Exam Y

Python Algorithms for Industrial Engineers

Hadas Lapid, PhD

סטודנטים.ות יקרים.ות,

המבחן עוסק בבסיס נתונים בשם "athletes_short.csv".

בסיס הנתונים מכיל סיכום נתוני תחרויות אתלטיקה מארבע מדינות.

בסיס הנתונים מכיל את העמודות הבאות:

Gender – משתנה קטגורי המיצג את מגדר הספורטאי

Age – משתנה מספרי המייצג גיל

Height - משתנה מספרי המייצג גובה

Weight - משתנה מספרי המייצג משקל

Team – משתנה קטגורי המייצג מדינה

Year – משתנה מספרי, המיצג את השנה

Medal – משתנה קטגורי המייצג זכיה במדליה.

המבחן כולל 5 שאלות, אותן יש להטמיע בפונקציות המתאימות בשלד באופן הבא:

יש לממש את הפונקציות **במקום** המילה pass, כולל הקלט והפלט של כל אחת ואחת מהן.

יש לבדוק את מימושן באמצעות שורות הבדיקה התואמות ב-main.

באם צלח המימוש, השאירו את הבדיקות עובדות. באם לא צלח המימוש, יש להסתירן בכדי לאפשר לקובץ לרוץ באופן תקין.

5 נקודות מתוך 100 שמורות לתקינות ריצת הפתרון. (לקובץ קוד שלא רץ יורדו 5 נק')

במידה ולא הצלחתם לממש שאלה כלשהי, ויש תלות בעיבוד קודם לצורך הפתרון, ישנם קבצי עזר, אותם ניתן להעלות ע"י שימוש בשורות המוסתרות לפני הבדיקות של שאלות אלו.

בכל מקרה אין לשנות את שמות הפונקציות ו/או את הבדיקות המוצעות ב-main ואין לכתוב הערות בעברית.

העלאת בסיס הנתונים לטבלה (df), שהיא מאפיין של המחלקה (Data_Process), ממומשת עבורכם ב-main. כאשר מדפיסים את אובייקט המחלקה ((print(data) , ניתן לצפות בראשית הטבלה:

	Gender	Age	Height	Weight	Team	Year	Medal
0	M	30.0	187.0	76.0	France	2012	NaN
1	M	22.0	189.0	80.0	France	1976	NaN
2	M	21.0	NaN	NaN	France	1956	NaN
3	M	21.0	NaN	NaN	France	1956	Gold
4	M	21.0	198.0	90.0	Italy	2016	Bronze

טבלת סיכום הניקוד:

שאלה	ניקוד יחסי	ציון חלקי והערות לסטודנט.ית
קובץ רץ/לא רץ	5	
1	10	
2	15	
3	15	
4	25	
5	5	
6	25	
סה"כ	100	

1. ממשו את הפונקציה modify_col (10 נק)

הפונקציה מקבלת שם של עמודה (col).

הפונקציה מחליפה ערכים בעמודה col בטבלת הנתונים df לפי כלל המרה מוגדר.

לדוגמא, עבור עמודה 'Medal', וכלל ההחלפה הבא:

כל np.nan מוחלף בספרה 0,

כל "Gold" מוחלף בספרה 3,

כל "Silver" מוחלף ב-2,

וכל "Bronze" מוחלף ב-1.

הפונקציה אינה מחזירה אף ערך, אבל טבלת df משתנה בהתאם.

בדיקה: אחרי הפעלת modify_col על "Medal" לפי הכלל הנ"ל, df נראית כך:

	Gender	Age	Height	Weight	Team	Year	Medal
0	M	30.0	187.0	76.0	France	2012	0
1	M	22.0	189.0	80.0	France	1976	0
2	M	21.0	NaN	NaN	France	1956	0
3	M	21.0	NaN	NaN	France	1956	3
4	M	21.0	198.0	90.0	Italy	2016	1

רמז: ניתן לעשות שימוש בפונקציה replace עם מילון.

