**מטלה 4**

**Python Algorithms for Industrial Engineers**

by Hadas Lapid, PhD

סטודנטים.ות יקרים.ות,

במטלה זו תתבקשו לנתח בסיס נתונים בשם “employment.csv”.

בסיס הנתונים מכיל נתוני העסקה של 1000 עובדים ממחלקות שונות.

בסיס הנתונים מכיל את העמודות הבאות:

Gender – משתנה בינארי (Male/Female)

Salary – משתנה רציף (float)

Bonus % - משתנה רציף (float)

Team – משתנה קטגורי (בן 11 קטגוריות)

בכל השאלות עליכם להטמיע את הפונקציות הכתובות בשלד.

המטלה מכילה 6 שאלות ו-7 פונקציות למימוש (פונקציה אחת היא פונקציית עזר).

יש לממש את הפונקציות **במקום** המילה pass, כולל הקלט והפלט של כל אחת ואחת מהן, ולבדוק את מימושן באמצעות שורות הבדיקה התואמות ב-main.

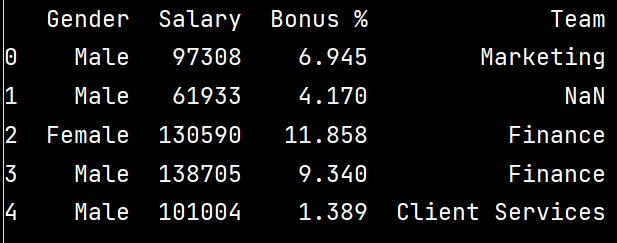
באם צלח המימוש, השאירו את הבדיקות עובדות. באם לא צלח המימוש, יש להסתירו בכדי לאפשר לקובץ לרוץ באופן תקין.

5 נקודות מתוך 100 שמורות לתקינות ריצת הפתרון. (לפתרון לא רץ יורדו 5 נק' באופן אוטומטי)

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 2, 5 ו-6.

**בכל מקרה אין לשנות את שמות הפונקציות ו/או את הבדיקות המוצעות ב-main.**

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



1. **ממשו את הפונקציה omit\_zeros**

הפונקציה מקבלת רשימת עמודות, colnames, מחליפה את החוסרים (nans) בעמודות אלו באפסים, ומוחקת את השורות בהן מופיעים אפסים. העיבוד מתייחס לטבלה df של האובייקט data. הפונקציה לא מחזירה אף פרמטר.

**הערה:** לצורך הפתרון ניתן לעשות שימוש בפונקציות fillna, drop או בחיתוך תואם.

**בדיקות:**

גודל df אחרי העיבוד: (855, 4).

בבדיקת סכום האפסים בעמודות "Gender" , "Salary" ו- "Bonus %", עליהן מופעלת הפונקציה, יופיע 0 בכל אחת מהעמודות בהתאמה.

1. **ממשו את הפונקציה calc\_total**

הפונקציה מקבלת שם של עמודה ראשונה (col1), שם של עמודה שניה (col2) ושם של עמודה חדשה (newcol).

באמצעות פונקציית עזר, f(), המופיעה בתוכה, הפונקציה יוצרת עמודה חדשה בשם newcol בטבלת df, כך שבעמודה החדשה מופיע החישוב של המשכורת הכוללת:

שכר הבסיס + הבונוס באחוזים משכר הבסיס.

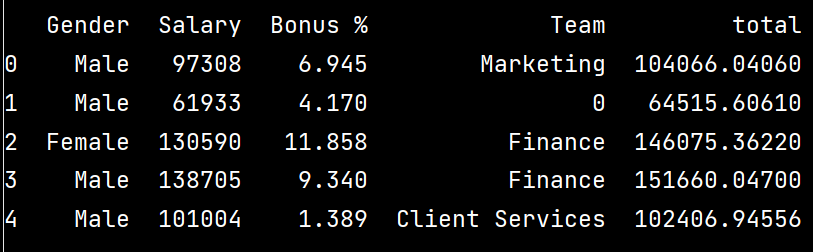
או בנוסחא col1+col1\*(col2/100)

כאשר col1 מייצגת בבדיקה את עמודת השכר ("Salary"), col2 מייצגת את עמודת הבונוס באחוזים ("Bonus %"), והשכר הכולל מיוצג על ידי העמודה החדשה newcol (“total”).

