισμένες με κόμμα, με ονόματα A, B, C, και 1000 γραμμές. Οι τιμές των B και C είναι συνάρτηση αυτών της A. Συγκεκριμένα, για κάθε τιμή x στη στήλη A, η αντίστοιχη τιμή στις στήλες B και C είναι $f_B(x) \cdot (1+\epsilon)$ και $f_C(x) \cdot (1+\epsilon)$ αντίστοιχα όπου ϵ είναι τυχαίος θόρυβος (διαφορετικός για κάθε x και για κάθε στήλη). Ο στόχος σας είναι να προσδιορίσετε τις συναρτήσεις f_B και f_C.

Για να προσδιορίσετε τις συναρτήσεις, φορτώστε τα δεδομένα σε ένα Pandas data frame και δημιουργήστε γραφικές παραστάσεις των B και C ως προς το A, όπως είδαμε την τάξη, καθώς και όποια άλλη γραφική παράσταση χρειάζεστε. Παραδώσετε ένα Notebook το οποίο θα περιέχει τον κώδικα για την επεξεργασία των δεδομένων, τις γραφικές παραστάσεις και τους υπολογισμούς που κάνατε, καθώς και μία αναφορά με τα συμπεράσματα σας.

Ερώτηση 3

Τα τελευταία χρόνια, η ανάλυση δεδομένων σε σπορ έχει γίνει ένα επιστημονικό πεδίο από μόνη της. Ένα άθλημα στο οποίο έχει μεγάλη εφαρμογή είναι το μπάσκετ. Στην ερώτηση αυτή θα μελετήσουμε δεδομένα από το NBA (National Basketball Association). Ο στόχος μας είναι να κάνουμε κάποιες παρατηρήσεις πάνω στα δεδομένα, να βρούμε ενδιαφέρουσες συσχετίσεις και να ερευνήσουμε κάποιες υποθέσεις. Επίσης, να εξασκηθείτε με την χρήση των Pandas για ανάλυση δεδομένων.

Μπορείτε να κατεβάσετε τα δεδομένα που θα χρησιμοποιήσουμε από τη σελίδα Ασκήσεις του μαθήματος. Έχουμε δύο συλλογές δεδομένων: Το dataset1 (πηγή: Kaggle) και το dataset2 (πηγή: Kaggle).

Το **dataset1** περιέχει τρία αρχεία. Μπορείτε να διαβάσετε περισσότερα γι αυτά στο Kaggle link. Εμείς θα χρησιμοποιήσουμε μόνο το αρχείο Seasons_Stats.csv. Περιέχει τα συγκεντρωτικά στατιστικά των παιχτών για διαφορετικές σαιζόν, ξεκινώντας από το 1950. Εδώ είναι ένα λεξικό για τα πεδία (στήλες) του αρχείου.

Για την άσκηση θα κάνετε την εξής προεπεξεργασία των δεδομένων:

1. Θα κρατήσετε τις χρονιές (Year) από το 1981 και μετά.
2. Θα κρατήσετε μόνο παίχτες (γραμμές) με ηλικία (Age) μεγαλύτερη από 18 και μικρότερη από 40.
3. Θα κρατήσετε μόνο τις σαιζόν ενός παίχτη (γραμμές) που έχει παίξει περισσότερα από 500 λεπτά (MP).

4. Θα κρατήσετε μόνο τις θέσεις (Pos) ‘PG’ (Point Guard – θέση 1), ‘SG’ (Scoring Guard – θέση 2), ‘SF’ (Small Forward – θέση 3), ‘PF’ (Power Forward – θέση 4), ‘C’ (Center – θέση 5). (Θα σας είναι χρήσιμη η μέθοδος isin για μια στήλη).

Δεν θα χρησιμοποιήσουμε όλες τις στήλες. Οι στήλες που θα χρησιμοποιήσουμε κατά κύριο λόγο είναι οι εξής: ‘G’ (Games), ‘PTS’ (Points), ‘AST’ (Assists), ‘TRB’ (Total Rebounds), ‘BLK’ (Blocks), ‘PER’ (Player Efficiency Rating – μια μετρική για τη γενική αξιολόγηση ενός παίχτη), ‘TS%’ (True Shooting Percentage). Μπορείτε να δείτε τον ορισμό αυτών των πεδίων στο λεξικό. Θα είναι χρήσιμο να δημιουργήσετε και κάποια δικά σας πεδία όπως για παράδειγμα τα στατιστικά ανά παιχνίδι (π.χ., πόντοι ανά παιχνίδι – Points Per Game). Όταν χρησιμοποιείτε κάποια πεδία βεβαιωθείτε ότι αφαιρέσατε τις null τιμές.

