מטלה 3
pandas

Python Algorithms for Industrial Engineers

by Hadas Lapid, PhD

הקדמה:

המחלקה Data_Preprocess מטפלת בבסיס הנתונים שנמצא בקובץ המכווץ 'data.pickle'.

ה-constructor של המחלקה ממומש, והוא מאתחל את בסיס הנתונים לטבלה בשם df (public instance dataframe).

ב-main תמצאו את כל הבדיקות הדרושות למימוש המחלקה.

בפרט, אובייקט המחלקה Data_Preprocess מאותחל ב-main תחת השם data.

יש לממש את הפונקציות בהתאם להוראות הניתנות בשאלות להלן.

בכל מקרה אין לשנות את שמות הפונקציות במחלקה, ואין להוסיף הערות בעברית.

במידת הצורך, ניתן להוסיף הערות באנגלית בלבד בגוף הפונקציות.

יש למחוק את ה-pass בכל פונקציה שמממשים.

יש להסיר את ההערות מהבדיקות של כל פונקציה כדי להוכיח נכונות ריצה. הבדיקות תואמות לפלט המצוטט בהסברים להלן.

1. ממשו את הפונקציה intro. הפונקציה מדפיסה את שמות העמודות של df, את מימדי הטבלה, ואת האינפורמציה הבסיסית על תוכן עמודות הטבלה. לפונקציה אין קלט מלבד האובייקט העצמי והיא לא מחזירה פלט.
בהפעלה שלה ב-main היא מדפיסה את הפלטים הבאים למסך:

```
Index(['Area', 'Item', 'Element', 'Year', 'Unit', 'Value'], dtype='object')
```

```
(10418605, 6)
```

```
Data columns (total 6 columns):
 #   Column   Dtype
---  ------   -----
 0   Area     object
 1   Item     object
 2   Element  object
 3   Year     int64
 4   Unit     object
 5   Value    float64
dtypes: float64(1), int64(1), object(4)
```

2. ממשו את הפונקציה describe_object. הפונקציה מתארת עמודות שהן מסוג אובייקט (קטגוריות או strings). הפונקציה מקבלת את האובייקט העצמי ושם עמודה (colname).
הפונקציה מדפיסה את מספר המופעים הייחודי בעמודה (כמות האיברים ב-set של העמודה) ואת כמות החזרות שמופיע כל מופע ייחודי בעמודה.
הפונקציה מחזירה את כמות החזרות שמופיע כל מופע ייחודי בעמודה (פלט הפונקציה הוא מסוג pd.Series).
למשל, עבור העמודה "Area" הפונקציה מדפיסה את המספר 210 ככמות המופעים הייחודיים בעמודה, ומדפיסה ומחזירה את כמות המופעים של כל ערך ייחודי בעמודה:

```
Austria                          79288
Germany                          79182
China                            78970
United Kingdom                   78758
Spain                            78758
                                 ...
Nauru                             6360
Norfolk Island                    5936
Eritrea                           2730
Sudan                             1742
Falkland Islands (Malvinas)        636
Name: Area, Length: 210, dtype: int64
```

3. ממשו את הפונקציה הרקורסיבית describe_all.

הפונקציה מקבלת מספר (num), המסמל מיקום עמודה בטבלת הנתונים df של המחלקה.
הפונקציה עוברת על כל העמודות בבסיס הנתונים ומתארת אותן לפי סוגן.
אם העמודה היא נומרית, הפונקציה מדפיסה את התפלגות האחוזונים שלה.
אחרת, הפונקציה עושה שימוש בפונקציה describe_object (משאלה 2) על מנת לתאר את תוכן העמודה.
שימו לב, כי ניתן לעשות שימוש בפונקציה המובנית np.issubdtype מספריית numpy על מנת להבדיל עמודות נומריות לעמודות שאינן נומריות.
לדוגמא, עבור העמודה הקטגורית "Element", מספר האיברים הייחודיים הוא 4, ומספר המופעים של כל אחד מהאיברים בהתאמה הם:

```
4
Import Value       2974779
Import Quantity    2896046
Export Value       2312019
Export Quantity    2235761
Name: Element, dtype: int64
```

עבור העמודה המספרית "Year", הפונקציה מדפיסה את האומדנים הסטטיסטיים הבאים:

```
count    1.041860e+07
mean     1.987901e+03
std      1.538010e+01
min      1.961000e+03
25%      1.975000e+03
50%      1.989000e+03
75%      2.001000e+03
max      2.013000e+03
Name: Year, dtype: float64
```