**הערה: חובה** **להשתמש בפונקציה apply** של dataframe, ולעשות שימוש בפונקציית העזר f. פתרונות שלא יעשו שימוש בפונקציית העזר f לא יקבלו ניקוד.

**בדיקות:**

אחרי הפעלת data.calc\_total("Salary", "Bonus %", "total"), ראשית הטבלה df תראה כך:



בבדיקת תוכן העמודה “total”, שהתווספה בעקבות ההפעלה, יתקבל שוויון לחישוב מפורש של הנוסחא:

np.all(np.round((data.df.loc[:,"Salary"]\*(1+(data.df.loc[:,"Bonus %"]/100))),1) == np.round(data.df.loc[:,"total"],1))

יחזיר True.

1. **ממשו את הפונקציה select\_group**

הפונקציה מקבלת שם של עמודה קטגורית (group\_col), ושם של קטגוריה (group).

הפונקציה מחזירה את השורות שבהן מופיעה הקטגוריה group בעמודה group\_col.

**הערה:** לצורך הפתרון יש לעשות שימוש בפונקציה groupby ובאחת מהפונקציות שלה, על מנת לשלוף את תת-הקבוצה המתאימה.

**הערה 2:** הסוג המוחזר מהפונקציה הוא dataframe גם כן.

**בדיקות:**

בשליפת כל ה-female מעמודה “Gender”, תוחזר טבלה שמידותיה (431, 5)

בשליפת כל ה-male מעמודה “Gender”, תוחזר טבלה שמידותיה (424, 5)

1. **ממשו את הפונקציה summ\_by\_group**

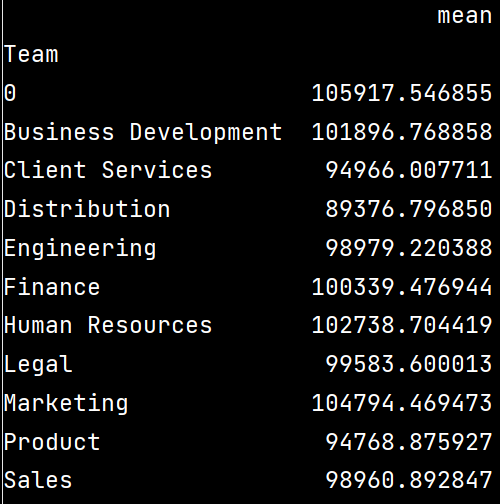
הפונקציה מקבלת טבלת נתונים מסוג DataFrame (d), רשימת שמות עמודות קטגוריות (group\_cols), רשימת שמות עמודות ספרתיות (val\_col) ורשימת פונקציות לחישוב (funcs).

הפונקציה מחזירה טבלה מסכמת ובה חישוב הפונקציות funcs על העמודות הספרתיות val\_col, פר קטגוריה מתוך העמודות הקטגוריאליות group\_cols.

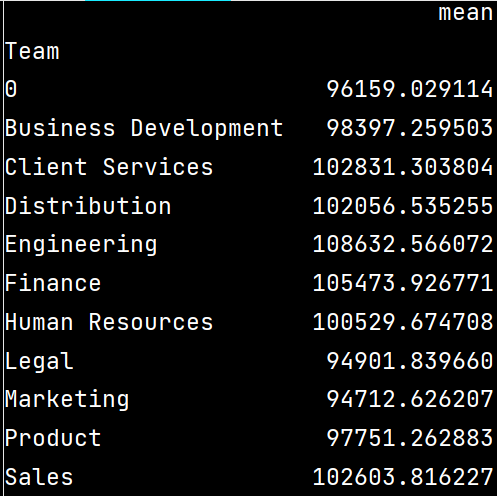
**הערה:** לצורך החישוב יש לעשות שימוש בפונקציה groupby ובה יש לעשות שימוש בפונקציה agg.