Για την προεπεξεργασία χρησιμοποιήστε μεθόδους της βιβλιοθήκης Pandas.

Το **dataset2** περιέχει πιο λεπτομερή δεδομένα για τις σαιζόν 2014-2021 (μέχρι πριν μερικές μέρες). Έχει δεδομένα για τις ομάδες (teams.csv), τους παίχτες (players.csv), τα παιχνίδια (games.csv), την τρέχουσα κατάταξη των ομάδων σε κάθε παιχνίδι (ranking.csv) και τα στατιστικά των παιχτών ανά παιχνίδι (games_details.csv). Τα στατιστικά είναι ένα υποσύνολο αυτών που εμφανίζονται στο dataset1. Θα χρησιμοποιήσουμε κατά κύριο λόγο τα αρχεία games.csv και games_details.csv.

Η άσκηση αποτελείται από τα παρακάτω κομμάτια. Ο στόχος είναι να υλοποιήσετε τα παρακάτω φορτώνοντας τα δεδομένα σε Pandas dataframes και χρησιμοποιώντας κατά κύριο λόγο μεθόδους της Pandas (συν δικές σας συναρτήσεις που θα εφαρμόσετε με apply).

Α. Στο κομμάτι αυτό μας ενδιαφέρει να καταλάβουμε την κατανομή που ακολουθεί ο συνολικός αριθμός πόντων που έχουν σκοράρει οι παίχτες. Θα κάνετε τα εξής γραφήματα (plots):

1. Ένα ιστόγραμμα των πόντων με 100 κάδους (bins) χρησιμοποιώντας έτοιμη συνάρτηση της βιβλιοθήκης Pandas
2. Ένα ιστόγραμμα του **λογαρίθμου** των πόντων με 100 κάδους χρησιμοποιώντας πάλι μεθόδους της βιβλιοθήκης Pandas
3. Ένα ιστόγραμμα των πόντων με 100 ισομεγέθεις κάδους που θα κατασκευάσετε εσείς. Στον Χ άξονα θα έχετε το κάτω άκρο του κάδου, και στον Υ τον αριθμό των σημείων που πέφτουν σε αυτό τον κάδο. Θα κάνετε plot το Υ ως προς το Χ παίρνοντας λογαριθμική κλίμακα και στους δύο άξονες.
4. Το Zipf plot της κατανομής. Το Zipf plot κατασκευάζεται έχοντας στον Υ άξονα τις τιμές (πόντους στην περίπτωση μας) και στο Χ την τάξη (rank) των τιμών. Για παράδειγμα ο μέγιστος αριθμός των πόντων έχει rank 1, ο δεύτερος μεγαλύτερος 2, κλπ. Το plot θα είναι σε λογαριθμική κλίμακα και τους δύο άξονες.

Παρουσιάστε τα γραφήματα σας σε ένα grid και σχολιάστε την κατανομή

Σημείωση: Δεν υπάρχει σαφές συμπέρασμα για την κατανομή που ακολουθούν οι πόντοι αλλά μπορείτε να κάνετε κάποιες παρατηρήσεις για το σχήμα της κατανομής. Μπορείτε επίσης να προσθέσετε κάποιο δικό σας plot αν πιστεύετε ότι θα σας βοηθήσει. Τα βήματα 3,4 είναι πιο δύσκολο να υλοποιηθούν χρησιμοποιώντας Pandas (ειδικά το Βήμα 3), μπορείτε αν θέλετε να τα υλοποιήσετε μεταφέροντας τα δεδομένα σε λίστες.