4. ממשו את הפונקציה omit_zeros.
הפונקציה פועלת על בסיס הנתונים df של המחלקה.
הפונקציה מוחקת את כל השורות שבהן מופיע 0 בעמודה "Value".
לאחר מחיקת השורות, הפונקציה מאתחלת את ה-index של הטבלה (מפתחות השורות).
הפונקציה מחזירה True לכשמסתיימת.
בדיקות לאחר הפעלת הפונקציה, מימדיי הטבלה וראשיתה נראים בהתאמה כך:

```
after omitting zeros df shape is  (5216095, 6)
```

	Area	Item	Element	Year	Unit	Value
0	Afghanistan	Almonds shelled	Export Quantity	1976	tonnes	642.0
1	Afghanistan	Almonds shelled	Export Quantity	1977	tonnes	286.0
2	Afghanistan	Almonds shelled	Export Quantity	1978	tonnes	518.0
3	Afghanistan	Almonds shelled	Export Quantity	1979	tonnes	1100.0

5. ממשו את הפונקציה filt_top_areas_by_unit.
הפונקציה מקבלת קטגוריה אפשרית מהעמודה 'Unit', (u) ומספר (n) ומצמצמת את טבלת הנתונים df של המחלקה.
בהתאם לחקר קודם של מבנה הנתונים, u יכולה לקבל אחד משני הערכים הבאים:
'1000 US$' או 'tonnes', המופיעים בעמודה זו.
הפונקציה מצמצמת את הטבלה df כך שתכיל רק את השורות בהן היחידות ב-"Unit" הן u.
לדוגמא, אם u הוא "tonnes", לאחר הסינון הראשון מימדי הטבלה יהיו: **(2522557,6)**

לאחר מכן, הפונקציה בודקת כמה מופעים יש בטבלה מכל מדינה (תחת העמודה 'Area'),
ומצמצמת את df כך שתכיל רק את n המדינות, שמספר המופעים שלהן הוא הגדול ביותר.
לדוגמא, בהנחה ש: n=5, חמש המדינות בעלות מספר המופעים הרב ביותר בטבלה הן:

```
top areas are:
 Index(['France', 'Germany', 'Italy', 'United Kingdom', 'Netherlands'], dtype='object')
```

לסיום, הפונקציה מאתחלת מחדש את index השורות לאחר שני מהלכי הצמצום הנ"ל.
לאחר הסינון השני, מימדי הטבלה יהיו (169022,6) וראשיתה תראה כך:

```
5. After filtering product tonnes from 5 most reported areas, df shape is  (169022, 6)
```

	Area	Item	Element	Year	Unit	Value
0	France	Alfalfa meal and pellets	Import Quantity	1961	tonnes	25.0
1	France	Alfalfa meal and pellets	Import Quantity	1962	tonnes	1352.0
2	France	Alfalfa meal and pellets	Import Quantity	1963	tonnes	654.0
3	France	Alfalfa meal and pellets	Import Quantity	1964	tonnes	70.0
4	France	Alfalfa meal and pellets	Import Quantity	1965	tonnes	147.0

הפונקציה אינה מחזירה אף ערך.

6. ממשו את הפונקציה drop_cols.
הפונקציה מקבלת רשימת שמות של עמודות, ומסירה עמודות אלו ללא החזרה מבסיס הנתונים של המחלקה, df.
לדוגמא, אם הקלט של הפונקציה הוא ["Item","Unit"], מימדי הטבלה לאחר הסרה יהיו:

```
df shape after cols reduction (169022, 4)
```

7. ממשו את הפונקציה calc_stats_by_factors.
הפונקציה מקבלת טבלת נתונים (DataFrame),
רשימת שמות של עמודות קטגוריאליות (factors),
שם עמודה נומרית (vals),
ורשימת פונקציות סטטיסטיות (funcs).
הפונקציה מחזירה טבלת סיכום סטטיסטי (אוסף הפלטים של הפונקציות funcs), של הערכים הנומריים (vals) מחולקת לפי קטגוריות (factors).