**בדיקות:**

בהפעלת הפונקציה על רשימת משכורות הנשים (females\_salaries), עם חלוקה קטגורית לפי מחלקות (["Team"]), בניתוח השכר הכולל (["total"]), ובחישוב הממוצע פר קטגוריה ([np.mean]), תוחזר סדרה בת 11 איברים שמכילה את הממוצעים הבאים:



בהפעלת הפונקציה על רשימת משכורות הגברים (males\_salaries), עם חלוקה קטגורית לפי מחלקות (["Team"]), בניתוח השכר הכולל (["total"]), ובחישוב הממוצע פר קטגוריה ([np.mean]), תוחזר סדרה בת 11 איברים שמכילה את הממוצעים הבאים:



1. **ממשו את הפונקציה concat\_dfs**

הפונקציה מקבלת רשימה של שתי סדרות (dfs) עם ערך ברירת מחדל רשימה ריקה [ ], שם של עמודה ראשונה (col1) עם ערך ברירת מחדל "", שם של עמודה שניה (col2) עם ערך ברירת מחדל "", ושם של עמודה חדשה (newcol) עם ערך ברירת מחדל “diff”.

הפונקציה מחברת את הסדרות בציר העמודות בהתאם לשמות השורות כולל כל הערכים הקיימים בכל אחת מהסדרות.

הפונקציה משנה את שמות העמודות כך שהעמודה הראשונה נקראת col1 והעמודה השניה נקראת col2.

הפונקציה מייצרת עמודה חדשה בשם newcol ובה ההפרש בין הערכים של עמודה col1 לערכים של עמודה col2.

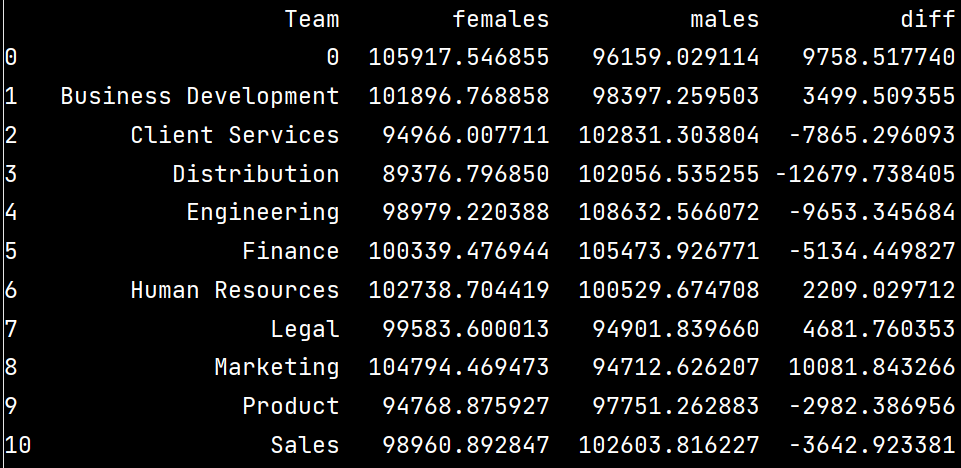
לסיום, הפונקציה מאפסת את מספרי השורות (index) כך שיהיו בסדר עולה מ-0 ועד למספר השורות פחות 1.

הפונקציה מחזירה את הטבלה המאוחדת בעלת עמודת ההפרש החדשה.

אם לא הצלחתם לממש את שאלה 4, ניתן להעלות את הקבצים “females\_salaries.csv”, ואת “males\_salaries.csv” לבצע השמה לסדרות females\_salaries ו- males\_salaries בהתאמה, כפי שממומש בשורות הנסתרות בקוד הבדיקה של שאלה זו.