Β. Στο κομμάτι αυτό μας ενδιαφέρει η εξέλιξη των μετρικών στο χρόνο. Συγκεκριμένα θα εξετάσουμε πως εξελίσσεται η απόδοση των παιχτών με την ηλικία. Συγκεκριμένα από τα υπάρχοντα δεδομένα θα κοιτάξουμε τα πεδία PER και TS%, και θα υπολογίσετε επίσης τα πεδία Points Per Game (PPG), Assists Per Game (APG), Total Rebounds Per Game (RPG), και Blocks Per Game (BPG) ανά σαιζόν. Χρησιμοποιήστε το lineplot της βιβλιοθήκης seaborn για να κάνετε γραφήματα της μέσης τιμής αυτών των στατιστικών ως συνάρτηση της ηλικίας. Παρουσιάστε τα γραφήματα σέ ένα grid 2X3. Τι παρατηρείτε? Σε ποια ηλικία πιάνουν την κορυφή της απόδοσης τους οι παίχτες? Πως διαφέρει ανά στατιστικό? Σχολιάστε τα αποτελέσματα.

Στη συνέχεια θα εξετάσουμε αν η απόδοση των παιχτών με την ηλικία εξαρτάται από την θέση τους. Κάνετε τα ίδια plots ξεχωρίζοντας τα στατιστικά ανά θέση (χρησιμοποιήστε την παράμετρο hue της lineplot). Βλέπετε κάποια διαφορά ανάλογα με την θέση? Σχολιάστε τα αποτελέσματα.

Γ. Στο κομμάτι αυτό μας ενδιαφέρει να ερευνήσουμε αν υπάρχει συσχέτιση μεταξύ των βασικών στατιστικών των παιχτών. Υπολογίστε ξανά τα στατιστικά PPG, APG, RPG, BPG που υπολογίσατε στο Μέρος Β, αυτή τη φορά όχι ανά σαιζόν, αλλά χρησιμοποιώντας τα δεδομένα από όλες τις σαιζόν. Δημιουργήστε ένα γράφημα με όλα τα scatter plots των πεδίων ανά δύο (χρησιμοποιήστε το pairplot της seaborn), και δύο πίνακες 4X4 με τα Pearson correlation coefficients και τα αντίστοιχα p-values. (Εναλλακτικά μπορείτε να παρουσιάσετε τα αποτελέσματα χρησιμοποιώντας heatmaps με τις τιμές). Σχολιάστε τα αποτελέσματα. Παρατηρείτε κάποια ενδιαφέρουσα συσχέτιση? Μια συσχέτιση είναι ενδιαφέρουσα αν έχει μεγάλο coefficient και p-value μικρότερο του 0.05.

Bonus: Δημιουργήστε τα ίδια plots και μετρήσεις ανά θέση (χρησιμοποιείστε το hue για τα plots). Σχολιάστε τις διαφορές.

Δ. Στο κομμάτι αυτό μας ενδιαφέρει να μελετήσουμε αν υπάρχει διαφορά μεταξύ της απόδοσης παιχτών που παίζουν σε διαφορετική θέση. Χρησιμοποιήστε τα στατιστικά PPG, APG, RPG, BPG που υπολογίσατε στο Γ, και δημιουργήστε barplots με τις μέσες τιμές για τις διαφορετικές θέσεις, με 95%-confidence intervals (τοποθετήστε τα σε ένα grid 4 θέσεων). Μελετήστε τα οπτικά και αναφέρετε τα συμπεράσματα σας.

Στη συνέχεια κάνετε το ίδιο για το μέσο PER των παιχτών από όλες τις σαιζόν. Οι διαφορές δεν είναι πλέον τόσο ξεκάθαρες. Χρησιμοποιείστε το t-test για να αποφασίσετε ποιες διαφορές είναι στατιστικά σημαντικές. Δημιουργήστε ένα 5X5 πίνακα με τα p-values για όλους τους συνδυασμούς θέσεων, και σχολιάστε τι παρατηρείτε.

Ε. Ο Russell Westbrook είναι ένας παίχτης που διχάζει με τον τρόπο που παίζει. Παρότι έχει καταγράψει πολλαπλά triple-doubles (διψήφιο αριθμό από πόντους, ασίστς και ριμπάουντ σε ένα παιχνίδι), πολλοί πιστεύουν ότι τα επιτυγχάνει αυτά σε βάρος της ομάδας του. Υπολογίστε την δεσμευμένη πιθανότητα να κερδίσει η ομάδα του Westbrook όταν αυτός πετυχαίνει triple-double, και συγκρίνετε την με την πιθανότητα να κερδίσει η ομάδα του ασχέτως triple-double, όταν παίζει ο Westbrook. Χρησιμοποιήστε το χ^2-test για να εξετάσετε αν το να έχει triple-double ο Westbrook και το να κερδίσει η ομάδα του είναι ανεξάρτητα. Σχολιάστε τα αποτελέσματα σας.