לדוגמא, כאשר נבדקים הממוצע וסטיית התקן (np.mean,np.std) של נתוני היבוא והיצוא השנתיים (חלוקה לקבוצות לפי ["Year","Element"]) על טבלת הנתונים המספריים בעמודה Value מתוך בסיס הנתונים df של המחלקה, מתקבלת הטבלה הבאה:

```
annual mean and std of export and import quantities
                                mean           std
Year Element
1961 Export Quantity   36634.504831  1.841040e+05
     Import Quantity  113466.270250  5.053679e+05
1962 Export Quantity   35449.094279  1.671386e+05
     Import Quantity  123184.873874  5.596212e+05
1963 Export Quantity   40331.279842  2.280826e+05
                              ...           ...
2011 Import Quantity  270829.209289  9.344836e+05
2012 Export Quantity  242214.809989  1.229958e+06
     Import Quantity  271445.985177  9.412441e+05
2013 Export Quantity  249860.530787  1.368142e+06
     Import Quantity  279329.985160  9.636368e+05
```

8. ממשו את הפונקציה **norm_by_factors**.

הפונקציה מקבלת כקלט **רשימת עמודות** קטגוריות (cols).

הפונקציה מבצעת נרמול של העמודה "Value" פר קטגורייה מהקטגוריות המופיעות בעמודות cols.

הפונקציה מוסיפה עמודה בשם "normed_val", המכילה את הערכים המנורמלים, לטבלת הנתונים df של המחלקה.

הפונקציה לא מחזירה אף ערך בסיומה.

רמז 1: לצורך הפתרון ניתן להשתמש בפונקציה groupby ובפונקציה transform, על טבלת הנתונים df.

רמז 2: נרמול היא פעולה שבה מחסירים מכל ערך באוסף התצפיות את ממוצע התצפיות, ומחלקים בסטיית התקן המדגמית, בהתאם למשוואה:

$$x_i' = \frac{x_i - \bar{x}}{s}$$

כאשר $\bar{x}$ מתייחס לממוצע המדגם, ו-s מתייחס לסטיית תקן שלו.

לדוגמא, בהפעלת הפונקציה עם הרשימה ["Year"] כקלט, df תראה כדלקמן:

```
after z-score normalization by Year


     Area          Element  Year   Value  normed_val
0  France  Import Quantity  1961    25.0   -0.197115
1  France  Import Quantity  1962  1352.0   -0.188760
2  France  Import Quantity  1963   654.0   -0.205212
3  France  Import Quantity  1964    70.0   -0.204510
4  France  Import Quantity  1965   147.0   -0.198772
5  France  Import Quantity  1966    13.0   -0.198826
6  France  Import Quantity  1967    48.0   -0.201112
```

ומימד הטבלה החדש יהיה:

```
(169022, 5)
```

9. ממשו את הפונקציה **split_by_factor**.
הפונקציה מקבלת שם של עמודה קטגוריאלית (factor), וערך אפשרי מעמודה זו (val).
הפונקציה מחלצת מתוך טבלת הנתונים של המחלקה (df) את השורות שבהן מופיע הערך val בעמודה factor.
הפונקציה שומרת העתק מקומי של חיתוך זה מטבלת האם, **ומחזירה אותו כפלט**.
לפני החזרת הפלט, הפונקציה מאתחלת את מספור השורות (index) של הטבלה המצומצמת, כך שהם יהיו מספרים רצים, ללא שמירת מיקומם המקורי.
לדוגמא, כאשר בוחרים את שורות הייצוא ("Export Quantity") מעמודה "Element", מימדי טבלת היצוא הם:

```
Export dataframe shape is  (83403, 5)
```

וראשית הטבלה נראה כך:

```
Export dataframe head:
      Area          Element  Year    Value  normed_val
0  France  Export Quantity  1961  31757.0   -0.115040
1  France  Export Quantity  1962  27132.0   -0.127562
2  France  Export Quantity  1963  45902.0   -0.087230
3  France  Export Quantity  1964  82818.0   -0.000130
4  France  Export Quantity  1965  88253.0    0.000026
```

ואילו כאשר בוחרים את שורות הייבוא ("Import Quantity") מעמודה "Element", מימדי טבלת הייבוא הם:

```
Import dataframe shape is  (85619, 5)
```

וראשיתה נראית כך:

```
Import dataframe head:
      Area          Element  Year   Value  normed_val
0  France  Import Quantity  1961    25.0   -0.197115
1  France  Import Quantity  1962  1352.0   -0.188760
2  France  Import Quantity  1963   654.0   -0.205212
3  France  Import Quantity  1964    70.0   -0.204510
4  France  Import Quantity  1965   147.0   -0.198772
```

שימו לב, שבבדיקת הקוד ב-main בוצעה השמה של טבלאות נתוני היצוא והיבוא הנ"ל במשתנים ציבוריים של האובייקט data (data.export_df ו-data.import_df בהתאמה).

