# מטלה 2: pandas + Regular Expressions

נתון הדאטה-סט בקובץ: matala2\_cosmetics\_2019-Nov.csv.

את הדאטה אפשר להוריד מכאן:

<https://drive.google.com/file/d/1ILH7VyMQa2rqGs7KZ0ELaWs1PFB6Wk4o/view?usp=sharing>

יש לקרוא את הנתונים ל – dataframe בשם data.

כל שורה ב – dataset מתארת אירוע (פעולה) שביצע משתמש מסוים בזמן מסוים במהלך ביקור (סשן) באתר קניות ברשת.

כל משתמש (user\_id) גולש במהלך ביקוריו באתר (ביקור מזוהה ע"י user\_session) ומבצע פעולות שונות (מזוהות ע"י event\_id) כגון צפיה במוצר (view), הוספת מוצר לעגלה (cart), הסרת מוצר מעגלה (remove\_from\_cart) על מוצרים שונים (product\_id).

משימות:

1. יש להוסיף ל – dataframe עמודה חדשה בשם duration\_to\_next\_event שתחזיק את הזמן בשניות מאירוע עד האירוע הבא אחריו באותו ביקור. לאירוע האחרון בביקור הערך בשדה זה יהיה 0 (כי במקרה הזה אין אירוע נוסף ולכן לא ניתן לחשב משך לאירוע הבא בביקור).
2. יש להוסיף עמודה בשם funnel\_number. פאנל הוא רצף של ביקורים (סשנים) של אותו משתמש באתר שאין ביניהם יותר מ – 5 ימים הפרש. למשל אם משתמש X ביקר באתר ב – 01.01.2023, ב – 02.01.2023, ב – 05.01.2023, ב – 11.01.2023, וב – 12.01.2023, נאמר שהיו לו 2 פאנלים כאשר את הפאנל הראשון של משתמש X (שכולל את הביקורים ב – 01.01.2023, ב – 02.01.2023, ב – 05.01.2023) נסמן כ – funnel\_number=1, ואת הפאנל השני (שכולל את הביקורים ב – 11.01.2023, וב – 12.01.2023) נסמן כ – funnel\_number = 2.
3. יש להוסיף גם עמודה בשם index\_in\_funnel שתכיל את המספר של הביקור (הסשן) באותו הפאנל

למשל, בהמשך לדוגמא הקודמת, למשתמש X, שלושת הביקורים בפאנל הראשון יקבלו את המספרים 1, 2, 3 ושני הביקורים בפאנל השני יקבלו את המספרים 1, 2.

כמובן שאם יש מספר אירועים באותו ביקור – כולם יקבלו את אותו ערך index\_in\_funnel.

1. בעמודת price אמורים להופיע מחירים (float) אך בחלק מהשורות מופיעות מחרוזות כמו למשל 2.44 sale, net 5.44. יש להסיר מהעמודה טקסטים מעבר למחיר עצמו ולהמיר את הערך בשדה ל – float. יש להשתמש ב- regular expressions לשם כך.
2. מה הם סוגי האירועים הקיימים? צור תרשים שמתאר את מספר האירועים מכל סוג
3. יש ליצור dataframe חדש בשם session\_data שבו כל שורה תייצג ביקור באתר ותכלול את העמודות הבאות:
   1. User\_id
   2. User\_session – מזהה ה – session
   3. funnel\_number
   4. index\_in\_funnel
   5. מספר האירועים הכולל שביצוע המשתמש באותו ביקור
   6. משך הביקור באתר
   7. list\_of\_viewed - רשימת מוצרים שבהם צפה המשתמש בביקור
   8. list\_of\_added\_to\_cart – רשימת המוצרים שנוספו לעגלה בביקור
   9. list\_of\_purchased – רשימת המוצרים שנרכשו בביקור