Για το μέρος αυτό θα χρησιμοποιήσετε το dataset2. Το PLAYER_ID του Westbrook είναι 201566. Η πληροφορία για το ποια ομάδα νίκησε βρίσκεται στο games.csv, ενώ τα στατιστικά του Westbrook για κάθε παιχνίδι είναι στο games_details.csv.

Bonus: Οι haters του Westbrook θα πουν ότι πετυχαίνει τα triple-double απέναντι σε εύκολους αντιπάλους. Χρησιμοποιήστε τα δεδομένα για να εξετάσετε αυτή την υπόθεση.

Ζ. Στο κομμάτι αυτό θα θέσουμε θα κάνουμε κάποιες υποθέσεις και θα τις εξετάσουμε στα δεδομένα.

Υπόθεση 1: Υπάρχει συσχέτιση μεταξύ του αριθμού των assist που δίνουν οι περιφερειακοί παίχτες σε μια ομάδα (the backcourt - PG, SG) και των πόντων που σκοράρουν οι ψηλοί στην ομάδα (the frontcourt – SF, PF, C).

Χρησιμοποιήστε το dataset1 και το Pearson Correlation Coefficient για να εξετάσετε την υπόθεση.

Υπόθεση 2: Οι ομάδες σκοράρουν κατά μέσο όρο περισσότερους πόντους εντός έδρας (home) από ότι εκτός έδρας (away).

Για να εξετάσετε αυτή την υπόθεση χρησιμοποιήστε το dataset2. Θα εξετάσουμε δύο ομάδες: Τους Boston Celtics (BOS - TEAM_ID = 1610612738) και τους Minnesota Timberwolves (MIN – TEAM_ID = 1610612750). Θα κάνετε δύο πειράματα για να εξετάσετε την υπόθεση:

1. Στο πρώτο πείραμα θα χρησιμοποιήσετε το t-test για να εξετάσετε την υπόθεση.
2. Στο δεύτερο θα κάνετε ένα permutation test. Δοθείσας της ομάδας, για κάθε παιχνίδι της μπορούμε να υπολογίσουμε τους πόντους που σκόραρε, και μια ετικέτα «Εντός» (home) ή «Εκτός» (away). Μπορούμε τώρα να υπολογίσουμε την διαφορά μεταξύ του μέσου αριθμού πόντων εντός και εκτός έδρας. Αυτή είναι η παρατηρούμενη διαφορά, ή παρατηρούμενη τιμή. Στη συνέχεια, κρατήστε την στήλη με τους πόντους σταθερή, και ανακατέψτε τις τιμές στην στήλη με τις ετικέτες δημιουργώντας μια τυχαία αναδιάταξη. Υπολογίστε την διαφορά των μέσων τιμών όπως πριν. Επαναλάβετε αυτή τη διαδικασία 1000 φορές, και πάρετε 1000 διαφορές. Υπολογίστε το εμπειρικό p-value για την παρατηρούμενη τιμή (ποσοστό των 1000 πειραμάτων που έχουν διαφορά μεγαλύτερη ή ίση από παρατηρούμενη διαφορά), Η παρατηρούμενη διαφορά είναι στατιστικά σημαντική αν το εμπειρικό p-value μικρότερο από 0.05. Αναφέρετε τα αποτελέσματα σας για τις δύο ομάδες.
   Δημιουργήστε επίσης το ιστόγραμμα των 1000 τιμών που υπολογίσατε. Τοποθετήστε την παρατηρούμενη τιμή σαν μια κάθετη γραμμή στο ιστόγραμμα, ώστε να δείξετε και οπτικά το αποτέλεσμα του πειράματος σας (χρησιμοποιήστε τη μέθοδο axvline της pyplot για να τοποθετήσετε μια κάθετη γραμμή στο σημείο της παρατηρούμενης διαφοράς).

Η. Διατυπώστε μια δική σας υπόθεση και εξετάστε την χρησιμοποιώντας τα δεδομένα.

Παραδώσετε 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. Βάλετε headers ώστε να ξεχωρίζουν τα διαφορετικά κομμάτια της άσκησης.