2. ממשו את הפונקציה replace_by_mean (15 נק)

הפונקציה מקבלת רשימת שמות עמודות קטגוריות (group_cols) ושם של עמודה רציפה (val). הפונקציה ממלאת את החוסרים בעמודה val בהתאם לממוצע הקבוצתי של עמודות group_cols. במילים אחרות, בכל שורה שבה מופיע חוסר (Nan) בעמודה val, החוסר יוחלף בממוצע ערכי val, עבור הקטגוריות המתאימות מעמודות group_cols.

לדוגמא, בהנחה שהעמודות הקטגוריות הן ["Gender","Team"], ועמודת התוכן המספרי היא "Height". החוסרים בשורות 2 ו-3 שמופיעים בעמודה Height, משוייכים לקטגוריות M ו-France מעמודות Gender ו-Team בהתאמה. לכן, חוסרים אלו יושלמו על ידי הגובה הממוצע של הספורטאים הגברים מצרפת. או במקרה זה, 178.29.

בדיקות:

לאחר הפעלת הפונקציה על עמודות "Height", "Weight" ו-"Age", עם העמודות הקטגוריות ["Gender","Team"], שורות 20-19 בהן הופיעו חוסרים קודם יראו עתה כך:

	Gender	Age	Height	Weight	Team	Year	Medal
17	M	26.000000	183.000000	75.000000	United States	1928	0
18	F	30.000000	163.000000	52.000000	United States	2016	0
19	M	26.729748	178.293243	73.157444	France	1952	0
20	M	20.000000	178.293243	73.157444	France	1952	0
21	M	23.000000	178.000000	61.000000	United States	2000	0
22	M	27.000000	178.000000	61.000000	United States	2004	0
23	M	31.000000	178.000000	61.000000	United States	2008	0
24	M	35.000000	178.000000	61.000000	United States	2012	0

רמז 1: מומלץ להשתמש ב-groupby ובאינדקסים של הקבוצות שלו עבור חיתוך הטבלה בהתאם לדרישות.

רמז 2: ניתן להשתמש בפונקציה transform עם פונקציית עזר על מנת לחשב את הממוצעים הנדרשים להשלמה.

3. ממשו את הפונקציה data_select (15)

הפונקציה מבצעת חיתוך של טבלת df בהתאם לדרישות הבאות:

הפונקציה מקבלת כקלט:

שם של עמודה קטגורית, cat_col

ערך קטגורי, cat

שם של עמודה מספרית, num_col

ערך סף מספרי, threshold

ורשימת שמות של עמודות, cols.

הפונקציה מחלצת את השורות שבהן מופיע בעמודה הקטגורית cat_col הערך cat.

מתוך אותן השורות, הפונקציה בוחרת רק בשורות בהן בעמודה המספרית num_col מופיעים ערכים הגדולים או שוים ל-threshold, ועבור שורות נבחרות אלו, הפונקציה בוחרת רק את העמודות cols.

הפונקציה מחזירה True בלבד, אך משנה את טבלת df של המחלקה להיות הגרסה המצומצמת בהתאם לדרישות החיתוך הנ"ל.

בדיקה: לאחר הפעלת הפונקציה data_select עם בחירה בשורות בהן מופיע F בעמודה Gender, ובחירה בשנים 2000 ומעלה, וצמצום העמודות ל- ["Team", "Year", "Medal"] בלבד, תהיה הטבלה בעלת מימדים: (2598,3) בלבד, וראשיתה תראה כך:

	Team	Year	Medal
18	United States	2016	0
64	United States	2000	0
65	United States	2004	0
66	United States	2008	0
67	United States	2012	0

רמז: ניתן לפתור את הדרישה בשורת קוד אחת, הכוללת חיתוך שורות על ידי שימוש במשפט תנאי מורכב, וחיתוך עמודות לפי הדרישה.

4. ממשו את הפונקציה calc_counts (25)

הפונקציה מקבלת רשימת שמות של עמודות קטגוריות (grps) ושם של עמודת קטגורית נוספת, שרוצים למדוד את התפלגות המופעים שבה (val).

הפונקציה מייצרת טבלת אחוזונים ובה סיכום של אחוז המופעים בכל קטגוריה ב-val, בכל קבוצה מתוך grps.

הפונקציה מחזירה את טבלת אחוזונים זו.