**בדיקה:**

כאשר מפעילים את concat\_dfs עם הסדרות [females\_salaries,males\_salaries], "females" בתור col1, "males" בתור col2 ו-“diff” בתור newcol, מתקבלת טבלה בת (11,4) שנראית כך:



1. **ממשו את הפונקציות merge\_sort ואת פונקציית העזר merge**

**הפונקציה הרקורסיבית merge\_sort** מקבלת:

- טבלה-DataFrame בשם L

- שם של עמודה (col) בעלת ערך ברירת מחדל "diff".

הפונקציה ממיינת בצורת פיצול ומיזוג (merge sort) את טבלת הקלט L, בהתאם לערכים המופיעים בעמודה col בטבלה.

בתחילה, merge\_sort מפצלת את L לשניים לפי מימד השורות, עד שמגיעה לשורות בודדות.

בתנאי העצירה, היא מחזירה את הטבלאות המכילות שורה בודדת.

לאחר הפיצול merge\_sort ממזגת את הטבלאות המפוצלות על ידי הפעלת הפונקציה merge.

פונקציית המיזוג, merge, מקבלת כקלט:

- תת-טבלה מסוג DataFrame בשם left, ממויינת לפי העמודה col

- תת-טבלה מסוג DataFrame בשם right, ממויינת לפי העמודה col

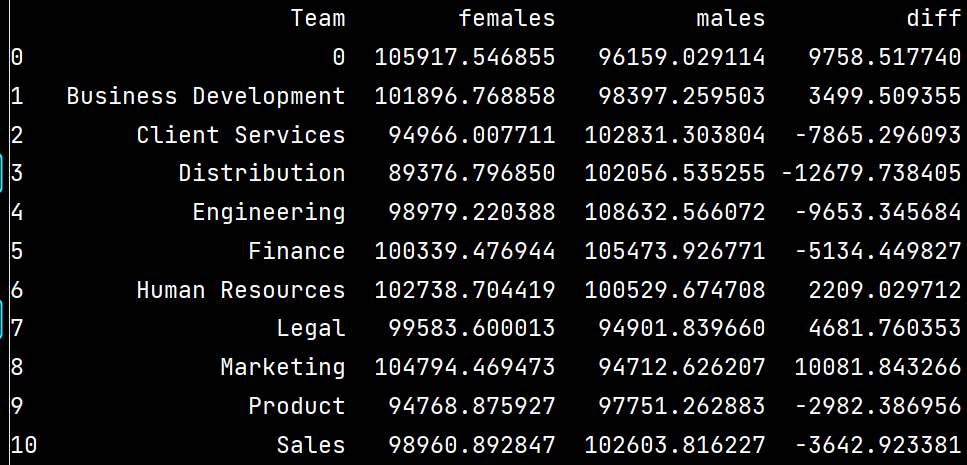
- שם עמודה בשם col, לפיה מתבצע המיזוג. ערך ברירת המחדל של col הוא “diff”

הפונקציה merge מחברת בין השורות של left לבין השורות של right ומחזירה טבלה אחת המכילה את כל השורות של left ושל right, ממויינות בסדר עולה בהתאם לערכים המופיעים בעמודה col.

**הערה:** המימוש האלגוריתמי מתבצע בדיוק כפי שהודגם בכיתה עבור רשימה, רק שכאן יש לפצל ולמזג טבלאות (DataFrames) לפי הערכים המופיעים בעמודה ספציפית.

**הערה 2:** אם הפלט של שאלה 5 לא תואם את הדרישה, ניתן להעלות אותו באופן עצמאי מהקובץ “join\_dfs.csv” על ידי הפעלת השורה join\_dfs = pd.read\_csv("join\_dfs.csv").

**בדיקה:** אם הקלט של merge\_sort הוא טבלת הפלט של שאלה 5, ו-col לפיה מתבצע המיון היא העמודה diff.

:

אז פלט הפונקציה merge\_sort לאחר אתחול האינדקסים של טבלת הפלט (reset\_index) יראה כך:



בדקו היטב את פתרונכם, ודאו שאתם מעלים קובץ רץ, עם הת.ז. שלכם, שמכיל את הפתרון ולא מכיל אף הערה בעברית או בכל שפה אחרת מלבד אנגלית / פייתון.

בהצלחה רבה,

צוות הקורס