**תרגיל בית 2 – הגורם האנושי**

**מאת: גיל קפלן, יעקב פרידמן והלל טריפ**

בחרנו את הדאטאסט [Zomato Restaurants Dataset](https://www.kaggle.com/datasets/abhijitdahatonde/zomato-restaurants-dataset), שמכיל נתונים ודרוגים של מסעדות.

**שאלה 1:** בחרנו את המשתנים הבאים:

* הדירוג –שזה הדירוג הממוצע של המסעדה בסקאלה של 1 עד 5.

סקאלה: אינטרוולית

ממוצע: 3.51

חציון: 3.5

סטיית תקן: 0.46

שכיח: 2.8

מינימום: 1.8

מקסימום: 4.9

אחוז נתונים חסרים: 0.96%

* המחיר הממוצע לזוג אנשים, ברופי

סקאלה: מנה

הסבר: המחיר הממוצע לארוחה עבור שני אנשים

ממוצע: 540.29

חציון: 400

סטיית תקן: 462.9

שכיח: 300

מינימום: 40

מקסימום: 6000

אחוז נתונים חסרים: 0.80%

* מספר הדירוגים

סקאלה: מנה

הסבר: כמות הדירוגים של המסעדה

ממוצע: 188.9

חציון: 40

סטיית תקן: 592.17

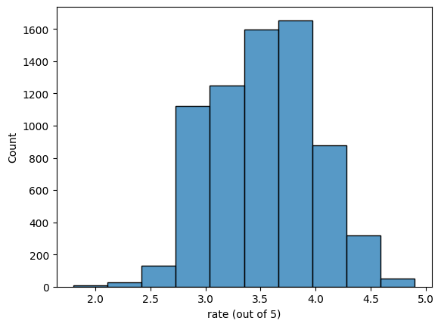
שכיח: 40

מינימום: 1

מקסימום: 16345

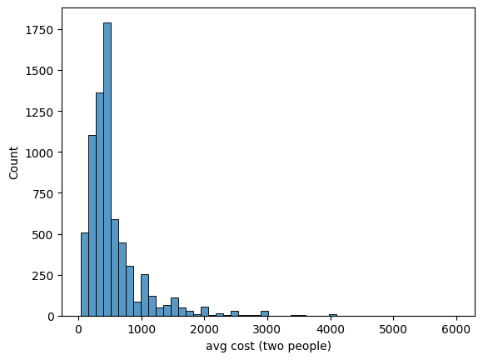
אחוז נתונים חסרים: 0%

**שאלה 2 - וויזואליזציה:**

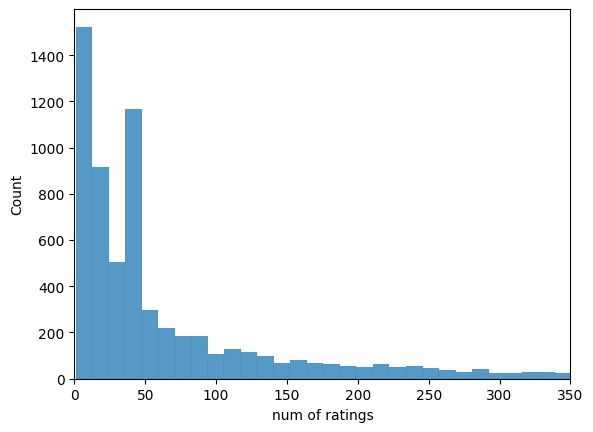
****היסטוגרמה של הדירוג:

רוב הערכים מפוזרים בין 2.5 ל-4.5, כאשר רוב המסעדות הן בעלות דירוג 3.5 עד 4.

היסטוגרמה של המחיר הממוצע לזוג:



מדובר בהתפלגות עם זנב ימני ארוך מאוד שיש בו מעט ערכים, כאשר רוב הערכים נמצאים באזור ה 500.

היסטוגרמה של כמות הדירוגים בתחום של 0 עד 350:

הרוב המוחלט של הערכים הוא בין 0 ל 50. חתכנו את ההיסטוגרמה בערך של 350, מכיוון שיש זנב ימני ארוך מאוד שמגיע עד 16345.

**שאלה 3 – ערכים קיצוניים ונתונים חסרים:**

1. יש ערכים קיצוניים רבים למחיר הממוצע לזוג ולכמות הדירוגים. אלו התפלגויות עם זנב בעל משקל קטן אבל מאוד ארוך.
2. מתוך המשתנים שבחרנו יש כמה עשרות ערכים חסרים בלבד עבור הדירוג והמחיר הממוצע לזוג. הערכים החסרים של הדירוג הם תמיד עבור כמות של 40 דירוגים. הערכים החסרים של החיר הממוצע לזוג נמתאים רובם עבור 4 או 40 דירוגים.

40 הוא הערך השכיח בדאטאסט (שאולי נובע מהדרך שבה הדאטא נאסף). לא ראינו קשר בין הנתונים החסרים לנתונים אחרים.

1. אנו מציעים להוריד את הנתונים החסרים, מכיוון שהם נדירים מאוד ביחס לדאטאסט.
2. כמות הנתונים החסרים היא מסדר גודל של 1%, ולכן אינה משמעותית בניתוח סטטיסטי.