

מבחן פנדס

לפניכם מבחן פיתון העוסק בספרייה pandas
השימוש ב-`AI` לכל צורתי אסור בהחלט.
 מותר להשתמש באתרים הבאים :

<https://docs.python.org/3/>
<https://www.geeksforgeeks.org/>
<https://www.w3schools.com/>
<https://stackoverflow.com/>
<https://realpython.com/>
<https://pandas.pydata.org/docs/>

אין להשתמש בספריה אחרת מלבד pandas וספריות מבוססות של פיתון
אנא קראו היטב את כל ההנחיות לפני התחלת ביצוע המשימה.

היעד במבחן זה: ליצור קובץ CSV המציג את כל השינויים וההגדרות המפורטים בהמשך על אוסף הנתונים הנוכחי.

הנחיות הגשה:

יש להגיש למודול:

1. את כתובות repo כמויד'י מבחן, יש לדאוג שכל הקבצים הבאים יעלו ל: repo
 - a. קבצי המקור - קובץ פונקציות [utils.py](#) וקובץ [main.py](#), ללא הפרדה לתיקיות.
 - b. קובץ [readme](#)
 - c. קובץ הנתונים המקורי [order_sample.json](#)
2. יש להעלות למטרלה במודול את קובץ התוצאה בשם [clean_orders_\[ID_NUMBER\].csv](#).

יש לתכנן את הקוד בחלוקת מודולרית לפונקציות (אין חובה למחוקות ואובייקטים), יש להפריד לקובץ פונקציות [utils.py](#) וקובץ [main.py](#), ללא הפרדה לתיקיות.
 יש להקפיד על תעוד, שמות משתנים משמעתיים, פונקציות קצרות ונכונות.

בכל שאלה תתבקשו לבצע את הפעולות הבאות או חלק מהן, כפי שיפורט בסוגרים בכל שלב:

1. שינוי עמודה קיימת ב-DF (החלפת עמודה לאחר ניקוי או שינוי הנתונים)
2. הוספה עמודה על סמך נתונים מעמודה קיימת ב-DF
3. שינוי סדר הרשומות ב-DF (מיון)
4. מחיקת חלק מהרשומות (פילטר)

תוצר נדרש (פלט):

- יש להציג ולבנות DF שעבר את כל פעולות העבודה שלנו (**למעט פעולות הדפסה למסך**)
- הטבלה הסופית תכיל את כל העמודות מהטבלה המקורי, חלקן לאחר עיבוד, ובនוסף עמודות נוספות שיצרנו, לאחר שהסרנו רשותנו שפילטרנו החוצה ובסדר המין כפי שהוגדר במהלך המטללה.
- ה DF ישמר בסוף התהליך כקובץ במבנה טבלאי מסוג **CSV**.

קלט:

- אוסף הנתונים لمבחן מבוסס על קובץ [ch03_order_sample.json](#) בשם [order_sample.json](#) הוא דата-טט עסק' יומיומי של עסק אינטרנטני.
- כל אובייקט (רשומה) מייצג זמנה אחת באתר ומכיל פרטים רבים אודוטיה.

שלב 0 - (טעינת קובץ)

טען את קובץ ה-hosz לזכרת DataFrame.

בשלב זה עליכם להבין מול איזה נתונים אתם עובדים, איך הנתונים נקלטו ומבנה כללי של הרשומה (יש לוודא שמספר השורות והעמודות תואם לזה שבקובץ).

המלצתה (רשות): בכדי לבצע את המטלה שלפניכם, מומלץ לבדוק את הפרטים הבאים (בכל מקרה **אסור להכניס את הנתונים הנלן לקובץ הסופי**):

10 שורות ראשונות

מספר שורות ועמודות

סוגי העמודות כפי שנקלטו בתהליך טעינת הקובץ

בדקן עבור כל עמודה כמה ערכים ריקים/חסרים

בדקן עבור כל עמודה כמה ערכים "חוויים"

בדקן עבור כל עמודה סטטיסטיות מספריות

שלב 1 - (החלפת עמודות קיימות)

במידת הצורך המר את העמודות הבאות לשוגן נתונים המתאימים (בצעו פעולות ניקיון והמרה):

Total_amount: float (\$)

Shipping_days: int

Customer_age: int

Rating: float

Order_date: datetime

שלב 2 - (החלפת עמודה קיימת)

הסר את סימוני ה-HTML מעמודת items_html כך שתכיל טקסט נקי בלבד (טור הקפה שיישארו הרווחים בין מילים שונים).

שלב 3 - (החלפת עמודה קיימת)

בעמודת used_coupon החלף ערך ריק ("") ב"חסоп כוח" אחרית השאר ערך קיימים.

שלב 4 - (יצירת עמודה חדשה בסוף הטבלה)

צור עמודה חדשה בשם order_month המכילה את חודש ההזמנה ערך מספרי בין 1 ל-12 על פי .order_date

שלב 5 - (יצירת עמודה חדשה בסוף הטבלה ומין הטבלה)

- צור עמודה חדשה בשם order_high_value_high_value_order_true אם ה-total_amount��於 total_amount True ואם ה-total_amount_low_value_order_low_value_order_False False במידה ולא.

- מין את הטבלה מה-total_amount הגובה לנמוך.

שלב 6 - (יצירת עמודה חדשה בסוף הטבלה)

צור עמודה חדשה הכוללת את ממוצע הדירוג rating בעבור כל מדינה country.

שלב 7 - (מחיקת חלק מהשירותות - פילטר)

סנן את הזמנות המופיעות בטבלה, כך שהרשומות הנותרות יעדדו בתנאים הבאים:
הערך של total_amount גובה מ-1000 **וגם** Rating גובה מ-4.5

שלב 8 - (יצירת עמודה חדשה בסוף הטבלה)

צור עמודה חדשה בשם delivery_status המכילה את הערכיהם הבאים:
אם shipping_days גדול מ-7 יהיה הערך "delayed" ובמידה ולא, יהיה הערך "on time".

שלב 9 - (שמירת הטבלה כקובץ)

שמרו את ה DF שיצרתם כקובץ CSV בשם **.clean_orders_[ID_NUMBER].csv**

בצלחה!