בדיקה: עבור עמודת המדידה (val) "Medal", ועבור העמודה הקטגורית (grps) "Team", התפלגות האחוזונים תסוכם בפלט הפונקציה בטבלה הבאה:

female_scores

	0	3	1	2
France	93.04511278195488	1.3157894736842106	3.7593984962406015	1.8796992481203008
Great Britain	85.95679012345678	4.320987654320987	4.012345679012346	5.709876543209877
Italy	91.13636363636364	3.8636363636363638	2.5	2.5
United States	60.22494887525562	18.50715746421268	8.895705521472392	12.372188139059304

לדוגמא, בצרפת 93.0% מהספורטאיות לא זכו במדליה, 1.32% זכו במקום ראשון, 3.76% זכו במקום שלישי, ו1.88% זכו במקום שני (לפי סדר הופעת העמודות משמאל לימין).

הערה: שימו לב, שהערכים היחודיים בעמודה val מופיעים כשמות העמודות בטבלת האחוזונים המסכמת.

הערה 2: שימו לב, ששמות הקטגוריות ב-grps מופיעות באינדקסים של טבלת האחוזונים המסמכת.

הערה 3: שימו לב כי סך האחוזים בכל שורה (כל קטגוריה מ-grps) הוא 100.

5. ממשו את הפונקציה rename_index (5 נק')

הפונקציה מקבלת טבלה (d), ושם של עמודה חדשה (new_colname)

הפונקציה מאתחלת את שמות השורות (האינדקסים) בטבלה לספרור רץ (מ-0 ועד למספר השורות בטבלה פחות אחד), כאשר היא שומרת את שמות האינדקסים המקוריים בעמודה חדשה ששמה new_colname.

הפונקציה מחזירה את הטבלה המתוקנת.

הערה: ניתן להשתמש בפונקציות reset_index ו-rename.

בדיקה: אחרי הפעלת הפונקציה על הטבלה female_scores תתקבל הטבלה הבאה:

	Country	0	3	1	2
0	France	93.045113	1.315789	3.759398	1.879699
1	Great Britain	85.95679	4.320988	4.012346	5.709877
2	Italy	91.136364	3.863636	2.5	2.5
3	United States	60.224949	18.507157	8.895706	12.372188

6. ממשו את הפונקציה הרקורסיבית selection_sort (25 נק')

הפונקציה מקבלת טבלת נתונים (d), שם של עמודה (col) ומשתנה בוליאני בשם ascending שברירת המחדל שלו היא False.

הפונקציה מממשת את אלגוריתם המיון selection sort (רמת סיבוכיות $O(n^2)$), ומחזירה את הטבלה d, ממויינת לפי עמודה col, בסדר **יורד/עולה** בהתאם לדרישה במשתנה ascending (False זה מיון בסדר יורד, True זה מיון בסדר עולה).

לדוגמא, אם ascending הוא False, והמיון נעשה לפי עמודה col=3 (מייצג זכיה במדליית זהב), טבלת הפלט תראה כך:

	Country	0	3	1	2
0	United States	60.224949	18.507157	8.895706	12.372188
1	Great Britain	85.95679	4.320988	4.012346	5.709877
2	Italy	91.136364	3.863636	2.5	2.5
3	France	93.045113	1.315789	3.759398	1.879699

דוגמא 2:

אם ascending הוא True, והמיון נעשה לפי עמודה col=3 (מייצג זכיה במדליית זהב), טבלת הפלט תראה כך:

	Country	0	3	1	2
0	France	93.045113	1.315789	3.759398	1.879699
1	Italy	91.136364	3.863636	2.5	2.5
2	Great Britain	85.95679	4.320988	4.012346	5.709877
3	United States	60.224949	18.507157	8.895706	12.372188

לסיום, בדקו היטב את פתרונכם, ודאו שאתם מעלים קובץ רץ, בשם id_XXX.py, כאשר XXX מייצג את הת.ז. שלכם, וודאו שהקובץ אכן מכיל את הפתרון שלכם ולא מכיל אף הערה בעברית או בכל שפה אחרת מלבד אנגלית / פייתון.

עלו והצליחו,

צוות הקורס