10. ממשו את הפונקציה **merge_dfs.**

הפונקציה מקבלת שתי סדרות בעלות אינדקסים משותפים (s1 ו-s2), ורשימת שמות עמודות בת שני איברים (colnames).

הפונקציה מאחדת את שתי הסדרות ליצירה של טבלה חדשה בת שתי עמודות, ובעלת כל המופעים המשותפים לשתי הסדרות (inner join). הפונקציה מחזירה את הטבלה המאוחדת.

לדוגמא, ממוצע נתוני היצוא השנתיים של כל מדינה סוכמו בסדרה, על ידי שימוש בפונקציה calc_stats_by_factors על טבלת export_df.

עמודות ["Area","Year"] שימשו כעמודות קטגוריאליות לסינון,

"Value" משמשת כעמודה המספרית לביצוע החישוב, ו-np.mean כפונקצייה המסכמת.

כתוצאה מכך, מתקבלת סדרה בת 265 שורות שנראית כך:

data.export_df

France/1961	65710.48276
France/1962	58547.96127
France/1963	78198.07877
France/1964	88947.28378
France/1965	100436.97980

ואילו ממוצע נתוני היבוא השנתיים של כל מדינה סוכמו בסדרה, על ידי שימוש בפונקציה calc_stats_by_factors על טבלת import_df.

עמודות ["Area","Year"] שימשו כעמודות קטגוריאליות לסינון,

"Value" משמשת כעמודה המספרית לביצוע החישוב, ו-np.mean כפונקצייה המסכמת.

המתקבלת סדרה בת 265 שורות שנראית כך:

data.import_df

France/1961	57258.07719
France/1962	67966.00345
France/1963	64453.91611
France/1964	68197.67774
France/1965	68895.10333

הפעלת **merge_dfs**, עם שתי הסדרות הללו כקלט, והרשימה ["Import","Export"] כרשימת שמות העמודות, מייצרת את הטבלה המאוחדת שמושמת ל-data.merged.

merged מכילה 265 שורות ושתי עמודות ונראית כך:

		Import	Export
Area	Year		
France	1961	57258.077193	65710.482759
	1962	67966.003448	58547.961268
	1963	64453.916107	78198.078767
	1964	68197.677741	88947.283784
	1965	68895.103333	100436.979798

11. ממשו את הפונקציה **apply_diff_cols**.

הפונקציה מקבלת טבלת נתונים, d, שמות של שתי עמודות מספריות, c1 ו-c2, ושם של עמודה נוספת, newcol.

הפונקציה מפעילה את פונקציית העזר diff_cols, על ידי שימוש במתודת apply על טבלת הנתונים d, כאשר הקלטים ל-diff_cols הם c1 ו-c2, והמימוש הוא על ציר העמודות.

את הפלט של apply עם diff_cols יש להציב בעמודה חדשה בשם newcol בטבלת d.

הפונקציה מחזירה את הטבלה d, אליה נוספה העמודה החדשה newcol.

לצורך שימוש בתת הפונקציה diff_cols, יש לממשה גם.

diff_cols מקבלת כקלט טבלת נתונים, dat, ושני שמות של עמודות col1 ו-col2.

diff_cols מחזירה את סדרת ההפרשים בין העמודות col1 ו-col2 מ-dat.

כאשר מפעילים את apply_diff_cols עם הטבלה הממוזגת משאלה 10, data.merged, המשתנה c1 מקבל את הערך "Export", המשתנה c2 מקבל את הערך "Import", והמשתנה newcol מקבל את הערך 'GNI', אזי הטבלה המוחזרת מהפונקציה נראית כך:

	Import	Export	GNI
France/1961	57258.07719	65710.48276	8452.40557
France/1962	67966.00345	58547.96127	-9418.04218
France/1963	64453.91611	78198.07877	13744.16266
France/1964	68197.67774	88947.28378	20749.60604
France/1965	68895.10333	100436.97980	31541.87646

הערה: ניתן להעזר בתיעוד הפורמלי של פונקציית apply.

שימו לב לסוגי הקלט של apply: func, args ו-axis הדרושים לפתרון השאלה.